(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 104537028 A
(43)申请公布日 2015.04.22

(21)申请号 201410804222.9

(22)申请日 2014.12.19

(71)申请人 百度在线网络技术(北京)有限公司
地址 100085 北京市海淀区上地十街10号百度大厦三层

(72)发明人 王岳 徐明泉 张琦 秦敏 黄绍建 王玉瑶 崔代锐 邝卓聪

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 路凯 胡彬

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

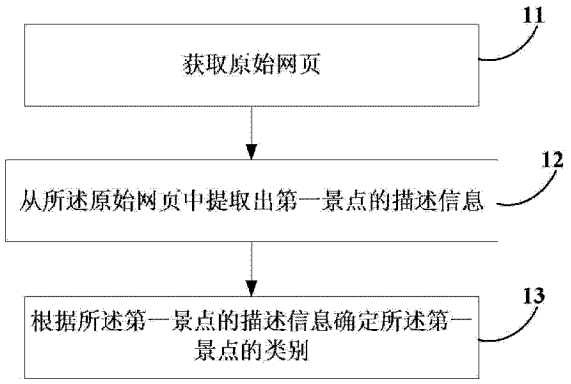
权利要求书5页 说明书16页 附图6页

(54)发明名称

一种网页信息处理方法及装置

(57)摘要

本发明公开了一种网页信息处理方法及装置。所述方法该包括:获取原始网页;从所述原始网页中提取出第一景点的描述信息;根据所述第一景点的描述信息确定所述第一景点的类别。所述装置包括:网页获取模块,用于获取原始网页;第一景点描述信息获取模块,用于从所述原始网页中提取出第一景点的描述信息;第一景点类别确定模块,用于根据所述第一景点的描述信息确定所述第一景点的类别,解决了现有技术中旅游网站提供的景点信息不准确的问题,提高了景点信息的准确性。



CN 104537028 A

1. 一种网页信息处理方法，其特征在于，包括：

获取原始网页；

从所述原始网页中提取出第一景点的描述信息；

根据所述第一景点的描述信息确定所述第一景点的类别。

2. 根据权利要求 1 所述的方法，其特征在于，根据所述第一景点的描述信息确定所述第一景点的类别之前，所述方法还包括：

从所述原始网页中获取第二景点的类别信息和描述信息。

3. 根据权利要求 2 所述的方法，其特征在于，从所述原始网页中获取第二景点的类别信息，包括：

从所述原始网页中获取包含有所述第二景点的旅游路线信息，所述旅游路线信息包括旅游路线及其标签；

统计所述第二景点出现在标注有标签的旅游路线中的次数；

根据统计的次数，将第一标签、第二标签和第三标签作为所述第二景点的类别，其中，所述第二景点出现在标注有所述第一标签的旅游线路中的次数最多，出现在标注有所述第二标签的旅游线路中的次数仅次于标注有所述第一标签的旅游线路，出现在标注有所述第三标签的旅游线路中的次数仅次于标注有所述第一标签和第二标签的旅游线路。

4. 根据权利要求 2 所述的方法，其特征在于，根据所述第一景点的描述信息确定所述第一景点的类别，包括：

根据所述第二景点的类别信息和描述信息以及所述第一景点的描述信息，确定所述第一景点的类别。

5. 根据权利要求 4 所述的方法，其特征在于，根据所述第二景点的类别信息和描述信息以及所述第一景点的描述信息，确定所述第一景点的类别，包括：

利用所述第二景点的类别信息和描述信息训练贝叶斯分类器；

利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类。

6. 根据权利要求 5 所述的方法，其特征在于，利用所述第二景点的类别信息和描述信息训练贝叶斯分类器，包括：

对所述第二景点的描述信息分词，得到训练描述词；

利用所述训练描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第二景点的所有训练描述词，列为所述第二景点的不同训练描述词；

利用所述向量空间模型训练贝叶斯分类器。

7. 根据权利要求 6 所述的方法，其特征在于，利用所述训练描述词，建立向量空间模型，包括：

根据词频 - 逆向文本频率 tf-idf 算法将所述训练描述词去除一半；

利用剩余的训练描述词建立所述向量空间模型。

8. 根据权利要求 5 所述的方法，其特征在于，利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类之前，所述方法还包括：

从所述原始网页中获取第三景点的类别信息和描述信息；

利用所述第三景点的类别信息和描述信息，对所述训练后的贝叶斯分类器进行验证；

验证通过后，触发所述利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所

述第一景点进行分类。

9. 根据权利要求 5 所述的方法，其特征在于，利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类，包括：

对所述第一景点的描述信息分词，得到分类描述词；

利用所述分类描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第一景点的所有分类描述词，列为所述第一景点的不同分类描述词；

利用所述训练后的贝叶斯分类器根据所述向量空间模型对所述第一景点进行分类。

10. 根据权利要求 5 所述的方法，其特征在于，利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类，包括：

将所述训练后的贝叶斯分类器根据所述第一景点的描述信息，得出的第一类别、第二类别和第三类别作为所述第一景点的类别，其中，所述第一类别的后验概率值最大，所述第二类别的后验概率值仅次于所述第一类别，所述第三类别的后验概率值仅次于所述第一类别和第二类别。

11. 根据权利要求 1-10 任一项所述的方法，其特征在于，从所述原始网页中提取出第一景点的描述信息之后，所述方法还包括：

根据所述第一景点的描述信息确定所述第一景点的建议访问时间。

12. 根据权利要求 11 所述的方法，其特征在于，根据所述第一景点的描述信息确定所述第一景点的建议访问时间，包括：

根据所述第一景点的描述信息确定所述第一景点的建议访问月份和建议访问天内时间中的至少一项信息，其中所述建议访问天内时间包括上午和下午中的至少一个时段。

13. 根据权利要求 12 所述的方法，其特征在于，根据所述第一景点的描述信息确定所述第一景点的建议访问月份，包括：

根据所述第一景点的历史被访问时间，统计所述第一景点在不同的月份的历史被访问次数；

利用所述第一景点在不同的月份的历史被访问次数，得到所述第一景点在不同的月份的历史被访问的熵值；

根据所述第一景点在不同的月份的历史被访问的熵值，确定所述第一景点的建议访问月份。

14. 根据权利要求 13 所述的方法，其特征在于，根据所述第一景点在不同的月份的历史被访问的熵值，确定所述第一景点的建议访问月份，包括：

当所述第一景点在不同的月份的历史被访问的熵值之和小于阈值时，将所述第一景点在不同的月份的历史被访问概率中最大的两个月份作为所述建议访问月份。

15. 根据权利要求 12 所述的方法，其特征在于，根据所述第一景点的描述信息确定所述第一景点的建议访问天内时间，包括：

根据所述第一景点在景点序列中的位置及建议访问时长，统计所述第一景点分别在上午和下午的历史被访问次数；

根据所述第一景点分别在上午和下午的历史被访问次数，确定所述第一景点的上午访问指数和下午访问指数；

将确定的上午访问指数和下午访问指数中值最大的访问指数对应的时段，作为所述建

议访问天内时间。

16. 根据权利要求 15 所述的方法，其特征在于，根据所述第一景点在景点序列中的位置及建议访问时长，统计所述第一景点分别在上午和下午的历史被访问次数，包括：

当所述第一景点在一个所述景点序列中排在第一位或第二位，且所述建议访问时长小于预设值时，则将所述第一景点在上午的历史被访问次数加 1；

当所述第一景点在一个所述景点序列中排在倒数第一位或倒数第二位时，则将所述第一景点在下午的历史被访问次数加 1。

17. 根据权利要求 1-10 任一项所述的方法，其特征在于，根据所述第一景点的描述信息确定所述第一景点的类别之后，所述方法还包括：

对应存储所述第一景点的类别和描述信息。

18. 一种网页信息处理装置，其特征在于，包括：

网页获取模块，用于获取原始网页；

信息提取模块，用于从所述原始网页中提取出第一景点的描述信息；

类别确定模块，用于根据所述第一景点的描述信息确定所述第一景点的类别。

19. 根据权利要求 18 所述的装置，其特征在于，所述装置还包括：

第一信息获取模块，用于在所述类别确定模块根据所述第一景点的描述信息确定所述第一景点的类别之前，从所述原始网页中获取第二景点的类别信息和描述信息。

20. 根据权利要求 19 所述的装置，其特征在于，所述第一信息获取模块具体用于：

从所述原始网页中获取包含有所述第二景点的旅游路线信息，所述旅游路线信息包括旅游路线及其标签；

统计所述第二景点出现在标注有标签的旅游路线中的次数；

根据统计的次数，将第一标签、第二标签和第三标签作为所述第二景点的类别，其中，所述第二景点出现在标注有所述第一标签的旅游线路中的次数最多，出现在标注有所述第二标签的旅游线路中的次数仅次于标注有所述第一标签的旅游线路，出现在标注有所述第三标签的旅游线路中的次数仅次于标注有所述第一标签和第二标签的旅游线路。

21. 根据权利要求 19 所述的装置，其特征在于，所述类别确定模块具体用于：

根据所述第二景点的类别信息和描述信息以及所述第一景点的描述信息，确定所述第一景点的类别。

22. 根据权利要求 20 所述的装置，其特征在于，所述类别确定模块包括：

第一训练子模块，用于利用所述第二景点的类别信息和描述信息训练贝叶斯分类器；

第一分类子模块，用于利用所述第一训练子模块训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类。

23. 根据权利要求 22 所述的装置，其特征在于，所述第一训练子模块包括：

第一分词子模块，用于对所述第二景点的描述信息分词，得到训练描述词；

第一模型建立子模块，用于利用所述训练描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第二景点的所有训练描述词，列为所述第二景点的不同训练描述词；

第二训练子模块，用于利用所述向量空间模型训练贝叶斯分类器。

24. 根据权利要求 23 所述的装置，其特征在于，所述第一模型建立子模块具体用于：

根据词频－逆向文本频率 tf-idf 算法将所述训练描述词去除一半；

利用剩余的训练描述词建立所述向量空间模型。

25. 根据权利要求 22 所述的装置，其特征在于，所述装置还包括：

第二信息获取模块，用于在所述第一分类子模块利用所述第一训练子模块训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类之前，从所述原始网页中获取第三景点的类别信息和描述信息；

验证模块，用于利用所述第三景点的类别信息和描述信息，对所述训练后的贝叶斯分类器进行验证；

触发模块，用于在所述验证模块对所述训练后的贝叶斯分类器的验证通过后，触发所述第一分类子模块利用所述训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类。

26. 根据权利要求 22 所述的装置，其特征在于，所述第一分类子模块包括：

第二分词子模块，用于对所述第一景点的描述信息分词，得到分类描述词；

第二模型建立子模块，用于利用所述分类描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第一景点的所有分类描述词，列为所述第一景点的不同分类描述词；

第二分类子模块，用于利用所述训练后的贝叶斯分类器根据所述向量空间模型对所述第一景点进行分类。

27. 根据权利要求 22 所述的装置，其特征在于，所述第一分类子模块具体用于：

将所述训练后的贝叶斯分类器根据所述第一景点的描述信息，得出的第一类别、第二类别和第三类别作为所述第一景点的类别，其中，所述第一类别的后验概率值最大，所述第二类别的后验概率值仅次于所述第一类别，所述第三类别的后验概率值仅次于所述第一类别和第二类别。

28. 根据权利要求 18-27 任一项所述的装置，其特征在于，所述装置还包括：

时间确定模块，用于在所述信息提取模块从所述原始网页中提取出第一景点的描述信息之后，根据所述第一景点的描述信息确定所述第一景点的建议访问时间。

29. 根据权利要求 28 所述的装置，其特征在于，所述时间确定模块具体用于：

根据所述第一景点的描述信息确定所述第一景点的建议访问月份和建议访问天内时间中的至少一项信息，其中所述建议访问天内时间包括上午和下午中的至少一个时段。

30. 根据权利要求 29 所述的装置，其特征在于，所述时间确定模块具体用于：

根据所述第一景点的历史被访问时间，统计所述第一景点在不同的月份的历史被访问次数；

利用所述第一景点在不同的月份的历史被访问次数，得到所述第一景点在不同的月份的历史被访问的熵值；

根据所述第一景点在不同的月份的历史被访问的熵值，确定所述第一景点的建议访问月份。

31. 根据权利要求 30 所述的装置，其特征在于，所述时间确定模块具体用于：

当所述第一景点在不同的月份的历史被访问的熵值之和小于阈值时，将所述第一景点在不同的月份的历史被访问概率中最大的两个月份作为所述建议访问月份。

32. 根据权利要求 29 所述的装置，其特征在于，所述时间确定模块具体用于：

根据所述第一景点在景点序列中的位置及建议访问时长，统计所述第一景点分别在上午和下午的历史被访问次数；

根据所述第一景点分别在上午和下午的历史被访问次数，确定所述第一景点的上午访问指数和下午访问指数；

将确定的上午访问指数和下午访问指数中值最大的访问指数对应的时段，作为所述建议访问天内时间。

33. 根据权利要求 32 所述的装置，其特征在于，所述时间确定模块具体用于：

当所述第一景点在一个所述景点序列中排在第一位或第二位，且所述建议访问时长小于预设值时，则将所述第一景点在上午的历史被访问次数加 1；

当所述第一景点在一个所述景点序列中排在倒数第一位或倒数第二位时，则将所述第一景点在下午的历史被访问次数加 1。

34. 根据权利要求 18-27 任一项所述的装置，其特征在于，所述装置还包括：

存储模块，用于在所述类别确定模块根据所述第一景点的描述信息确定所述第一景点的类别之后，对应存储所述第一景点的类别和描述信息。

一种网页信息处理方法及装置

技术领域

[0001] 本发明实施例涉及信息处理技术，尤其涉及一种网页信息处理方法及装置。

背景技术

[0002] 随着互联网与旅游业的不断发展，人们可以随时随地从旅游网站上了解各种旅游信息。

[0003] 但是，目前旅游网站的景点详情信息，由不同的旅游编辑编辑，而每个旅游编辑可能只是熟悉某一个或者某几个目的地，且提供的信息具有很大的主观性，导致同一个景点被标注上不同甚至是互斥的标签。如同一景点可能被打上独行和家庭游等互斥标签，导致提供的信息不客观也不准确。

发明内容

[0004] 本发明实施例提供一种网页信息处理方法及装置，以提高景点信息的准确性。

[0005] 第一方面，本发明实施例提供了一种网页信息处理方法，包括：

[0006] 获取原始网页；

[0007] 从所述原始网页中提取出第一景点的描述信息；

[0008] 根据所述第一景点的描述信息确定所述第一景点的类别。

[0009] 第二方面，本发明实施例还提供了一种网页信息处理装置，包括：

[0010] 网页获取模块，用于获取原始网页；

[0011] 信息提取模块，用于从所述原始网页中提取出第一景点的描述信息；

[0012] 类别确定模块，用于根据所述第一景点的描述信息确定所述第一景点的类别。

[0013] 本发明实施例提供的一种网页信息处理方法及装置，通过获取原始网页，从所述原始网页中提取出第一景点的描述信息，并根据所述第一景点的描述信息确定所述第一景点的类别，解决了现有技术中旅游网站提供的景点信息不准确的问题，提高了景点信息的准确性和客观性。

附图说明

[0014] 图 1 为本发明实施例一提供的一种网页信息处理方法的流程图；

[0015] 图 2 为本发明实施例二提供的一种网页信息处理方法的流程图；

[0016] 图 3 为本发明实施例三提供的网页信息处理方法中训练贝叶斯分类器的流程图；

[0017] 图 4 为本发明实施例四提供的网页信息处理方法中利用训练后的贝叶斯分类器对第一景点进行分类的流程图；

[0018] 图 5 为本发明实施例五提供的网页信息处理方法中对训练后的贝叶斯分类器进行验证的流程图；

[0019] 图 6 为本发明实施例六提供的网页信息处理方法中确定建议访问月份方法的流程图；

[0020] 图 7 为本发明实施例七提供的网页信息处理方法中确定建议访问天内时间方法的流程图；

[0021] 图 8 为本发明实施例八提供的网页信息处理方法中确定景点的类别的流程示意图；

[0022] 图 9 为本发明实施例九提供的网页信息处理方法中确定景点的建议访问月份的流程示意图；

[0023] 图 10 为本发明实施例十提供的网页信息处理方法中确定建议访问天内时间的流程示意图；

[0024] 图 11 为本发明实施例十一提供的一种网页信息处理装置的结构示意图。

具体实施方式

[0025] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释本发明，而非对本发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本发明相关的部分而非全部结构。

[0026] 本发明实施例的网页信息处理方法可以由网页信息处理装置在线下或离线状态下执行，该装置可通过硬件和 / 或软件的方式实现，并一般可集成于服务端所在的终端设备如服务器中，或作为服务端的子程序。

[0027] 实施例一

[0028] 参见图 1，本实施例提供的网页信息处理方法具体包括：操作 11- 操作 13。

[0029] 操作 11 中，获取原始网页。

[0030] 例如，可以获取各旅游网站的原始网页，或者旅游论坛的原始网页。去哪网，携程网，百度旅游等网站的原始网页大多为旅游编辑手动编辑，或者由游客自行根据网站提供的模板编辑的游记，记录了行程概要、旅游攻略和景点图片等。

[0031] 优选的，在获取原始网页时，选择包含有结构化较好的游记的原始网页，如游记提供了详细的行程概要，类似于：第一天：景点 1--> 景点 2--> 景点 3；第二天：景点 1--> 景点 2。选择包含有结构化较好的游记的原始网页可以节省数据挖掘时间。

[0032] 操作 12 中，从所述原始网页中提取出第一景点的描述信息。

[0033] 例如，可以从原始网页中获取结构化较好的游记或游记攻略，然后可以利用语义分析技术，从旅游攻略、游记中提取出第一景点的描述信息，还可以从旅游网站如百度旅游网站为每个景点编辑的信息介绍的原始网页中，直接提取第一景点的描述信息。

[0034] 需要说明的是，景点的描述信息应尽量的有区分度，如很受欢迎、很好等等描述词就没有区分度，如红叶很多、水流很急等描述词则具有区分度。

[0035] 其中，第一景点中的第一并无特殊含义，只是为了更清楚的描述技术方案。

[0036] 操作 13 中，根据所述第一景点的描述信息确定所述第一景点的类别。

[0037] 假设第一景点为十渡，上述操作 12 中获取的十渡的描述信息为“漂流是十渡旅游的灵魂等”，根据描述信息确定所述十渡的类别为漂流。具体地，根据所述第一景点的描述信息确定所述第一景点的类别的方式，可以是语义语法分析，还可以是根据分类器分类等，本发明实施例对实现方式不作限制。

[0038] 本实施例提供的网页信息处理方法，通过获取原始网页，并从所述原始网页中提

取出第一景点的描述信息，根据所述第一景点的描述信息确定所述第一景点的类别，解决了现有技术中旅游网站提供的景点信息不准确的问题提高了景点信息的准确性和客观性。

[0039] 示例性的，上述根据所述第一景点的描述信息确定所述第一景点的类别之前，本发明实施例提供的网页信息处理方法还包括：

[0040] 从所述原始网页中获取第二景点的类别信息和描述信息。

[0041] 其中，第二景点中的第二以及下述的第三等词并无特殊含义，只是为了更清楚的描述技术方案。

[0042] 示例性的，上述从所述原始网页中获取第二景点的类别信息，包括：

[0043] 从所述原始网页中获取包含有所述第二景点的旅游路线信息，所述旅游路线信息包括旅游路线及其标签；

[0044] 统计所述第二景点出现在标注有标签的旅游路线中的次数；

[0045] 根据统计的次数，将第一标签、第二标签和第三标签作为所述第二景点的类别，其中，所述第二景点出现在标注有所述第一标签的旅游线路中的次数最多，出现在标注有所述第二标签的旅游线路中的次数仅次于标注有所述第一标签的旅游线路，出现在标注有所述第三标签的旅游线路中的次数仅次于标注有所述第一标签和第二标签的旅游线路。

[0046] 示例性的，上述根据所述第一景点的描述信息确定所述第一景点的类别，包括：

[0047] 根据所述第二景点的类别信息和描述信息以及所述第一景点的描述信息，确定所述第一景点的类别。

[0048] 示例性的，上述根据所述第二景点的类别信息和描述信息以及所述第一景点的描述信息，确定所述第一景点的类别，包括：

[0049] 利用所述第二景点的类别信息和描述信息训练贝叶斯分类器；

[0050] 利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类。

[0051] 示例性的，上述利用所述第二景点的类别信息和描述信息训练贝叶斯分类器，包括：

[0052] 对所述第二景点的描述信息分词，得到训练描述词；

[0053] 利用所述训练描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第二景点的所有训练描述词，列为所述第二景点的不同训练描述词；

[0054] 利用所述向量空间模型训练贝叶斯分类器。

[0055] 示例性的，上述利用所述训练描述词，建立向量空间模型，包括：

[0056] 根据词频－逆向文本频率 tf-idf 算法将所述训练描述词去除一半；

[0057] 利用剩余的训练描述词建立所述向量空间模型。

[0058] 示例性的，上述利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类之前，还包括：

[0059] 从所述原始网页中获取第三景点的类别信息和描述信息；

[0060] 利用所述第三景点的类别信息和描述信息，对所述训练后的贝叶斯分类器进行验证；

[0061] 验证通过后，触发所述利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类。

[0062] 示例性的，上述利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类，包括：

[0063] 对所述第一景点的描述信息分词，得到分类描述词；

[0064] 利用所述分类描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第一景点的所有分类描述词，列为所述第一景点的不同分类描述词；

[0065] 利用所述训练后的贝叶斯分类器根据所述向量空间模型对所述第一景点进行分类。

[0066] 示例性的，上述利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类，包括：

[0067] 将所述训练后的贝叶斯分类器根据所述第一景点的描述信息，得出的第一类别、第二类别和第三类别作为所述第一景点的类别，其中，所述第一类别的后验概率值最大，所述第二类别的后验概率值仅次于所述第一类别，所述第三类别的后验概率值仅次于所述第一类别和第二类别。

[0068] 示例性的，上述从所述原始网页中提取出第一景点的描述信息之后，本发明实施例提供的网页信息处理方法还包括：

[0069] 根据所述第一景点的描述信息确定所述第一景点的建议访问时间。

[0070] 示例性的，上述根据所述第一景点的描述信息确定所述第一景点的建议访问时间，包括：

[0071] 根据所述第一景点的描述信息确定所述第一景点的建议访问月份和建议访问天内时间中的至少一项信息，其中所述建议访问天内时间包括上午和下午中的至少一个时段。

[0072] 示例性的，上述根据所述第一景点的描述信息确定所述第一景点的建议访问月份，包括：

[0073] 根据所述第一景点的历史被访问时间，统计所述第一景点在不同的月份的历史被访问次数；

[0074] 利用所述第一景点在不同的月份的历史被访问次数，得到所述第一景点在不同的月份的历史被访问的熵值；

[0075] 根据所述第一景点在不同的月份的历史被访问的熵值，确定所述第一景点的建议访问月份。

[0076] 示例性的，上述根据所述第一景点在不同的月份的历史被访问的熵值，确定所述第一景点的建议访问月份，包括：

[0077] 当所述第一景点在不同的月份的历史被访问的熵值之和小于阈值时，将所述第一景点在不同的月份的历史被访问概率中最大的两个月份作为所述建议访问月份。

[0078] 示例性的，上述根据所述第一景点的描述信息确定所述第一景点的建议访问天内时间，包括：

[0079] 根据所述第一景点在景点序列中的位置及建议访问时长，统计所述第一景点分别在上午和下午的历史被访问次数；

[0080] 根据所述第一景点分别在上午和下午的历史被访问次数，确定所述第一景点的上午访问指数和下午访问指数；

[0081] 将确定的上午访问指数和下午访问指数中值最大的访问指数对应的时段，作为所述建议访问天内时间。

[0082] 示例性的，上述根据所述第一景点在景点序列中的位置及建议访问时长，统计所述第一景点分别在上午和下午的历史被访问次数，包括：

[0083] 当所述第一景点在一个所述景点序列中排在第一位或第二位，且所述建议访问时长小于预设值时，则将所述第一景点在上午的历史被访问次数加 1；

[0084] 当所述第一景点在一个所述景点序列中排在倒数第一位或倒数第二位时，则将所述第一景点在下午的历史被访问次数加 1。

[0085] 示例性的，上述根据所述第一景点的描述信息确定所述第一景点的类别之后，本发明实施例提供的网页信息处理方法还包括：

[0086] 对应存储所述第一景点的类别和描述信息。

[0087] 实施例二

[0088] 本实施例在上述各实施例的基础上提供了另一种网页信息处理方法。具体地，在根据所述第一景点的描述信息确定所述第一景点的类别之前，还包括从所述原始网页中获取第二景点的类别信息和描述信息。

[0089] 参见图 2，本实施例二提供的网页信息处理方法具体包括：操作 21- 操作 24。

[0090] 操作 21 中，获取原始网页。

[0091] 操作 22 中，从所述原始网页中提取出第一景点的描述信息。

[0092] 其中，操作 21 和操作 22 与实施例一中的操作 11 和操作 12 的实施过程相同，这里不再赘述。

[0093] 操作 23 中，从所述原始网页中获取第二景点的类别信息和描述信息。

[0094] 其中，从所述原始网页中获取第二景点的描述信息，与实施例一中从所述原始网页中获取第一景点的描述信息的实施过程相同，这里不再赘述。

[0095] 具体地，可以先从所述原始网页中获取包含有所述第二景点的旅游路线信息，所述旅游路线信息包括旅游路线及其标签，其中，标签表征旅游路线的特色，获取的原始网页的数量可以根据实际应用情况而定；然后统计所述第二景点出现在标注有标签的旅游路线中的次数；最后根据统计的次数，将第一标签、第二标签和第三标签作为所述第二景点的类别。其中，所述第二景点出现在标注有所述第一标签的旅游线路中的次数最多，出现在标注有所述第二标签的旅游线路中的次数仅次于标注有所述第一标签的旅游线路，出现在标注有所述第三标签的旅游线路中的次数仅次于标注有所述第一标签和第二标签的旅游线路。

[0096] 例如，不同的旅游网站给出北京欢乐谷的旅游特色标签不同，容易引起用户选择上的困扰。所以，可以统计景点北京欢乐谷出现在标注有各标签的旅游路线中的次数，按照次数大小对各标签进行排序，次数越多的标签其排名越靠前，从而为用户提供正确的导向，避免标签不同带来的困扰。假设北京欢乐谷这个景点在亲子游路线方案中出现了 5 次，在周边游出现了 2 次，在蜜月游中出现了 1 次，在红色旅游中出现了 0 次。那么，获取的景点北京欢乐谷的标签有：亲子，周边游，蜜月游，红色旅游等。各标签的次数依次为：亲子 5 次，周边游 2 次，蜜月游 1 次，红色旅游 0 次。根据上述统计的次数，将次数依次从高到低的前三个标签，作为景点北京欢乐谷的类别，即：亲子，周边游，蜜月游。这里，所选标签的个数仅为举例而非限制。

[0097] 操作 24 中，根据所述第二景点的类别信息和描述信息以及所述第一景点的描述信息，确定所述第一景点的类别。

[0098] 例如，利用所述第二景点的类别信息和描述信息训练贝叶斯分类器；利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类。

[0099] 其中，贝叶斯分类器的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。也就是说，贝叶斯分类器是最小错误率意义上的优化。目前研究较多的贝叶斯分类器主要有四种，分别是：Naive Bayes、TAN、BAN 和 GBN。将上述第二景点作为训练样本，利用所述第二景点的类别信息和描述信息训练贝叶斯分类器，然后，利用训练后的贝叶斯分类器根据所述第一景点的描述信息对对没有分类的第一景点进行分类。

[0100] 本实施例提供的网页信息处理方法，通过从原始网页中提取出第二景点的类别信息和描述信息，根据所述第二景点的类别信息和描述信息以及所述第一景点的描述信息，确定所述第一景点的类别，使得景点的特色标签更加客观准确，避免了不同的旅游网站对同一景点给出的旅游特色标签不同给用户带来的困扰，解决了现有技术中旅游网站提供的景点信息不准确的问题，提高了景点信息的准确性，节约了获取正确景点信息的时间和成本。

[0101] 在上述各个实施例的基础上，优选地，根据所述第一景点的描述信息确定所述第一景点的类别之后，本发明实施例提供的网页信息处理方法还包括：对应存储所述第一景点的类别和描述信息，以及对应存储上述第二景点的类别和描述信息，形成景点知识库或景点信息库，以供线上查询。

[0102] 实施例三

[0103] 本实施例以上述实施例为基础，给出了网页信息处理方法中利用第二景点训练贝叶斯分类器的方法。

[0104] 参见图 3，本发明实施例提供的训练贝叶斯分类器的方法具体包括：操作 31- 操作 33。

[0105] 操作 31 中，对所述第二景点的描述信息分词，得到训练描述词。

[0106] 假设景点的类别包括：休闲、亲子、情侣、历史、毕业、独行、家庭游、户外、摄影、姐妹游、艺术、民俗、宗教、徒步、蜜月、自驾游、探秘、踏青、骑行、赏花、购物游、文艺游、美食、避暑、漂流、滑雪、骑马、探险、民署、人文和购物等等，那么，每一类别中可至少选取一个第二景点，并将上述选取的所有第二景点的描述信息分词，得到训练描述词。该训练描述词为一初步的词表集合。例如，对于休闲、历史、亲子和情侣这四个类别，分别选取一个第二景点。假设休闲类别的景点选十渡，对应的训练描述词为水流湍急、河水蜿蜒、划船和散步；历史类别的景点选故宫，对应的训练描述词为皇宫、古建筑和世界文化遗产；亲子类别的景点选长城，对应的训练描述词为世界文化遗产和古建筑；情侣类别的景点选后海对应的训练描述词为古建筑、划船、散步和美食。

[0107] 操作 32 中，利用所述训练描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第二景点的所有训练描述词，列为所述第二景点的不同训练描述词。

[0108] 以操作 31 中的例子为例，得到的空间向量模型如下表所示：

[0109]

	水流湍急	河水蜿蜒	皇宫	古建筑	世界文化遗产	划船	散步	美食
十渡	1	1	0	0	0	1	1	0
故宫	0	0	1	1	1	0	0	0
长城	0	0	0	1	1	0	0	0
后海	0	0	0	1	0	1	1	1

[0110] 其中，1 和 0 为空间向量模型中的描述向量，与训练描述词相对应。

[0111] 操作 33 中，利用所述向量空间模型训练贝叶斯分类器。

[0112] 由于上述向量空间模型中，给出了景点的类别和描述词，那么训练贝叶斯分类器就是要得到各类别可能对应的描述词有哪些，从而以此为依据对已有描述词的景点进行分类。

[0113] 训练时，首先计算每一第二景点向量空间模型中的训练描述词，属于每一类的条件概率，将该概率作为某分类下某训练描述词出现的条件概率；再计算某一些训练描述词属于某一类的概率。例如训练描述词“水流湍急”属于“漂流”的分类概率更大，而训练描述词“古建筑”属于“历史”的概率更大。训练贝叶斯分类器通过计算第二景点向量空间模型中的训练描述词属于各类的概率，得到各类别所对应的训练描述词。其中，第二景点向量空间模型属于各类别的概率用符号 $P(\omega 1|x), P(\omega 2|x), \cdots, P(\omega n|x)$ 表示。比较这些条件概率，最大数值所对应的类别 ωi 就是该模式所属的类。其中，x 为向量空间模型中的训练描述词，ωi 为第 i 个类别 $(1 \leqslant i \leqslant n)$，n 为类别数量。

[0114] 本发明实施例提供的训练贝叶斯分类器的方法，通过得到所述第二景点的训练描述词，并利用所述训练描述词建立向量空间模型，然后利用所述向量空间模型训练贝叶斯分类器，实现对未标注标签的第一景点进行分类，提高了景点分类的准确性。

[0115] 在上述实施例的基础上，优选地，对操作 32 中获取的训练描述词进行过滤。例如，训练描述词中出现“这里”、“非常”、“好玩”、“的”等不具区分度的训练描述词，则需要过滤掉。

[0116] 优选地，过滤时可以根据 tf-idf(term frequency－inverse document frequency 词频－逆向文本频率）算法去除一半的训练描述词，利用剩余的训练描述词建立向量空间模型。tf-idf 算法是一种用于资讯检索与资讯探勘的常用加权技术，因此应用此算法所去除的训练描述词对于景点分类不具区分度，减小了训练分类器时的向量空间模型的维度，节省了计算时间。

[0117] 进一步地，训练时可根据非负矩阵分解 (Non-negative Matrix

Factorization,NMF) 得到一个景点的前 30 维最重要的训练描述词，然后利用该 30 维最重要的训练描述词训练贝叶斯分类器。这里，前 30 维最重要的训练描述词是指最具有区分度的描述词，例如描述信息中出现“这个景点很受欢迎”，“这个景点适合情侣出游”等，显然的上述两个描述信息中，后一句对于该景点的分类更具有区分度，因此删除“这个景点很受欢迎”，将剩余的训练描述词放到贝叶斯分类器训练，训练出一个分类器，以供后续对于未标注特色标签的第一景点进行分类。

[0118] 实施例四

[0119] 本实施例以上述实施例为基础，提供了网页信息处理方法中一种利用训练后的贝叶斯分类器对第一景点进行分类的方法。

[0120] 参见图 4，本发明实施例四提供的利用训练后的贝叶斯分类器对第一景点进行分类具体包括：操作 41- 操作 43。

[0121] 操作 41 中，对所述第一景点的描述信息分词，得到分类描述词。

[0122] 操作 42 中，利用所述分类描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第一景点的所有分类描述词，列为所述第一景点的不同分类描述词。这里，可以根据非负矩阵分解同样取第一景点的分类描述词中前 30 维最重要的特征，即选取最具区分度的 30 个分类描述词。

[0123] 其中，操作 41 和操作 42 与实施例三中的操作 31 和操作 32 的实施过程类似，这里不再赘述。

[0124] 操作 43 中，利用所述训练后的贝叶斯分类器根据所述向量空间模型对所述第一景点进行分类。

[0125] 例如，利用第二景点十渡训练的贝叶斯分类器，由于十渡景点（标签被标注为漂流、周边、……），十渡的描述信息中出现了周围水流湍急等等描述信息，因此，后验概率 p（漂流｜水流湍急）最大。如果所需分类的第一景点的描述中也出现了水流湍急等，那么第一景点的描述信息中，p（漂流｜第一景点）＝ p（漂流｜水流湍急）p（漂流｜分类描述词 2）p（漂流｜分类描述词 3）……，势必要大于其他类，比如似于 p（红色｜第一景点）等。因此，将第一景点分类为“漂流”。

[0126] 由于利用训练后的分类器对第一景点进行分类，得到的是第一景点属于各类别的概率，因此，还可以从第一景点属于各类别的概率中选取概率最高前三类，作为所述第一景点的类别，并对其进行标注，例如对景点故宫分类后，概率最高的类别依次是历史、休闲、家庭游。

[0127] 本实施例提供的分类方法，通过利用所述第一景点的描述信息分词建立向量空间模型，并利用所述训练后的贝叶斯分类器根据所述向量空间模型对所述第一景点进行分类，提高了景点类别的准确性和客观性。

[0128] 实施例五

[0129] 本实施例以上述实施例为基础，提供了对训练后的贝叶斯分类器进行验证的方法。具体地，在利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类之前，还包括对训练后的贝叶斯分类器进行验证。

[0130] 参见图 5，本发明实施例五提供的对训练后的贝叶斯分类器进行验证的方法具体包括：操作 51- 操作 53。

[0131] 操作 51 中，从所述原始网页中获取第三景点的类别信息和描述信息。

[0132] 从所述原始网页中获取第三景点的类别信息和描述信息，与上述实施例二中操作 23 的实施过程相似，这里不再赘述。

[0133] 操作 52 中，利用所述第三景点的类别信息和描述信息，对所述训练后的贝叶斯分类器进行验证。

[0134] 将第三景点作为验证集，对所述第三景点的描述信息分词，得到验证描述词；利用所述验证描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第三景点的所有验证描述词，列为所述第三景点的不同验证描述词。利用所述训练后的贝叶斯分类器根据所述向量空间模型对所述第三景点进行分类，验证分类器的性能。例如，将第三景点故宫的描述信息转化为向量空间模型的向量后放到这个分类器中去分类，若得到的结果是历史、休闲、家庭游，与其本身的类别信息相同，则说明分类器的准确率是 100％，召回率为 100％；若从分类器中得到的结果是亲子、蜜月、周边，与其本身的类别信息不相同，那么准确率为 0，召回率为 0。

[0135] 操作 53 中，验证通过后，触发所述利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类。

[0136] 对所述训练后的贝叶斯分类器进行验证的目的是验证训练的分类器是否可用，准确率召回率越高的分类器对第一景点的分类越准确。

[0137] 本实施例提供的对训练后的贝叶斯分类器进行验证的方法，通过利用第三景点的类别信息和描述信息对所述训练后的贝叶斯分类器进行验证，并在验证通过后，触发所述利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类，通过用验证通过的分类器对第一景点进行分类，进一步提高了所景点类别的准确性和客观性。

[0138] 实施例六

[0139] 本实施例在上述各实施例的基础上，提供了一种根据所述第一景点的描述信息确定所述第一景点的建议访问月份的方法。

[0140] 参见图 6，本实施例提供的确定建议访问月份的方法包括：操作 61- 操作 63。

[0141] 操作 61 中，根据所述第一景点的历史被访问时间，统计所述第一景点在不同的月份的历史被访问次数。

[0142] 例如，从各原始网页中提取出第一景点的历史被访问时间，然后统计出第一景点在不同月份被访问次数。例如：统计出故宫 1 月份被访问 300 次；2 月份被访问 300 次，…，12 月份被访问 50 次。

[0143] 操作 62 中，利用所述第一景点在不同的月份的历史被访问次数，得到所述第一景点在不同的月份的历史被访问的熵值。

[0144] 第一景点在不同的月份的历史被访问的熵值等于 p(x)log(p(x))，其中，p(x) 为所述第一景点在 x 月份的历史被访问的概率，所述第一景点在 x 月份的历史被访问的概率等于该月历史被访问次数除以 12 个月的次数之和。例如，故宫在 1 月份历史被访问概率的计算公式为：$P(1)=\frac{A(1)}{C}$，其中 P(1) 为故宫在 1 月份的历史被访问概率，A(1) 为故宫在 1 月份的历史被访问次数，C 为故宫 12 个月的历史被访问次数之和。得到每个月被访问的概率之后，计算该景点在每个月份的熵，例如故宫在 1 月份历史被访问熵为：p(1)log(p(1))。

[0145] 操作 63 中，根据所述第一景点在不同的月份的历史被访问的熵值确定所述第一景点的建议访问月份。

[0146] 例如，将某一景点各月份的熵相加，得到了这个景点的熵。由于熵反映一个事物的混乱程度，因此，如果这个值大于设定的阈值，（这里第一阀值设定为所有景点熵的加权平均值，热门景点的权值高）则认为该景点不具有区分性，认为四季皆宜；小于阈值则认为区分度明显。

[0147] 因此，当第一景点的熵小于阈值时，取所述第一景点在不同的月份的历史被访问概率中最大的两个月份作为所述建议访问月份，即作为该第一景点适宜游玩月份。仍以上述操作 62 中给出的故宫为例，假设其熵 $\sum_{i=1}^{12} p(i)\log(p(i))$ 小于上述阈值，且 p(9) 和 p(10) 最大，则将 9 月份和 10 月份作为故宫的建议访问月份。

[0148] 本发明实施例提供的确定建议访问月份的方法，通过在从所述原始网页中提取出第一景点的描述信息之后，根据所述第一景点的描述信息确定所述第一景点的建议访问月份，在获得上述各实施例的有益效果基础上节省了大量的人力物力，提供的第一景点的信息更加贴近用户的原始期望，方便用户直接参考。

[0149] 优选地，在根据所述第一景点在不同的月份的历史被访问的熵值确定所述第一景点的建议访问月份之后，还包括：对应存储所述第一景点的建议访问月份，这样设置的好处是：将所述第一景点的建议访问月份对应存储，能够进一步丰富景点知识库或景点信息库，以供线上查询。

[0150] 实施例七

[0151] 本实施例在上述各实施例的基础上，提供了一种根据所述第一景点的描述信息确定所述第一景点的建议访问天内时间的方法。

[0152] 参见图 7，本实施例提供的确定建议访问天内时间的方法包括：操作 71- 操作 73。

[0153] 操作 71 中，根据所述第一景点在景点序列中的位置及建议访问时长，统计所述第一景点分别在上午和下午的历史被访问次数。

[0154] 假设第一景点是故宫，从原始网页提取了相关的景点序列为：毛泽东纪念馆 - 故宫 - 鸟巢 - 水立方，可以得到故宫在景点序列中的位置为第 2 位。

[0155] 所述第一景点的建议访问时长可以从旅游攻略、游记网页中提取出，还可以从旅游网站为每个景点编辑的信息介绍的原始网页中直接提取。

[0156] 一般情况下，某一景点在景点序列中出现的位置越靠前，越适合在上午访问。但是，如果该景点访问的时间如果较长，则认为不适合上午访问。具体地，可以结合建议访问时长以及第一景点在序列中出现的位置，来统计所述第一景点分别在上午和下午的历史被访问次数。

[0157] 例如，一个景点序列：故宫 - 天安门 - 王府井 - 水立方中，故宫排在景点序列的第一位，但是故宫的建议访问时长是 6 个小时，那么故宫在该景点序列中的访问时间就延续到了中午，说明故宫并不是一定要上午访问，因此，将故宫在该景点序列中的上午访问不做统计，即不算入故宫的上午历史被访问次数中。另一个景点序列：天安门 - 故宫 - 王府井 - 水立方中，故宫排在景点序列的第二位，建议访问时长是 2 小时，则将故宫在该景点序列中的上午访问计入故宫的上午历史被访问次数中。这里，结合建议访问时长确定统计所

述第一景点分别在上午和下午的历史被访问次数可以放置噪声引入，目的是提高统计的准确性。

[0158] 操作 72 中，根据所述第一景点分别在上午和下午的历史被访问次数，确定所述第一景点的上午访问指数和下午访问指数。

[0159] 其中，所述第一景点上午访问指数是指所述第一景点的上午历史被访问次数除以下午的历史被访问次数，所述第一景点下午访问指数是指所述第一景点的下午历史被访问次数除以上午的历史被访问次数。

[0160] 例如：后海这个景点，在统计的景点序列中后海下午的历史被访问次数是 5，上午的历史被访问次数是 1，那么后海的上午指数是 1 除以 5 等于 0.2；后海的下午指数是 5 除以 1 等于 5。

[0161] 操作 73 中，将确定的上午访问指数和下午访问指数中值最大的访问指数对应的时段，作为所述建议访问天内时间。

[0162] 还是以操作 72 中后海的例子说明，后海的上午指数是 0.2；后海的下午指数是 5，所以后海上午访问指数和下午访问指数中值最大的访问指数对应的时段是下午，将下午作为后海的建议访问天内时间。

[0163] 显然，还可以直接将所述第一景点的上午和下午的历史被访问次数分别作为所述第一景点的上午访问指数和下午访问指数，将上午和下午的历史被访问次数中次数最大的访问时段，作为所述建议访问天内时间。

[0164] 本发明实施例通过在从所述原始网页中提取出第一景点的描述信息之后，根据所述第一景点的描述信息确定所述第一景点的建议访问天内时间，在获得上述各实施例的有益效果基础上节省了大量的人力物力，提供的第一景点的信息更加贴近用户的原始期望，方便用户直接参考。

[0165] 为了增加准确性，本实施例涉及的景点序列包含的景点数量应大于或等于 3。

[0166] 当所述第一景点在一个所述景点序列中排在第一位或第二位，且所述建议访问时长小于预设值时，则将所述第一景点在上午的历史被访问次数加 1；

[0167] 当所述第一景点在一个所述景点序列中排在倒数第一位或倒数第二位时，则将所述第一景点在下午的历史被访问次数加 1。

[0168] 进一步地，在确定所述第一景点的建议访问天内时间时，需获得该景点的上午访问指数和下午访问指数。优选地，设定第三阈值，当所述第一景点的上午访问指数大于所述第一景点的下午访问指数，且大于所述第三阈值时，则所述第一景点的建议访问天内时间只是上午；否则，所述第一景点的建议访问天内时间只是下午，这样设置的好处是，可以用于后续推荐景点的游玩顺序。所述第三阈值的设定，可以根据手动选出的一些上下午访问区分度明显的景点，计算出的景点的上午访问指数平均值和景点的下午访问指数平均值得到。

[0169] 优选地，在根据所述第一景点的描述信息确定所述第一景点的建议访问天内时间之后，还包括：对应存储所述第一景点的建议访问天内时间，这样设置的好处是：将所述第一景点的建议访问天内时间对应存储，能够进一步丰富景点知识库或景点信息库，以供线上查询。

[0170] 实施例八

[0171] 本实施例提供了另一种确定景点的类别的方法。

[0172] 参见图 8，本发明实施例八提供的确定所景点的类别的方法包括：操作 81- 操作 88。

[0173] 操作 81 中，获取标注有特色标签的景点序列。

[0174] 这里特色标签即类别。

[0175] 操作 82 中，获取景点的特色标签。

[0176] 具体地，从操作 81 获取的景点序列中获取一部分景点的特色标签，将该部分景点作为朴素贝叶斯分类器的训练集。训练集就是已经标注好标签的一系列的景点，例如故宫被标上历史、休闲和家庭游。

[0177] 训练集中的每个景点取标注次数最多的前三个标签。

[0178] 操作 83 中，建立向量空间模型。

[0179] 具体地，将操作 82 中获取的景点的描述信息进行分词，建立向量空间模型。

[0180] 操作 84 中，利用 NMF 提取景点的前 30 维特征。

[0181] 具体地，通过 NMF 算法对向量空间模型进行过滤，对每个景点取其前 30 维最重要的训练描述词。

[0182] 操作 85 中，训练朴素贝叶斯分类器。

[0183] 具体地，利用过滤后的向量空间模型训练朴素贝叶斯分类器。

[0184] 操作 86 中，验证训练后的朴素贝叶斯分类器

[0185] 具体地，利用操作 82 中获得的部分景点的特色标签对训练后的朴素贝叶斯分类器进行验证。

[0186] 例如，把上述训练集按照 9 比 1 分为训练集、测试集。用 9 成景点的数据训练好这个分类器之后，用 1 成的已标注好的景点去验证这个分类器的性能。假设故宫这个景点在这 1 成的测试集里，可以把故宫的描述信息转化为向量空间模型的向量后放到这个贝叶斯分类器中去分类。如果得到的结果是历史、休闲、家庭游，那么，说明分类器的准确率是 100%，召回率 100%；如果这个景点从分类器中得到的结果是亲子、蜜月、周边，那么说明分类器的准确率 0，召回率 0。

[0187] 验证准确率和召回率的目的是验证一下训练的分类器是否可用，只有准确率和召回率高的分类器才可用于对未标注标签的景点进行分类。

[0188] 操作 87 中，建立未标注景点的向量空间模型。

[0189] 具体地，对未标注标签的景点的描述信息进行分词，建立向量空间模型。

[0190] 操作 88 中，对未标注标签的景点进行分类。

[0191] 具体地，使用上述操作 86 中验证后的朴素贝叶斯分类器对操作 87 中的向量空间模型进行分类，即对未标注标签的景点标注标签。

[0192] 以故宫为例，假设利用分类器得到的概率值是 p（类别 1| 故宫）、p（类别 2| 故宫）等等，找出其中值最大的 3 个概率对应的类别作为故宫的标签对故宫进行标注。

[0193] 实施例九

[0194] 本实施例提供了一种确定景点的建议访问月份的方法。

[0195] 参见图 9，本发明实施例九提供的确定景点的建议访问月份的方法包括：操作 91- 操作 95。

[0196] 操作 91 中，提取景点序列及相应的出行时间。

[0197] 具体地，从格式化游记 1、格式化游记 2、…、格式化游记 n 中提取出的景点序列，以及游记提及的游览时间，对每一个景点分别统计出行在不同月份的次数，例如：故宫：1 月，300 次；2 月，300 次，…12 月，50 次。

[0198] 操作 92 中，确定景点在各月份出现的概率。

[0199] 具体地，使用某一景点在某月的出现次数除以在 12 个月出现的次数之和，即得到该景点在该月份出现的概率。

[0200] 操作 93 中，求出每个景点的熵。

[0201] 具体地，利用某一景点在某个月的概率之后，就可以得到这个景点在该月份的熵，将各月份的熵相加，即得到该景点的熵。

[0202] 操作 94 中，判断各景点的熵是否小于阈值。

[0203] 当小于阈值时，执行操作 95，否则，认为该景点四季皆可访问，不再做处理，结束流程。

[0204] 操作 95 中，标注景点适宜游玩的月份。

[0205] 具体地，将操作 92 得到的值最大的两个概率对应的月份作该景点的建议游玩月份。

[0206] 此外，可以将操作 91 中提取的景点序列中的部分景点作为测试样本标注其适宜游玩的月份，然后与上述操作 95 得到的结果进行比较，以验证经过上述操作标注景点的建议访问月份的合理性及准确性。

[0207] 实施例十

[0208] 本实施例提供了一种确定建议访问天内时间的方法。

[0209] 参见图 10，本发明实施例十提供的确定建议访问天内时间的方法具体包括：操作 101- 操作 105。

[0210] 操作 101 中，提取景点序列。

[0211] 具体地，可以从格式化游记 1、格式化游记 2、…、格式化游记 n 中提取出景点序列。

[0212] 操作 102 中，判断景点序列的长度是否大于 3，即判断景点序列中景点的数量是否大于 3。若是，则执行操作 103；否则舍弃该景点序列。

[0213] 操作 103 中，统计各景点上下午出现次数。

[0214] 具体地，统计出长度大于 3 的景点序列中各景点在上下午的出现次数。

[0215] 操作 104 中，求出各景点的上下午访问指数。

[0216] 具体地，用操作 103 得到的某一景点的上午游玩次数除以下午游玩次数，得到该景点的上午访问指数，用该景点的下午游玩次数除以上午游玩次数，得到该景点的下午访问指数。

[0217] 操作 105 中，判断上下午访问指数是否大于阈值。当上午指数大于这个阈值之后我们只认为该景点适合上午访问，同理某些景点只适于下午访问。否则，结束流程。

[0218] 其中，阈值可以由景点的上午访问指数或下午访问指数得到。具体地，

[0219] 可以从操作 101 提取的景点序列中选出的一些上下午访问区分度明显的景点，计算出的景点的上午访问指数平均值和景点的下午访问指数平均值，上午访问指数平均值是用于与上午访问指数比较的阈值，下午访问指数平均值是用于与下午访问指数比较的阈

值。

[0220] 实施例十一

[0221] 参见图 11，本实施例提供的一种网页信息处理装置具体包括：

[0222] 网页获取模块 111，用于获取原始网页；

[0223] 信息提取模块 112，用于从所述原始网页中提取出第一景点的描述信息；

[0224] 类别确定模块 113，用于根据所述第一景点的描述信息确定所述第一景点的类别。

[0225] 本实施例提供的网页信息处理装置，通过网页获取模块获取原始网页，并从所述原始网页中提取出第一景点的描述信息，根据所述第一景点的描述信息确定所述第一景点的类别，解决了现有技术中旅游网站提供的景点信息不准确的问题，提高了景点信息的准确性。

[0226] 示例性的，上述网页信息处理装置还包括：

[0227] 第一信息获取模块，用于在所述类别确定模块根据所述第一景点的描述信息确定所述第一景点的类别之前，从所述原始网页中获取第二景点的类别信息和描述信息。

[0228] 示例性的，上述第一信息获取模块具体用于：

[0229] 从所述原始网页中获取包含有所述第二景点的旅游路线信息，所述旅游路线信息包括旅游路线及其标签；

[0230] 统计所述第二景点出现在标注有标签的旅游路线中的次数；

[0231] 根据统计的次数，将第一标签、第二标签和第三标签作为所述第二景点的类别，其中，所述第二景点出现在标注有所述第一标签的旅游线路中的次数最多，出现在标注有所述第二标签的旅游线路中的次数仅次于标注有所述第一标签的旅游线路，出现在标注有所述第三标签的旅游线路中的次数仅次于标注有所述第一标签和第二标签的旅游线路。

[0232] 示例性的，上述类别确定模块具体用于：

[0233] 根据所述第二景点的类别信息和描述信息以及所述第一景点的描述信息，确定所述第一景点的类别。

[0234] 示例性的，上述类别确定模块包括：

[0235] 第一训练子模块，用于利用所述第二景点的类别信息和描述信息训练贝叶斯分类器；

[0236] 第一分类子模块，用于利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类。

[0237] 示例性的，上述第一训练子模块包括：

[0238] 第一分词子模块，用于对所述第二景点的描述信息分词，得到训练描述词；

[0239] 第一模型建立子模块，用于利用所述训练描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第二景点的所有训练描述词，列为所述第二景点的不同训练描述词；

[0240] 第二训练子模块，用于利用所述向量空间模型训练贝叶斯分类器。

[0241] 示例性的，上述第一模型建立子模块具体用于：

[0242] 根据词频-逆向文本频率 tf-idf 算法将所述训练描述词去除一半；利用剩余的训练描述词建立所述向量空间模型。

[0243] 示例性的，上述网页信息处理装置还包括：

[0244] 第二信息获取模块，用于在所述第一分类子模块利用所述第一训练子模块训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类之前，从所述原始网页中获取第三景点的类别信息和描述信息；

[0245] 验证模块，用于利用所述第三景点的类别信息和描述信息，对所述训练后的贝叶斯分类器进行验证；

[0246] 触发模块，用于在所述验证模块对所述训练后的贝叶斯分类器的验证通过后，触发所述第一分类子模块利用所述训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类。

[0247] 示例性的，上述分类子模块包括：

[0248] 第二分词子模块，用于对所述第一景点的描述信息分词，得到分类描述词；

[0249] 第二模型建立子模块，用于利用所述分类描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第一景点的所有分类描述词，列为所述第一景点的不同分类描述词；

[0250] 第二分类子模块，用于利用所述训练后的贝叶斯分类器根据所述向量空间模型对所述第一景点进行分类。

[0251] 示例性的，上述第一分类子模块具体用于：

[0252] 将所述训练后的贝叶斯分类器根据所述第一景点的描述信息，得出的第一类别、第二类别和第三类别作为所述第一景点的类别，其中，所述第一类别的后验概率值最大，所述第二类别的后验概率值仅次于所述第一类别，所述第三类别的后验概率值仅次于所述第一类别和第二类别。

[0253] 示例性的，上述网页信息处理装置还包括：

[0254] 时间确定模块，用于在所述信息提取模块从所述原始网页中提取出第一景点的描述信息之后，根据所述第一景点的描述信息确定所述第一景点的建议访问时间。

[0255] 示例性的，上述时间确定模块具体用于：根据所述第一景点的描述信息确定所述第一景点的建议访问月份和建议访问天内时间中的至少一项信息，其中所述建议访问天内时间包括上午和下午中的至少一个时段。

[0256] 示例性的，上述时间确定模块具体用于：

[0257] 根据所述第一景点的历史被访问时间，统计所述第一景点在不同的月份的历史被访问次数；

[0258] 利用所述第一景点在不同的月份的历史被访问次数，得到所述第一景点在不同的月份的历史被访问的熵值；根据所述第一景点在不同的月份的历史被访问的熵值确定所述第一景点的建议访问月份。

[0259] 示例性的，上述时间确定模块具体用于：

[0260] 当所述第一景点在不同的月份的历史被访问的熵值之和小于阈值，将所述第一景点在不同的月份的历史被访问概率中最大的两个月份作为所述建议访问月份。

[0261] 示例性的，上述时间确定模块具体用于：

[0262] 根据所述第一景点在景点序列中的位置及建议访问时长，统计所述第一景点分别在上午和下午的历史被访问次数；

[0263] 根据所述第一景点分别在上午和下午的历史被访问次数，确定所述第一景点的上

午访问指数和下午访问指数；

[0264] 将确定的上午访问指数和下午访问指数中值最大的访问指数对应的时段，作为所述建议访问天内时间。

[0265] 示例性的，上述时间确定模块具体用于：

[0266] 当所述第一景点在一个所述景点序列中排在第一位或第二位，且所述建议访问时长小于预设值时，则将所述第一景点在上午的历史被访问次数加 1；

[0267] 当所述第一景点在一个所述景点序列中排在倒数第一位或倒数第二位时，则将所述第一景点在下午的历史被访问次数加 1。

[0268] 示例性的，上述网页信息处理装置还包括：

[0269] 存储模块，用于对应存储所述第一景点的类别和描述信息，所述第一景点的建议范文月份信息，所述第一景点的建议访问天内时间信息。

[0270] 上述网页信息处理装置可执行本发明任意实施例所提供的网页信息处理方法，具备与网页信息处理方法中各操作相对应的功能模块和有益效果。

[0271] 注意，上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解，本发明不限于这里所述的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此，虽然通过以上实施例对本发明进行了较为详细的说明，但是本发明不仅仅限于以上实施例，在不脱离本发明构思的情况下，还可以包括更多其他等效实施例，而本发明的范围由所附的权利要求范围决定。

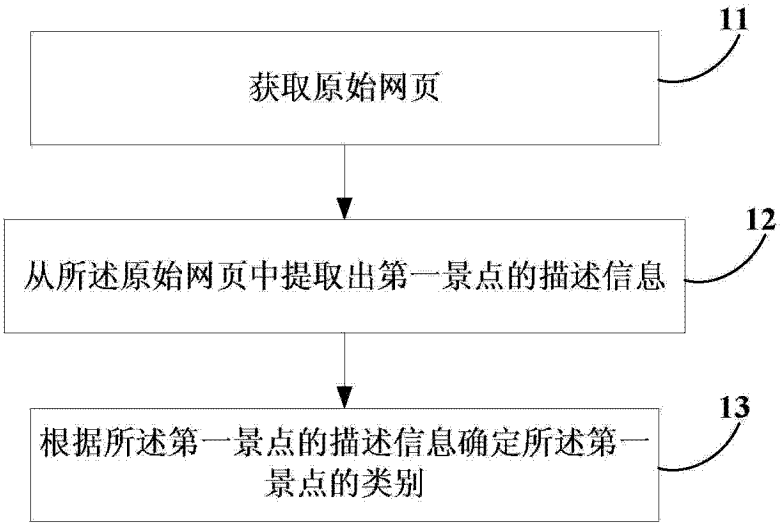

11
获取原始网页
12
从所述原始网页中提取出第一景点的描述信息
13
根据所述第一景点的描述信息确定所述第一景点的类别

图 1

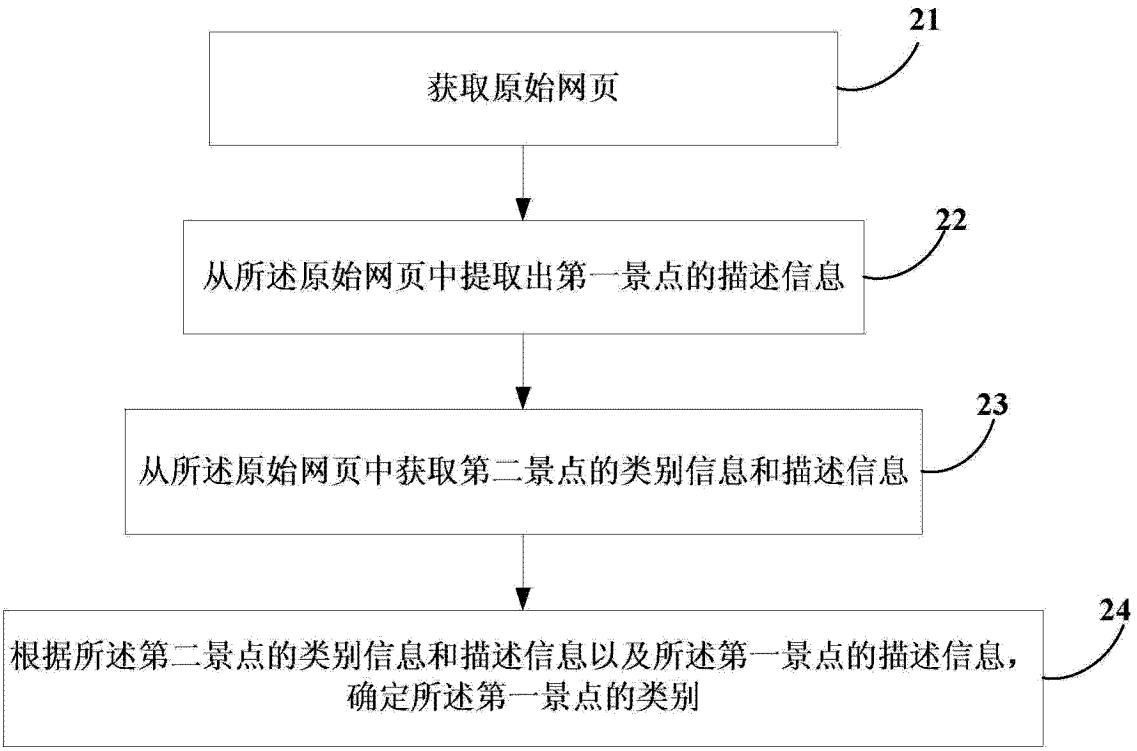

21
获取原始网页
22
从所述原始网页中提取出第一景点的描述信息
23
从所述原始网页中获取第二景点的类别信息和描述信息
24
根据所述第二景点的类别信息和描述信息以及所述第一景点的描述信息，确定所述第一景点的类别

图 2

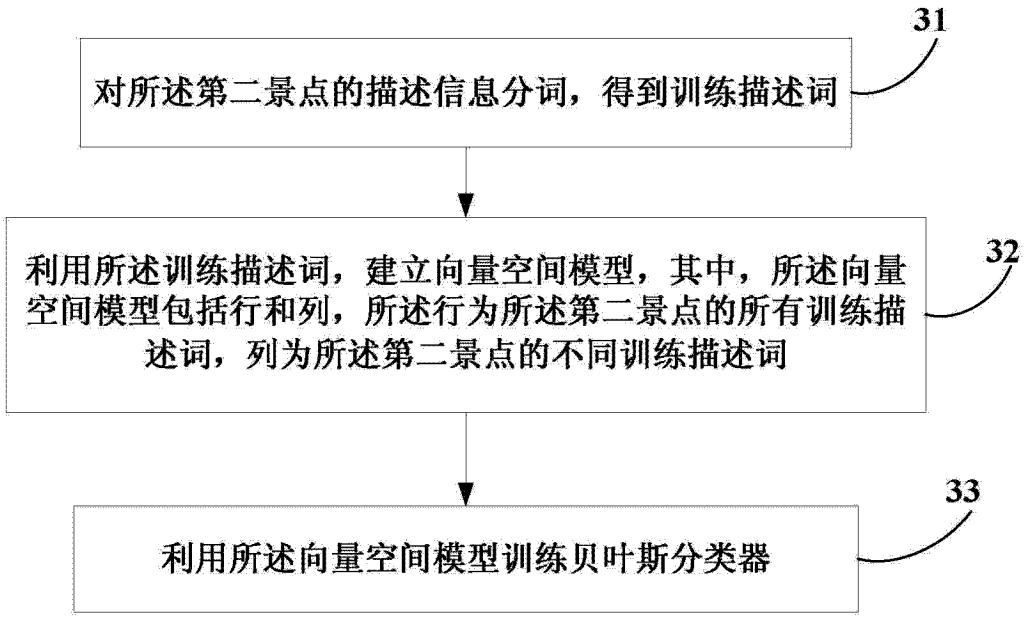

31
对所述第二景点的描述信息分词，得到训练描述词
32
利用所述训练描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第二景点的所有训练描述词，列为所述第二景点的不同训练描述词
33
利用所述向量空间模型训练贝叶斯分类器

图 3

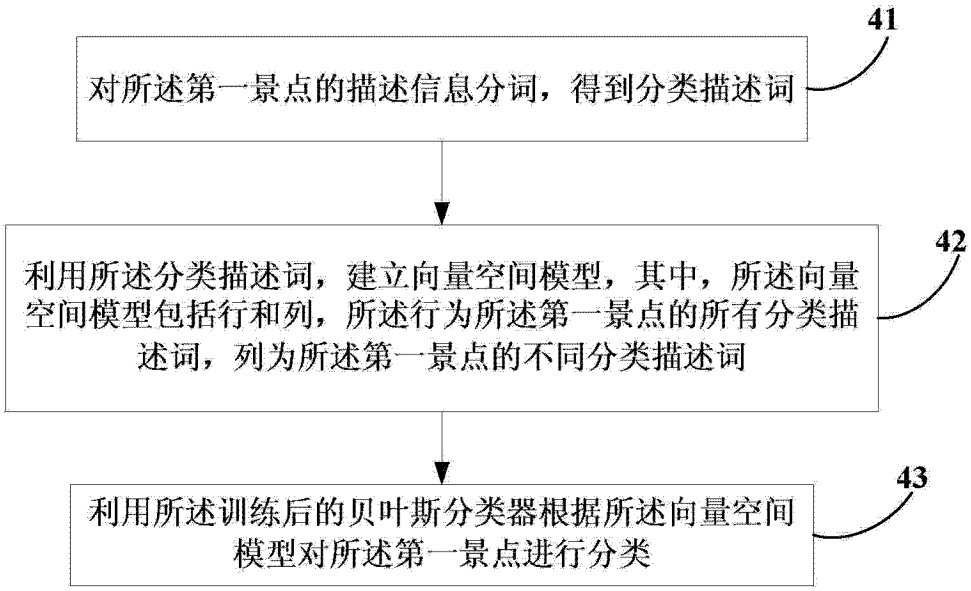

41
对所述第一景点的描述信息分词，得到分类描述词
42
利用所述分类描述词，建立向量空间模型，其中，所述向量空间模型包括行和列，所述行为所述第一景点的所有分类描述词，列为所述第一景点的不同分类描述词
43
利用所述训练后的贝叶斯分类器根据所述向量空间模型对所述第一景点进行分类

图 4

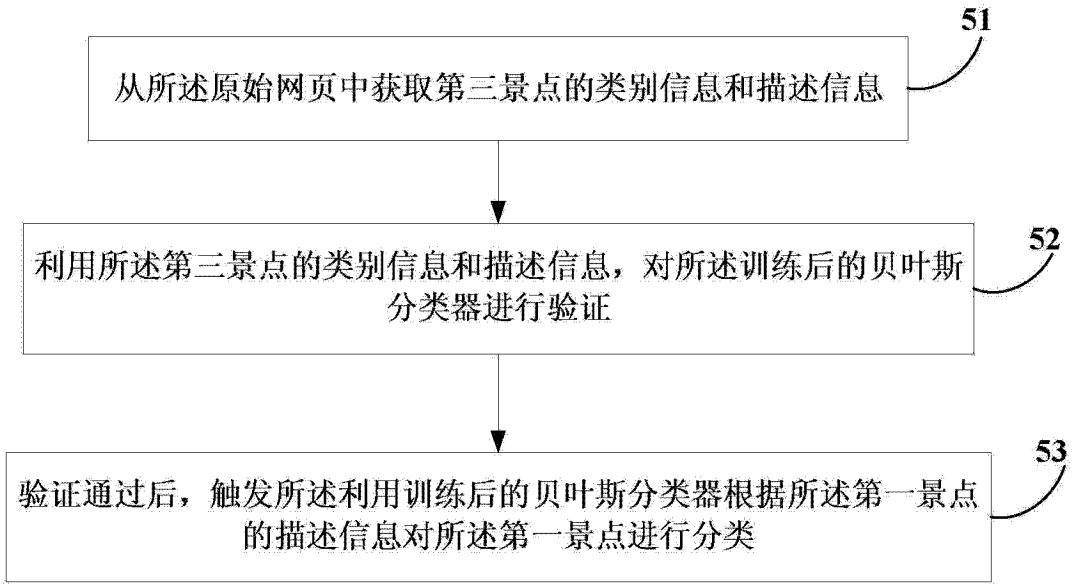

51
从所述原始网页中获取第三景点的类别信息和描述信息
52
利用所述第三景点的类别信息和描述信息，对所述训练后的贝叶斯分类器进行验证
53
验证通过后，触发所述利用训练后的贝叶斯分类器根据所述第一景点的描述信息对所述第一景点进行分类

图 5

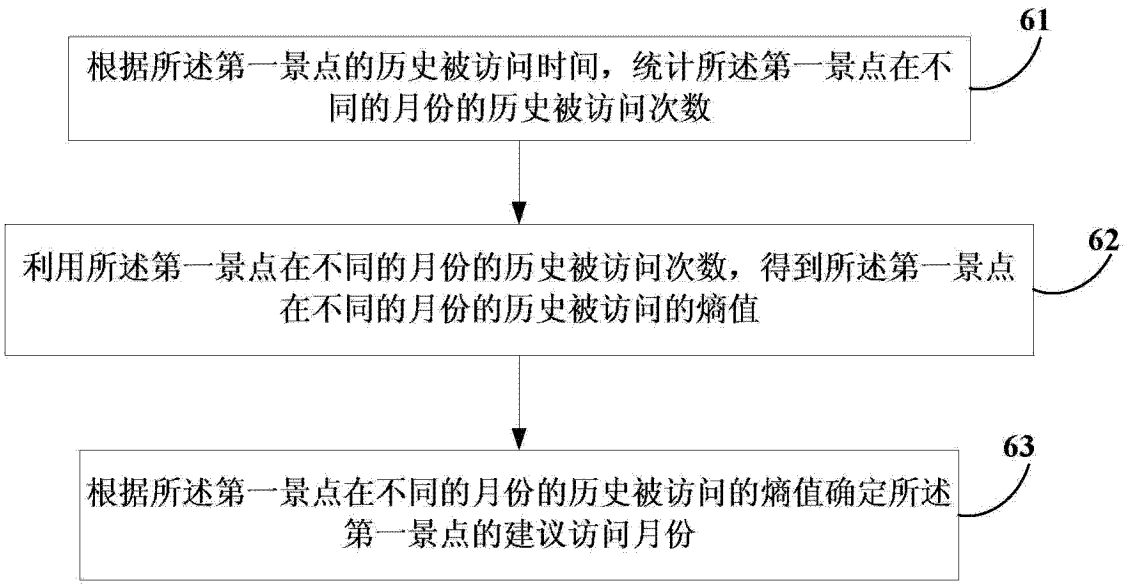

61
根据所述第一景点的历史被访问时间，统计所述第一景点在不同的月份的历史被访问次数
62
利用所述第一景点在不同的月份的历史被访问次数，得到所述第一景点在不同的月份的历史被访问的熵值
63
根据所述第一景点在不同的月份的历史被访问的熵值确定所述第一景点的建议访问月份

图 6

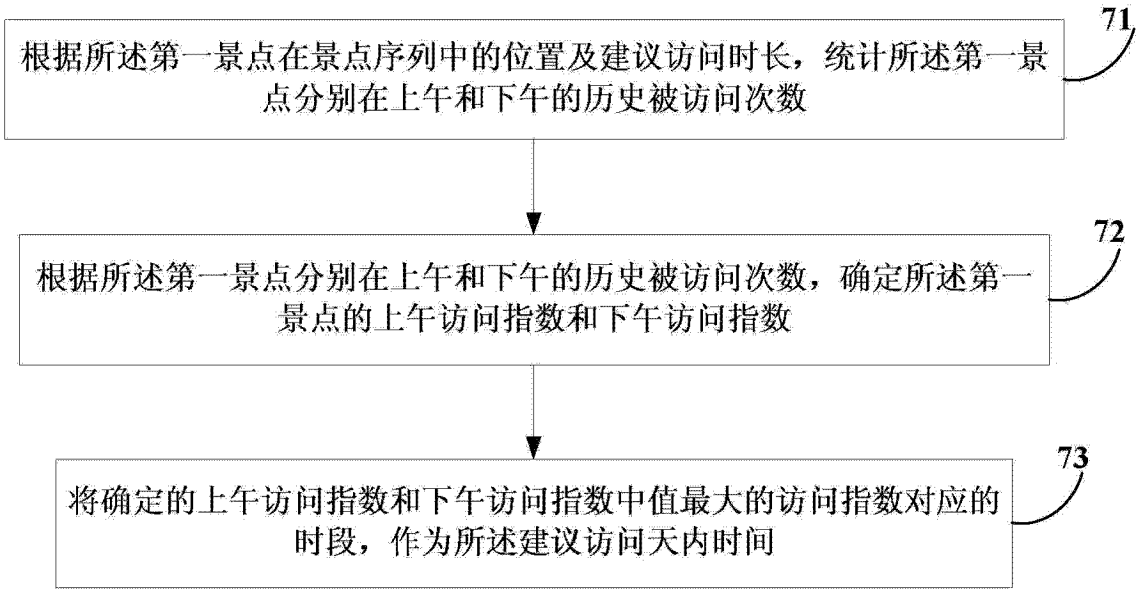

根据所述第一景点在景点序列中的位置及建议访问时长，统计所述第一景点分别在上午和下午的历史被访问次数
71
根据所述第一景点分别在上午和下午的历史被访问次数，确定所述第一景点的上午访问指数和下午访问指数
72
将确定的上午访问指数和下午访问指数中值最大的访问指数对应的时段，作为所述建议访问天内时间
73

图 7

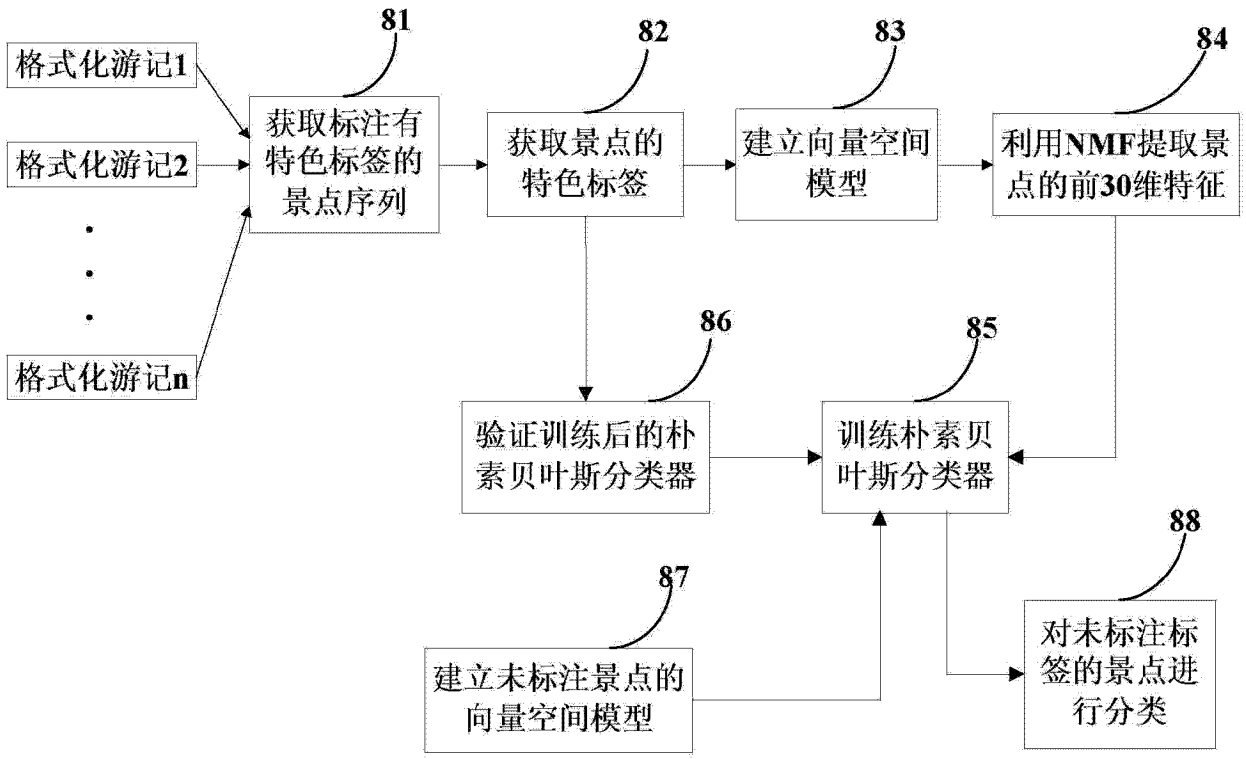

格式化游记1
格式化游记2
格式化游记n
81
获取标注有特色标签的景点序列
82
获取景点的特色标签
83
建立向量空间模型
84
利用NMF提取景点的前30维特征
86
验证训练后的朴素贝叶斯分类器
85
训练朴素贝叶斯分类器
87
建立未标注景点的向量空间模型
88
对未标注标签的景点进行分类

图 8

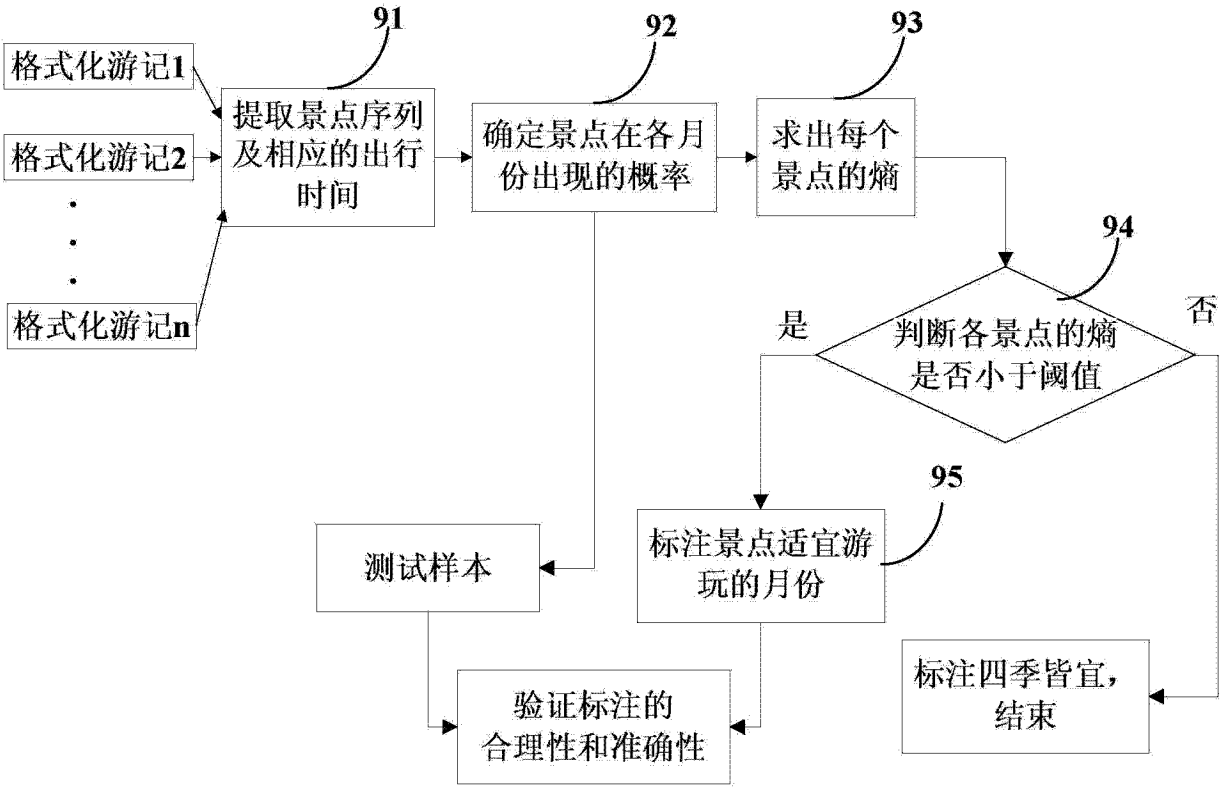

格式化游记1
格式化游记2
格式化游记n
91
提取景点序列及相应的出行时间
92
确定景点在各月份出现的概率
93
求出每个景点的熵
94
判断各景点的熵是否小于阈值
是
否
95
标注景点适宜游玩的月份
测试样本
验证标注的合理性和准确性
标注四季皆宜，结束

图 9

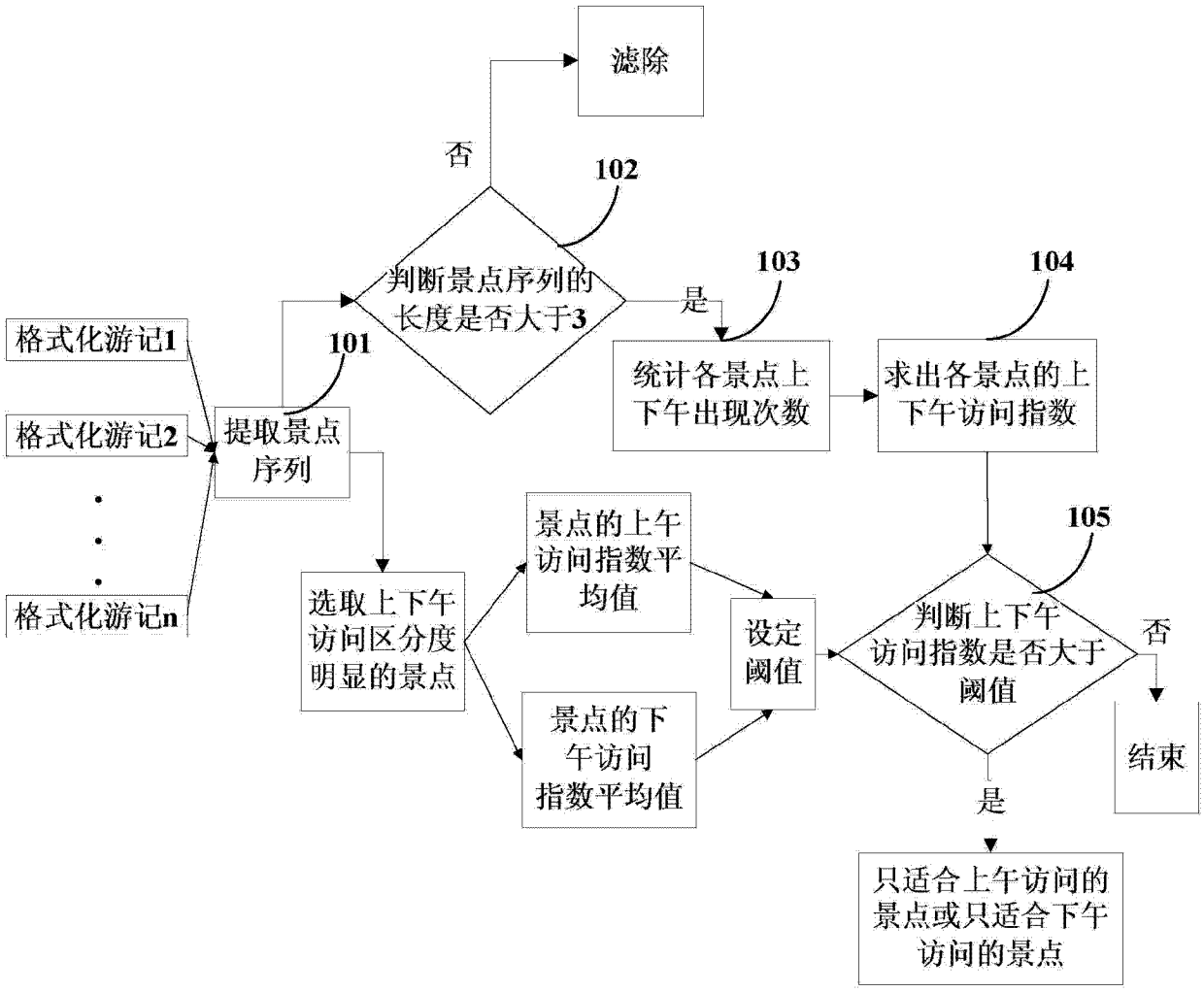

滤除
否
102
判断景点序列的长度是否大于3
是
103
104
格式化游记1
101
格式化游记2
提取景点序列
统计各景点上下午出现次数
求出各景点的上下午访问指数
格式化游记n
景点的上午访问指数平均值
105
选取上下午访问区分度明显的景点
设定阈值
判断上下午访问指数是否大于阈值
否
结束
景点的下午访问指数平均值
是
只适合上午访问的景点或只适合下午访问的景点

图 10

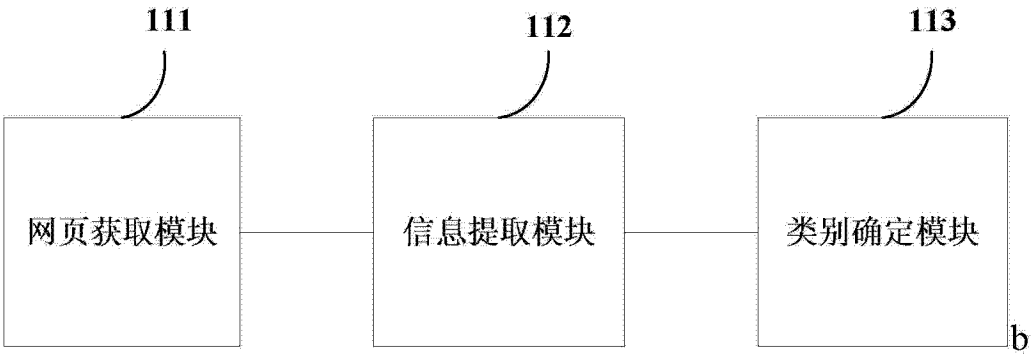

111
112
113
网页获取模块
信息提取模块
类别确定模块
b

图 11