(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 102521267 B
(45)授权公告日 2014.01.22

(21)申请号 201110372458.6

(22)申请日 2011.11.21

(73)专利权人 沈文策
地址 350003 福建省福州市鼓楼区软件园 A 区 25 号中金在线大厦

(72)发明人 沈文策

(74)专利代理机构 北京超凡志成知识产权代理事务所（普通合伙） 11371
代理人 李世喆

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

(56)对比文件
CN 102236719 A,2011.11.09,
CN 101763391 A,2010.06.30,
CN 1845104 A,2006.10.11,

审查员 王静

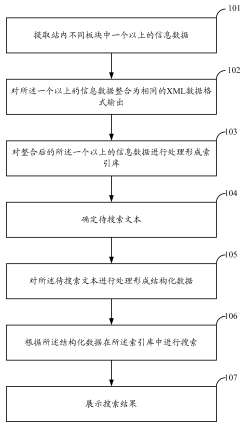
权利要求书3页 说明书9页 附图4页

(54)发明名称

站内信息搜索方法和搜索系统

(57)摘要

本发明涉及信息技术领域，具体为一种站内信息搜索方法和搜索系统。所述搜索方法包括：提取站内不同板块中一个以上的信息数据；对所述一个以上的信息数据整合为相同的 XML 数据格式输出；对整合后的所述一个以上的信息数据进行处理形成索引库；确定待搜索文本；对所述待搜索文本进行处理形成结构化数据；根据所述结构化数据在所述索引库中进行搜索；展示搜索结果。本发明能够加快站内信息搜索的速度。



CN 102521267 B

1. 一种站内信息搜索方法，其特征在于，包括：

提取站内不同板块中一个以上的信息数据；

对所述一个以上的信息数据整合为相同的 XML 数据格式输出；

对整合后的所述一个以上的信息数据进行处理形成索引库；

确定待搜索文本；

对所述待搜索文本进行处理形成结构化数据；

根据所述结构化数据在所述索引库中进行搜索；

展示搜索结果；

所述对所述待搜索文本进行处理形成结构化数据包括：

A. 载入与所述待搜索文本相匹配的词典；

B. 按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的一个待匹配子文本；

C. 按随机或设定的方式将所述待匹配子文本与所述词典进行匹配；

D. 当匹配成功时，记录所述待匹配子文本中匹配成功的待匹配关键词，并且判断所述待匹配文本是否存在未被抽取的文字；

当判断结果为存在时，再按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的下一个待匹配子文本，返回 C 步骤；

当判断结果为不存在时，进行 F 步骤；

E. 当匹配失败时，从所述待匹配子文本的任一端删减一个以上的文字，形成新的所述待匹配子文本，所述待匹配子文本进一步包括删减文字后的文本，返回 C 步骤；

F. 将所有所述待匹配关键词构成所述结构化数据。

2. 如权利要求 1 所述的站内信息搜索方法，其特征在于，所述对整合后的所述一个以上的信息数据进行处理形成索引库包括：

对所述一个以上的信息数据进行信息数据编号；

提取每一个所述信息数据中包含的一个以上的关键词；

对所述一个以上的关键词进行关键词编号；

记录每一个关键词在包含所述关键词的所述信息数据中的关键词位置；

记录每一个关键词在包含所述关键词的所述信息数据中出现的关键词次数；

根据所述信息数据编号、关键词编号、关键词位置、关键词次数之间的对应关系建立包含所述对应关系的所述索引库。

3. 如权利要求 1-2 任一项所述的站内信息搜索方法，其特征在于，所述对整合后的所述一个以上的信息数据进行处理形成索引库进一步包括：提取每一个所述信息数据中包含的一个以上的关键词；将所述关键词汇总成关键词词典；则，所述待搜索文本相匹配的词典为所述关键词词典或外部引入的词典。

4. 如权利要求 1 或 2 所述的站内信息搜索方法，其特征在于，

所述将所有所述待匹配关键词构成所述结构化数据包括：

统计所有所述待匹配关键词的数量；

统计所有重复的待匹配关键词及每一个重复的所述待匹配关键词的数量；

剔除所有重复的待匹配关键词；

将剔除重复的所述待匹配关键词后的其他所述待匹配关键词构成所述结构化数据。

5. 如权利要求 1-2 任一项所述的站内信息搜索方法，其特征在于，所述根据所述结构化数据在所述索引库中进行搜索包括：

将所述结构化数据中每一个待匹配关键词与所述索引库中的关键词进行匹配，并根据匹配结果确定搜索结果。

6. 如权利要求 5 所述的站内信息搜索方法，其特征在于，所述将所述结构化数据中每一个待匹配关键词与所述索引库中的关键词进行匹配包括：

a. 以随机或设定顺序选取一个所述待匹配关键词，将选定的所述待匹配关键词与所述索引库中的关键词进行匹配，并标定该选取的所述待匹配关键词为已选取；

b. 当匹配成功时，记录包含所述关键词的所述信息数据，并将关键词匹配次数加 1，将每一条包含所述关键词的所述信息数据的匹配次数加 1，设定匹配次数的初始值为 0；

c. 当匹配失败时，判断所述待匹配关键词中是否还有未标定为已选取的待匹配关键词，如果判断结果为有，则返回 a 步骤，如果判断结果为没有，则匹配结束；

则，

所述根据匹配结果确定搜索结果包括：

按照信息数据的匹配次数由大到小的顺序，选取设定数目的所述信息数据作为搜索结果。

7. 一种站内信息搜索系统，其特征在于，包括：

信息数据提取模块，用于提取站内不同板块中一个以上的信息数据；

格式输出模块，用于对所述一个以上的信息数据整合为相同的 XML 数据格式输出；

索引库形成模块，用于整合后的所述一个以上的信息数据进行处理形成索引库；

待搜索文本确定模块，用于确定待搜索文本；

结构化数据形成模块，用于对所述待搜索文本进行处理形成结构化数据；

搜索模块，用于根据所述结构化数据在所述索引库中进行搜索；

搜索结果展示模块，用于展示搜索结果；

所述结构化数据形成模块包括：词典载入子模块、待匹配子文本抽取子模块、匹配子模块、匹配成功子模块、匹配失败子模块、结构化数据构建子模块

所述词典载入子模块，用于载入与所述待搜索文本相匹配的词典；

所述待匹配子文本抽取子模块，用于按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的一个待匹配子文本；

所述匹配子模块，用于按随机或设定的方式将所述待匹配子文本与所述词典进行匹配；

所述匹配成功子模块，用于当匹配成功时，记录所述待匹配子文本中匹配成功的待匹配关键词，并且判断所述待匹配文本是否存在未被抽取的文字；当判断结果为存在时，再按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的下一个待匹配子文本，并将抽取的所述下一个待匹配子文本发送到所述匹配子模块；当判断结果为不存在时，将匹配成功的所有所述待匹配关键词发送给结构化数据构建子模块；

所述匹配失败子模块，用于当匹配失败时，从所述待匹配子文本的任一端删减一个以上的文字，形成新的所述待匹配子文本，所述待匹配子文本进一步包括删减文字后的文本，

并将删减过文字的待匹配子文本发送给所述匹配子模块；

所述结构化数据构建子模块，用于将所有所述待匹配关键词构成所述结构化数据。

8. 如权利要求 7 所述的站内信息搜索系统，其特征在于，所述索引库形成模块包括：

信息数据编号子模块，用于对所述一个以上的信息数据进行信息数据编号；

提取关键词子模块，用于提取每一个所述信息数据中包含的一个以上的关键词；

关键词编号子模块，用于对所述一个以上的关键词进行关键词编号；

关键词位置子模块，用于记录每一个关键词在包含所述关键词的所述信息数据中的关键词位置；

关键词次数子模块，用于记录每一个关键词在包含所述关键词的所述信息数据中出现的关键词次数；

对应关系子模块，用于根据所述信息数据编号、关键词编号、关键词位置、关键词次数之间的对应关系建立包含所述对应关系的所述索引库。

站内信息搜索方法和搜索系统

技术领域

[0001] 本发明涉及信息技术领域，具体为一种站内信息搜索方法和搜索系统。

背景技术

[0002] 目前，网络进入了千家万户，人们越来越喜欢在网上进行各种活动，网站也针对人们不同的需要开设了不同的版块，比如：新闻、体育、教育、出国、社区等等，人们在一家网站既可以得到多个版块的信息内容而且可以在多个版块进行互动性交流，而且，人们为了查找一些内容，越来越多的用到站内搜索，目前，站内搜索的局限性很大，由于版块众多，信息繁杂，站内搜索速度很慢，基本无法满足人们快节奏生活下的快节奏站内搜索的需要。

发明内容

[0003] 本发明提供了一种站内信息搜索方法和搜索系统，能够加快站内信息搜索的速度。

[0004] 本发明提供了一种站内信息搜索方法，包括：

[0005] 提取站内不同板块中一个以上的信息数据；

[0006] 对所述一个以上的信息数据整合为相同的 XML 数据格式输出；

[0007] 对整合后的所述一个以上的信息数据进行处理形成索引库；

[0008] 确定待搜索文本；

[0009] 对所述待搜索文本进行处理形成结构化数据；

[0010] 根据所述结构化数据在所述索引库中进行搜索；

[0011] 展示搜索结果。

[0012] 所述对整合后的所述一个以上的信息数据进行处理形成索引库优选为包括：

[0013] 对所述一个以上的信息数据进行信息数据编号；

[0014] 提取每一个所述信息数据中包含的一个以上的关键词；

[0015] 对所述一个以上的关键词进行关键词编号；

[0016] 记录每一个关键词在包含所述关键词的所述信息数据中的关键词位置；

[0017] 记录每一个关键词在包含所述关键词的所述信息数据中出现的关键词次数；

[0018] 根据所述信息数据编号、关键词编号、关键词位置、关键词次数之间的对应关系建立包含所述对应关系的所述索引库。

[0019] 所述对所述待搜索文本进行处理形成结构化数据优选为包括：

[0020] A. 载入与所述待搜索文本相匹配的词典；

[0021] B. 按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的一个待匹配子文本；

[0022] C. 按随机或设定的方式将所述待匹配子文本与所述词典进行匹配；

[0023] D. 当匹配成功时，记录所述待匹配子文本中匹配成功的待匹配关键词，并且判断所述待匹配文本是否存在未被抽取的文字；

[0024] 当判断结果为存在时，再按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的下一个待匹配子文本，返回 C 步骤；

[0025] 当判断结果为不存在时，进行 F 步骤；

[0026] E. 当匹配失败时，从所述待匹配子文本的任一端删减一个以上的文字，形成新的所述待匹配子文本，所述待匹配子文本进一步包括删减文字后的文本，返回 C 步骤；

[0027] F. 将所有所述待匹配关键词构成所述结构化数据。

[0028] 所述对整合后的所述一个以上的信息数据进行处理形成索引库优选为进一步包括：提取每一个所述信息数据中包含的一个以上的关键词；将所述关键词汇总成关键词词典；则，所述待搜索文本相匹配的词典优选为所述关键词词典或外部引入的词典。

[0029] 所述将所有所述待匹配关键词构成所述结构化数据优选为包括：

[0030] 统计所有所述待匹配关键词的数量；

[0031] 统计所有重复的待匹配关键词及每一个重复的所述待匹配关键词的数量；

[0032] 剔除所有重复的待匹配关键词；

[0033] 将剔除重复的所述待匹配关键词后的其他所述待匹配关键词构成所述结构化数据。

[0034] 所述根据所述结构化数据在所述索引库中进行搜索优选为包括：

[0035] 将所述结构化数据中每一个待匹配关键词与所述索引库中的关键词进行匹配，并根据匹配结果确定搜索结果。

[0036] 所述将所述结构化数据中每一个待匹配关键词与所述索引库中的关键词进行匹配优选为包括：

[0037] a. 以随机或设定顺序选取一个所述待匹配关键词，将选定的所述待匹配关键词与所述索引库中的关键词进行匹配，并标定该选取的所述待匹配关键词为已选取；

[0038] b. 当匹配成功时，记录包含所述关键词的所述信息数据，并将关键词匹配次数加 1，将每一条包含所述关键词的所述信息数据的匹配次数加 1，设定匹配次数的初始值为 0；

[0039] c. 当匹配失败时，判断所述待匹配关键词中是否还有未标定为已选取的待匹配关键词，如果判断结果为有，则返回 a 步骤，如果判断结果为没有，则匹配结束；

[0040] 则，

[0041] 所述根据匹配结果确定搜索结果优选为包括：

[0042] 按照信息数据的匹配次数由大到小的顺序，选取设定数目的所述信息数据作为搜索结果。

[0043] 本发明还提供了一种站内信息搜索系统，包括：

[0044] 信息数据提取模块，用于提取站内不同板块中一个以上的信息数据；

[0045] 格式输出模块，用于对所述一个以上的信息数据整合为相同的 XML 数据格式输出；

[0046] 索引库形成模块，用于整合后的所述一个以上的信息数据进行处理形成索引库；

[0047] 待搜索文本确定模块，用于确定待搜索文本；

[0048] 结构化数据形成模块，用于对所述待搜索文本进行处理形成结构化数据；

[0049] 搜索模块，用于根据所述结构化数据在所述索引库中进行搜索；

[0050] 搜索结果展示模块，用于展示搜索结果。

[0051] 所述索引库形成模块优选为包括：

[0052] 信息数据编号子模块，优选为用于对所述一个以上的信息数据进行信息数据编号；

[0053] 提取关键词子模块，优选为用于提取每一个所述信息数据中包含的一个以上的关键词；

[0054] 关键词编号子模块，优选为用于对所述一个以上的关键词进行关键词编号；

[0055] 关键词位置子模块，优选为用于记录每一个关键词在包含所述关键词的所述信息数据中的关键词位置；

[0056] 关键词次数子模块，优选为用于记录每一个关键词在包含所述关键词的所述信息数据中出现的关键词次数；

[0057] 对应关系子模块，优选为用于根据所述信息数据编号、关键词编号、关键词位置、关键词次数之间的对应关系建立包含所述对应关系的所述索引库。

[0058] 所述结构化数据形成模块优选为包括：词典载入子模块、待匹配子文本抽取子模块、匹配子模块、匹配成功子模块、匹配失败子模块、结构化数据构建子模块

[0059] 所述词典载入子模块，优选为用于载入与所述待搜索文本相匹配的词典；

[0060] 所述待匹配子文本抽取子模块，优选为用于按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的一个待匹配子文本；

[0061] 所述匹配子模块，优选为用于按随机或设定的方式将所述待匹配子文本与所述词典进行匹配；

[0062] 所述匹配成功子模块，优选为用于当匹配成功时，记录所述待匹配子文本中匹配成功的待匹配关键词，并且判断所述待匹配文本是否存在未被抽取的文字；当判断结果为存在时，再按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的下一个待匹配子文本，并将抽取的所述下一个待匹配子文本发送到所述匹配子模块；当判断结果为不存在时，将匹配成功的所有所述待匹配关键词发送给结构化数据构建子模块；

[0063] 所述匹配失败子模块，优选为用于当匹配失败时，从所述待匹配子文本的任一端删减一个以上的文字，形成新的所述待匹配子文本，所述待匹配子文本进一步包括删减文字后的文本，并将删减过文字的待匹配子文本发送给所述匹配子模块；

[0064] 所述结构化数据构建子模块，优选为用于将所有所述待匹配关键词构成所述结构化数据。

[0065] 通过本发明提供的一种站内信息搜索方法和搜索系统，能够达到如下效果：

[0066] 1. 加快站内信息搜索的速度。本发明对不同板块的信息整合为相同的 XML 数据格式输出，方便索引库的建立，由于格式统一，提高了索引库建立的速度，而且索引库的格式也相对统一，在进行搜索时，提高了搜索速度，和降低了确定搜索结果的用时，同时，对待搜索文本进行结构化数据处理，避免直接用待搜索文本进行搜索，同时将待搜索文本进行结构化处理后，形成了具有一定结构化的数据格式，其在索引库中进行搜索时缩短了搜索时间。

[0067] 2. 索引库易扩展。本发明对不同板块的信息整合为相同的 XML 数据格式输出，易于相同格式输出的外部信息数据的加入，同时，易于已建立的索引库根据信息数据的改进进行更新，索引库整体的扩展性强。

附图说明

[0068] 为了更清楚地说明本发明实施例或现有技术中的技术方案，以下将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，以下描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员而言，在不付出创造性劳动的前提下，还可以根据这些附图所示实施例得到其它的实施例及其附图。

[0069] 图 1 为本发明中站内信息搜索方法一个具体实施例的示意图。

[0070] 图 2 为本发明中站内信息搜索系统一个具体实施例的结构示意图。

[0071] 图 3 为本发明中站内信息搜索方法另一个具体实施例的示意图。

[0072] 图 4 为图 3 中在步骤 312- 步骤 317 间对匹配进一步步骤细化的示意图。

具体实施方式

[0073] 以下将结合附图对本发明各实施例的技术方案进行清楚、完整的描述，显然，所描述的实施例仅仅是本发明的一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所得到的所有其它实施例，都属于本发明所保护的范围。

[0074] 本发明提供一种站内信息搜索方法，如图 1，所示，包括：

[0075] 见步骤 101，提取站内不同板块中一个以上的信息数据；

[0076] 见步骤 102，对所述一个以上的信息数据整合为相同的 XML 数据格式输出；

[0077] 见步骤 103，对整合后的所述一个以上的信息数据进行处理形成索引库；

[0078] 见步骤 104，确定待搜索文本；

[0079] 见步骤 105，对所述待搜索文本进行处理形成结构化数据；

[0080] 见步骤 106，根据所述结构化数据在所述索引库中进行搜索；

[0081] 见步骤 107，展示搜索结果。

[0082] 本发明对不同板块的信息整合为相同的 XML 数据格式输出，方便索引库的建立，由于格式统一，提高了索引库建立的速度，而且索引库的格式也相对统一，在进行搜索时，提高了搜索速度，和降低了确定搜索结果的用时，同时，对待搜索文本进行结构化数据处理，避免直接用待搜索文本进行搜索，同时将待搜索文本进行结构化处理后，形成了具有一定结构化的数据格式，其在索引库中进行搜索时缩短了搜索时间。

[0083] 本发明还提供了一种站内信息搜索系统，如图 2，所示，包括：

[0084] 信息数据提取模块，用于提取站内不同板块中一个以上的信息数据；

[0085] 格式输出模块，用于对所述一个以上的信息数据整合为相同的 XML 数据格式输出；

[0086] 索引库形成模块，用于整合后的所述一个以上的信息数据进行处理形成索引库；

[0087] 待搜索文本确定模块，用于确定待搜索文本；

[0088] 结构化数据形成模块，用于对所述待搜索文本进行处理形成结构化数据；

[0089] 搜索模块，用于根据所述结构化数据在所述索引库中进行搜索；

[0090] 搜索结果展示模块，用于展示搜索结果。

[0091] 所述索引库形成模块优选为包括：

[0092] 信息数据编号子模块，优选为用于对所述一个以上的信息数据进行信息数据编号；

[0093] 提取关键词子模块，优选为用于提取每一个所述信息数据中包含的一个以上的关键词；

[0094] 关键词编号子模块，优选为用于对所述一个以上的关键词进行关键词编号；

[0095] 关键词位置子模块，优选为用于记录每一个关键词在包含所述关键词的所述信息数据中的关键词位置；

[0096] 关键词次数子模块，优选为用于记录每一个关键词在包含所述关键词的所述信息数据中出现的关键词次数；

[0097] 对应关系子模块，优选为用于根据所述信息数据编号、关键词编号、关键词位置、关键词次数之间的对应关系建立包含所述对应关系的所述索引库。

[0098] 所述结构化数据形成模块优选为包括：词典载入子模块、待匹配子文本抽取子模块、匹配子模块、匹配成功子模块、匹配失败子模块、结构化数据构建子模块

[0099] 所述词典载入子模块，优选为用于载入与所述待搜索文本相匹配的词典；

[0100] 所述待匹配子文本抽取子模块，优选为用于按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的一个待匹配子文本；

[0101] 所述匹配子模块，优选为用于按随机或设定的方式将所述待匹配子文本与所述词典进行匹配；

[0102] 所述匹配成功子模块，优选为用于当匹配成功时，记录所述待匹配子文本中匹配成功的待匹配关键词，并且判断所述待匹配文本是否存在未被抽取的文字；当判断结果为存在时，再按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的下一个待匹配子文本，并将抽取的所述下一个待匹配子文本发送到所述匹配子模块；当判断结果为不存在时，将匹配成功的所有所述待匹配关键词发送给结构化数据构建子模块；

[0103] 所述匹配失败子模块，优选为用于当匹配失败时，从所述待匹配子文本的任一端删减一个以上的文字，形成新的所述待匹配子文本，所述待匹配子文本进一步包括删减文字后的文本，并将删减过文字的待匹配子文本发送给所述匹配子模块；

[0104] 所述结构化数据构建子模块，优选为用于将所有所述待匹配关键词构成所述结构化数据。

[0105] 如图 3，所示，本发明的站内信息搜索方法的具体实施方式为：

[0106] 图 3，步骤 301，提取站内不同板块中一个以上的信息数据；

[0107] 由于本发明是针对一个网站不同板块信息数据的搜索，因此需要对站内不同板块的信息数据进行提取，而且信息数据的量非常多，格式，内容，均不统一；

[0108] 图 3，步骤 302，对所述一个以上的信息数据整合为相同的 XML 数据格式输出；

[0109] 针对步骤 301 提到的信息数据的格式不统一的问题，在本步骤中对其进行了整合，整合为相同的 XML 数据格式输出，采用哪一种相同的数据格式，不同的实施者可以根据需要或个性化需求进行定义；

[0110] 整合的好处包括：一、整合后的信息数据便于下一步信息数据的处理；二、便于其他外部信息数据的加入，因为外部信息数据只要同样将其信息数据整合成网站定义的数据格式则可以方便的集成到网站内；三、易于跨平台整合，易于对信息数据进行扩展性处理；

[0111] 图 3，步骤 303，对所述一个以上的信息数据进行信息数据编号；

[0112] 对信息数据进行编号，目的在于在建立索引库（类似目录）时，需要引用或建立连接，进行编号便于引用、建立连接和查找；

[0113] 图 3，步骤 304，提取每一个所述信息数据中包含的一个以上的关键词；

[0114] 本发明采用了关键词匹配搜索的方式，因此需要对每一个信息数据进行关键词提取，从而方便建立关键词和包含关键词的信息数据之间的对应关系；

[0115] 关键词是针对信息数据全文进行提取的，因此每一个信息数据中会包含很多关键词，每一个关键词有可能会出现很多次；

[0116] 这一步还包括：提取每一个所述信息数据中包含的一个以上的关键词；将所述关键词汇总成关键词词典；

[0117] 图 3，步骤 305，对所述一个以上的关键词进行关键词编号；

[0118] 针对步骤 304 对关键词的分析，在本步骤中对关键词进行编号，编号的好处与信息数据编号有相同之处；

[0119] 图 3，步骤 306，记录每一个关键词在包含所述关键词的所述信息数据中的关键词位置；

[0120] 针对步骤 304 对关键词的分析，本步骤中对关键词在包含关键词的信息数据中的位置进行记录，一旦关键词在包含该关键词的信息数据中的位置确定之后，更有利于建立关键词与信息数据的对应关系，同时利于关键词搜索过程中对关键词的查找；

[0121] 其中关键词位置包括：一、记录该关键词在包含该关键词的信息数据中的字符位置，优点在于定位块；二、记录该关键词在包含该关键词的信息数据中的第几个关键词，优点在于节约索引库占用的空间大小，查找快速；本发明的关键词位置优选为第二种位置。

[0122] 图 3，步骤 307，记录每一个关键词在包含所述关键词的所述信息数据中出现的关键词次数；

[0123] 本步骤主要是针对步骤 306 进行的处理；

[0124] 需要说明的是步骤 305-307 的顺序可以不分先后，采用何种顺序可以由实施者自由定义；

[0125] 图 3，步骤 308，根据所述信息数据编号、关键词编号、关键词位置、关键词次数之间的对应关系建立包含所述对应关系的所述索引库；

[0126] 步骤 303- 步骤 308 完成了图 1 中的步骤 103，即对整合后的所述一个以上的信息数据进行处理形成索引库；

[0127] 图 3，步骤 309，确定待搜索文本；

[0128] 待搜索文本可以来自于网络使用者的客户端的输入，或者网站开发者开发端输入或其他方式的输入；

[0129] 图 3，步骤 310，载入与所述待搜索文本相匹配的词典；

[0130] 本步骤采用的词典可以为步骤 304 形成的关键词词典，也可以为外部引入的词典；

[0131] 词典中包括各种语言、各种领域的词语、文字等；

[0132] 图 3，步骤 311，按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的一个待匹配子文本；

[0133] 设定的顺序可以为按从待搜索文本的首端开始连续选取或者从待搜索文本的末端开始连续选取；

[0134] 选取的待匹配子文本中包含的文字数目可以根据实施者的实施需要或个性化需求进行自由定义，通常情况下建议小于等于 7 个文字，包括标点符号；7 个文字比较符合中国现有词典的构成方式，因为中国现有词典中包含的词最多为 7 个文字；

[0135] 举例：比如待搜索文本为："按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的一个待匹配子文本"，则第一次选取的待搜索子文本可以为从首端抽取的"按随机或设定顺"或者从末端抽取的"个待匹配子文本"；

[0136] 图 3 或图 4，步骤 312，按随机或设定的方式将所述待匹配子文本与所述词典进行匹配；

[0137] 匹配的方法为将选取的待匹配子文本与词典中的现有词汇进行完全重合性的匹配；

[0138] 下面，将对图 3，步骤 312- 图 3，步骤 317 的匹配步骤进行细化，见图 4，所示：

[0139] 图 4，步骤 313，当匹配成功时，记录所述待匹配子文本中匹配成功的待匹配关键词，并且判断所述待匹配文本是否存在未被抽取的文字；

[0140] 因为匹配会出现匹配成功和匹配失败，本步骤即对匹配成功后的处理步骤进行说明；

[0141] 因为待搜索文本是有限的文字，因此抽取待搜索子文本的个数是有限的，需要判断是否还有未被选取的待搜索文本中的未抽取的文字，仅此需要进行判断；

[0142] 图 4，步骤 314，当判断结果为存在时，再按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的下一个待匹配子文本，返回步骤 312；

[0143] 本步骤是针对步骤 313 中的判断结果为不包括引起的下一步操作进行的说明；

[0144] 当判断结果为存在，说明待搜索文本中的文字还没有全部被抽取，因此还要继续进行抽取，所以继续进行步骤 312；

[0145] 图 4，步骤 315，当判断结果为不存在时，进行 317 步骤；

[0146] 本步骤是针对步骤 313 中的判断结果为包括引起的下一步操作进行的说明；

[0147] 当判断结果为不存在，说明待搜索文本中的文字被全部抽取，而且与词典的匹配进行完毕；

[0148] 图 4，步骤 316，当匹配失败时，从所述待匹配子文本的任一端删减一个以上的文字，形成新的所述待匹配子文本，所述待匹配子文本进一步包括删减文字后的文本，返回步骤 312；

[0149] 因为匹配会出现匹配成功和匹配失败，本步骤即对匹配失败后的处理步骤进行说明；

[0150] 结合步骤 311 的例子，匹配失败即为"按随机或设定顺"在词典中找不到与其相同的词语，因此，需要删减最后一个字，将待搜索子文本更改为"按随机或设定"，结合步骤 312，得到待匹配关键词为了"随机""设定""或"；

[0151] 图 3 或图 4，步骤 317，将所有所述待匹配关键词构成所述结构化数据；

[0152] 这种结构化数据的呈现形式有多种多样，可以将所有记录的匹配成功的待搜索子文本按随机的方式罗列，也可以按照设定的顺序罗列，其中包括按照从待搜索文本首端到

末端的出现顺序罗列；

[0153] 图 3，步骤 318，剔除所有重复的待匹配关键词；

[0154] 针对图 3 或图 4，步骤 317 对所有待匹配关键词的分析，可以将重复的待匹配关键词进行删除，按设定的顺序进行查找，遇到与以查找锅的待匹配关键词重复的则删除；

[0155] 具体处理可分为以下几步：

[0156] 统计所有所述待匹配关键词的数量；

[0157] 统计所有重复的待匹配关键词及每一个重复的所述待匹配关键词的数量；

[0158] 剔除所有重复的待匹配关键词；

[0159] 将剔除重复的所述待匹配关键词后的其他所述待匹配关键词构成所述结构化数据；

[0160] 则：结构化数据包括了剔除重复的所述待匹配关键词后的所有待匹配关键词。

[0161] 本步骤是为了将重复的关键词进行简短归一化；

[0162] 图 3，步骤 309- 图 3，步骤 318 完成了图 1 中步骤 105，即对所述搜索文本进行处理形成结构化数据；

[0163] 结构化数据处理的方式还可以为：正向最大匹配算法。

[0164] 图 3，步骤 319，以随机或设定顺序选取一个所述待匹配关键词，将选定的所述待匹配关键词与所述索引库中的关键词进行匹配，并标定该选取的所述待匹配关键词为已选取；

[0165] 图 3，步骤 320，当匹配成功时，记录包含所述关键词的所述信息数据，并将关键词匹配次数加 1，将每一条包含所述关键词的所述信息数据的匹配次数加 1，设定匹配次数的初始值为 0；

[0166] 图 3，步骤 321，当匹配失败时，判断所述待匹配关键词中是否还有未标定为已选取的待匹配关键词，如果判断结果为有，则返回 a 步骤，如果判断结果为没有，则匹配结束；

[0167] 图 3，步骤 322，按照信息数据的匹配次数由大到小的顺序，选取设定数目的所述信息数据作为搜索结果；

[0168] 步骤 319- 步骤 322 完成了图 1 中步骤 106，即根据所述结构化数据在所述索引库中进行搜索；

[0169] 图 3，步骤 323，展示搜索结果。

[0170] 图 3，步骤 301- 图 3，步骤 323 完成站内信息的搜索。

[0171] 通过本发明提供的一种站内信息搜索方法和搜索系统，能够达到如下效果：

[0172] 1. 加快站内信息搜索的速度。本发明对不同板块的信息整合为相同的 XML 数据格式输出，方便索引库的建立，由于格式统一，提高了索引库建立的速度，而且索引库的格式也相对统一，在进行搜索时，提高了搜索速度，和降低了确定搜索结果的用时，同时，对待搜索文本进行结构化数据处理，避免直接用待搜索文本进行搜索，同时将待搜索文本进行结构化处理后，形成了具有一定结构化的数据格式，其在索引库中进行搜索时缩短了搜索时间。

[0173] 2. 索引库易扩展。本发明对不同板块的信息整合为相同的 XML 数据格式输出，易于相同格式输出的外部信息数据的加入，同时，易于已建立的索引库根据信息数据的改进进行更新，索引库整体的扩展性强。

[0174] 本发明提供的各种实施例可根据需要以任意方式相互组合，通过这种组合得到的技术方案，也在本发明的范围内。

[0175] 显然，本领域技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若对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，则本发明也包含这些改动和变型在内。

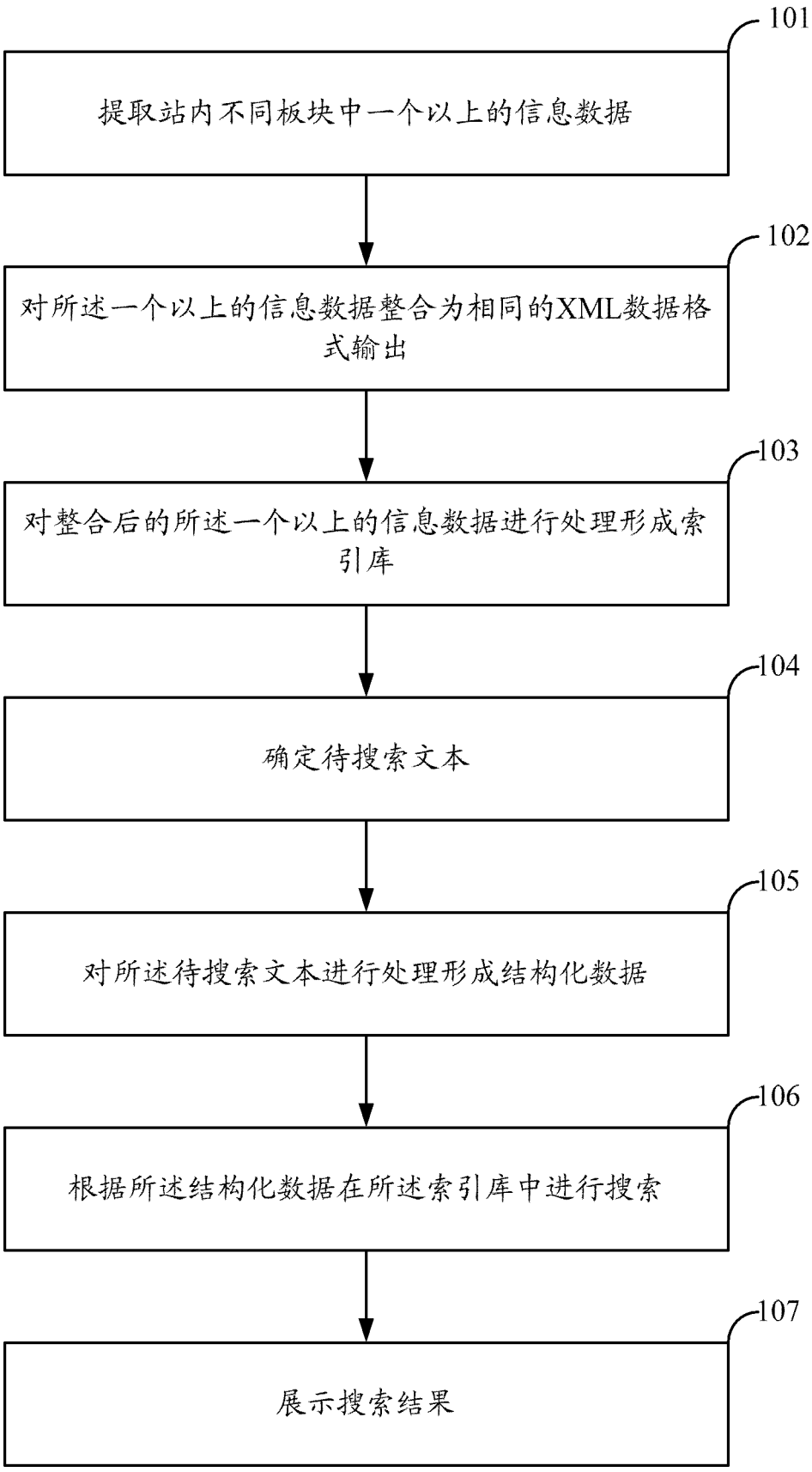

101
提取站内不同板块中一个以上的信息数据
102
对所述一个以上的信息数据整合为相同的XML数据格式输出
103
对整合后的所述一个以上的信息数据进行处理形成索引库
104
确定待搜索文本
105
对所述待搜索文本进行处理形成结构化数据
106
根据所述结构化数据在所述索引库中进行搜索
107
展示搜索结果

图 1

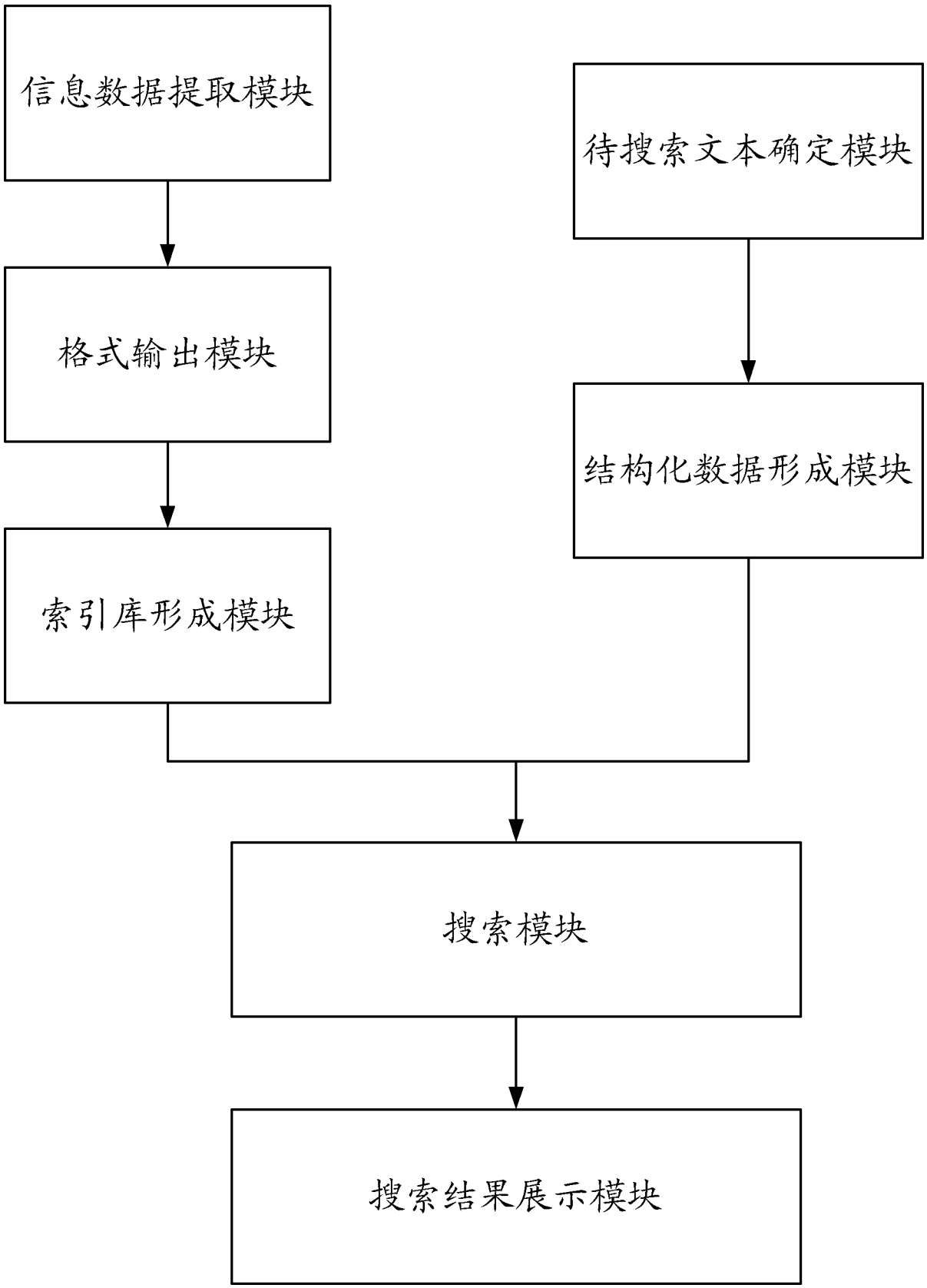

信息数据提取模块
待搜索文本确定模块
格式输出模块
结构化数据形成模块
索引库形成模块
搜索模块
搜索结果展示模块

图 2

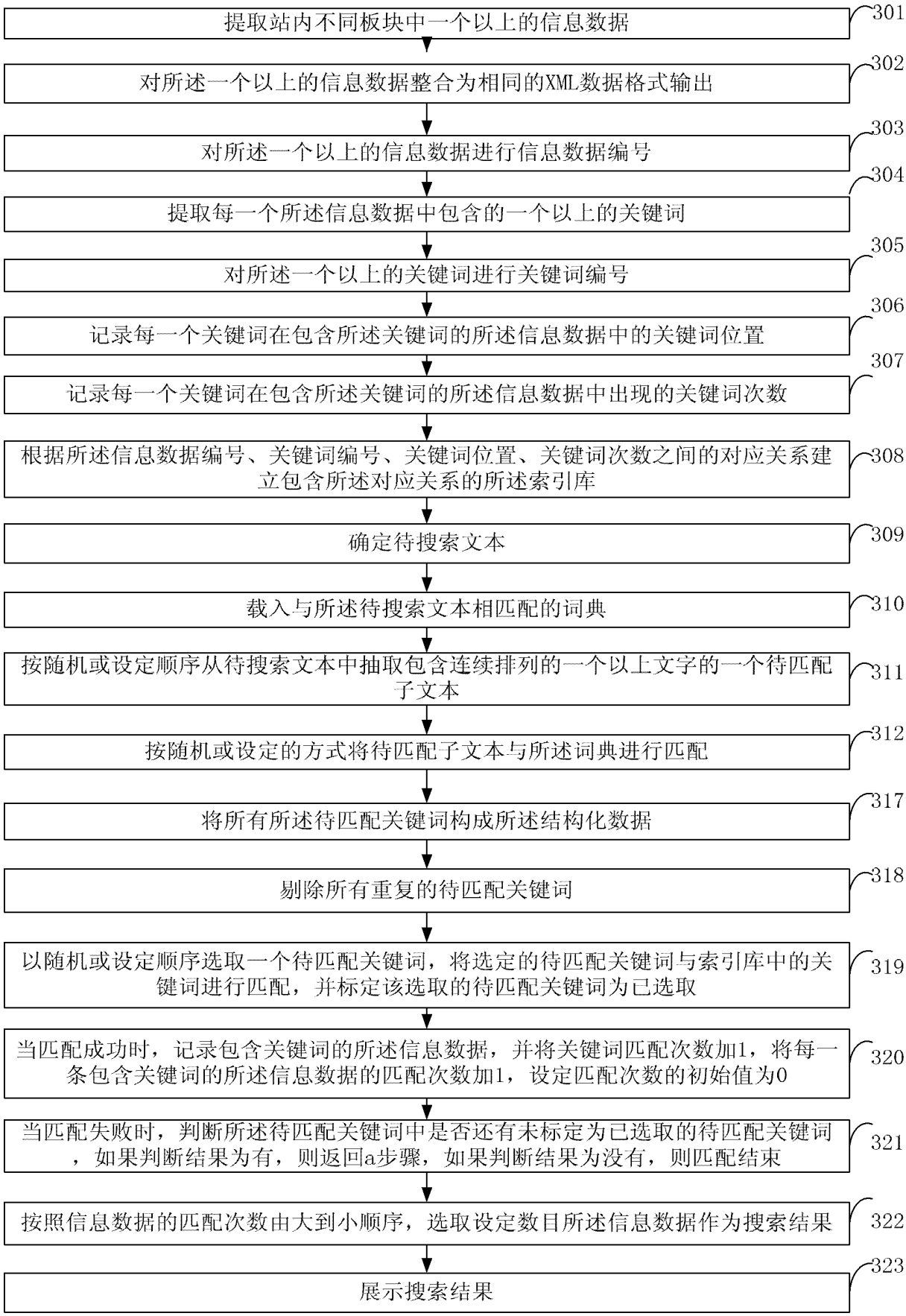

提取站内不同板块中一个以上的信息数据
301
对所述一个以上的信息数据整合为相同的XML数据格式输出
302
对所述一个以上的信息数据进行信息数据编号
303
提取每一个所述信息数据中包含的一个以上的关键词
304
对所述一个以上的关键词进行关键词编号
305
记录每一个关键词在包含所述关键词的所述信息数据中的关键词位置
306
记录每一个关键词在包含所述关键词的所述信息数据中出现的关键词次数
307
根据所述信息数据编号、关键词编号、关键词位置、关键词次数之间的对应关系建立包含所述对应关系的所述索引库
308
确定待搜索文本
309
载入与所述待搜索文本相匹配的词典
310
按随机或设定顺序从待搜索文本中抽取包含连续排列的一个以上文字的一个待匹配子文本
311
按随机或设定的方式将待匹配子文本与所述词典进行匹配
312
将所有所述待匹配关键词构成所述结构化数据
317
剔除所有重复的待匹配关键词
318
以随机或设定顺序选取一个待匹配关键词，将选定的待匹配关键词与索引库中的关键词进行匹配，并标定该选取的待匹配关键词为已选取
319
当匹配成功时，记录包含关键词的所述信息数据，并将关键词匹配次数加1，将每一条包含关键词的所述信息数据的匹配次数加1，设定匹配次数的初始值为0
320
当匹配失败时，判断所述待匹配关键词中是否还有未标定为已选取的待匹配关键词，如果判断结果为有，则返回a步骤，如果判断结果为没有，则匹配结束
321
按照信息数据的匹配次数由大到小顺序，选取设定数目所述信息数据作为搜索结果
322
展示搜索结果
323

图 3

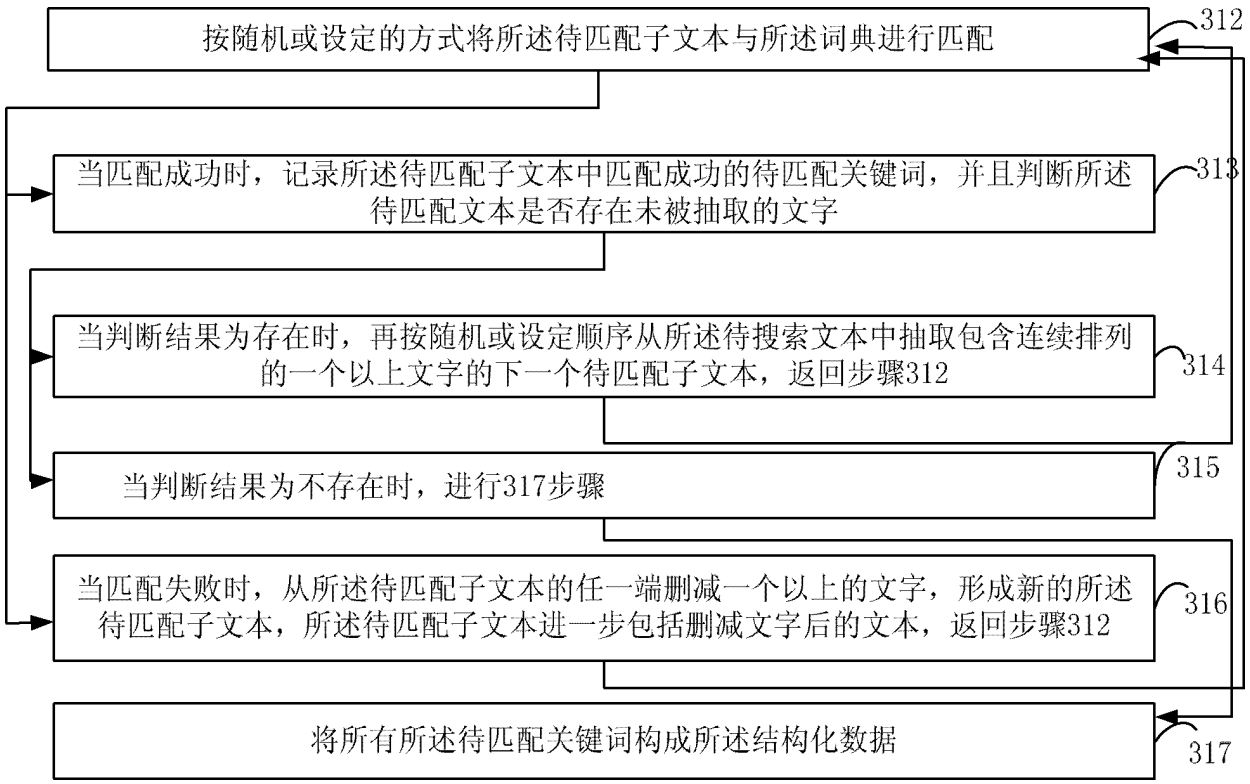

按随机或设定的方式将所述待匹配子文本与所述词典进行匹配
312
当匹配成功时，记录所述待匹配子文本中匹配成功的待匹配关键词，并且判断所述待匹配文本是否存在未被抽取的文字
313
当判断结果为存在时，再按随机或设定顺序从所述待搜索文本中抽取包含连续排列的一个以上文字的下一个待匹配子文本，返回步骤312
314
当判断结果为不存在时，进行317步骤
315
当匹配失败时，从所述待匹配子文本的任一端删减一个以上的文字，形成新的所述待匹配子文本，所述待匹配子文本进一步包括删减文字后的文本，返回步骤312
316
将所有所述待匹配关键词构成所述结构化数据
317

图 4