

[19] 中华人民共和国国家知识产权局

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



[12] 发明专利说明书

专利号 ZL 200410102046.0

[45] 授权公告日 2009 年 12 月 2 日

[11] 授权公告号 CN 100565503C

[22] 申请日 2004. 12. 15

[21] 申请号 200410102046.0

[30] 优先权

[32] 2003. 12. 15 [33] US [31] 10/735,999

[73] 专利权人 微软公司

地址 美国华盛顿州

[72] 发明人 C·B·威尔

[56] 参考文献

CN1216841A 1999. 5. 19

CN1336604A 2002. 2. 20

US6012053A 2000. 1. 4

Generation and Search of Clustered Files. G.

SALTON, A. WONG. ACM Transactions on Database Systems, Vol. 3 No. 4. 1978

审查员 郑守志

[74] 专利代理机构 上海专利商标事务所有限公司

代理人 谢喜堂

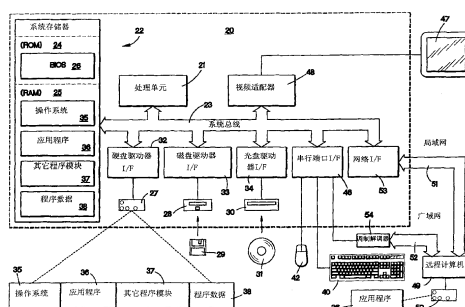
权利要求书 5 页 说明书 12 页 附图 3 页

[54] 发明名称

动态内容聚类

[57] 摘要

本发明提供了动态更新用于在因特网上发布的信息的一种方法和系统。 来自诸如新闻社的多个源的接收的信息中提取富有意义的内容。 该信息包括文本、声音、图像和视频。 可确定该接收的信息的一组表征性特征。 有共同表征性特征的信息被归组成多个聚类。 在归组步骤中获得的信息可用于确定如何基于客户的信息请求来发布聚类中所包含的信息。 这种客户请求可基于查询或分配给该客户的用户提问单。



1. 一种动态更新用于发布的信息的方法，其特征在于，它包括：
 - a)从接收的信息中抽取一套表征所述接收的信息的表征性特征，其中所述抽取包括根据在所述信息中权标出现的位置和频率提供相关性排序列表，该列表具有 K 个有最高相关性因子的权标；
 - b)根据权标-相关性对的内积将有共同表征性特征的接收的信息归组成多个聚类；以及
 - c)使用在归组步骤中的所获信息，基于客户的信息请求发布包含在一聚类中的信息。
2. 如权利要求 1 所述的方法，其特征在于，所述接收的信息包括文本数据、图形数据、或视频数据中的一种或多种的组合。
3. 如权利要求 1 所述的方法，其特征在于，所述接收的信息包括一给定类型的多个特征，且其中，在特征被提取时对所述多个特征按重要性进行了分级。
4. 如权利要求 3 所述的方法，其特征在于，一个聚类具有对聚类特征的归纳，且还包括：通过取对新接收的信息共同的特征与归纳所述聚类的特征的内积，来对归纳新接收的信息的特征和归纳于聚类中的特征进行比较，以及在所述内积超过一阈值时将所述新接收的信息并入聚类。
5. 如权利要求 1 所述的方法，其特征在于，对新接收信息的 K 个最高特征与一聚类的 K 个最高特征进行比较，以确定所述信息是否加入聚类中。
6. 如权利要求 5 所述的方法，其特征在于，每一特征都有一借以对所述特征进行衡量的相关性因子，且还包括在将所述接收的信息添加到一聚类中前，确定聚类和新接收信息是否至少有 L 个具有非零相关性因子的共同特征。
7. 如权利要求 1 所述的方法，其特征在于，还包括将具有共同特征的聚类归组，以产生一响应客户请求而全部发布的聚类的邻域。
8. 如权利要求 7 所述的方法，其特征在于，所述接收的信息是一包含

文档的文本，且邻域的相关性被用于确定是否对客户发布邻域中的文档。

9. 如权利要求 8 所述的方法，其特征在于，所述邻域的相关性取决于该文档被分配到该邻域的时间长短而变化。

10. 如权利要求 8 所述的方法，其特征在于，所述邻域的相关性根据信息请求中所包含的信息而变化。

11. 如权利要求 7 所述的方法，其特征在于，一条接收的信息可被归组到一个以上聚类中，但仅在一个邻域中发布。

12. 如权利要求 11 所述的方法，其特征在于，还包括维护一空邻域，并且在开始接收所述信息时，将接收的信息添加到所述空邻域中。

13. 如权利要求 11 所述的方法，其特征在于，还包括维护一空邻域，并且在因为所述邻域的重新构建而使邻域内容改变时，将接收的信息添加到所述空邻域中。

14. 如权利要求 11 所述的方法，其特征在于，还包括维护一空邻域，并且当接收的信息变成与一邻域不相关时，将接收的信息添加到所述空邻域中。

15. 如权利要求 11 所述的方法，其特征在于，所述归组步骤进一步包括：

a) 把所述信息中具有大于一阈值的共性的文档进行归组，以提供所述文档的多个文档聚类；所述归组执行如下：

i) 将一候选文档的 K 个权标与权标文档聚类的一个表征性权标组进行比较；

ii) 如果所述比较表示候选文档和文档聚类间有充分的相似度，把该候选文档添加到所述文档聚类中去；

iii) 考虑所添加的候选文档，更新文档聚类归纳；以及

b) 基于一请求发布分配到指定的一个或多个文档聚类中的文档。

16. 如权利要求 15 所述的方法，其特征在于，文档有文档类别，且所述抽取包括确定一文档类别中权标的类别频率，并基于所述类别频率将相关性因子分配给所述权标。

17. 如权利要求 16 所述的方法，其特征在于，基于在所述文档中权标的位置，向权标赋予相关性因子。

18. 如权利要求 15 所述的方法, 其特征在于, 如果所述候选文档不充分相似于一聚类, 它形成自己新聚类的基础。

19. 如权利要求 15 所述的方法, 其特征在于, 权标的相关性因子由关系式 $e^{(-\alpha * P_{0i})} * N_i * R_i$ 确定, 其中 α 为权标相似性的衰变率, 它是与文档 D 文本起始点距离的函数, P_{0i} 是权标 i 第一次出现在文本中的位置, N_i 是权标 i 的出现次数, 而 R_i 是权标 i 在文档 D 所属文档类别中文档频率倒数的对数。

20. 如权利要求 15 所述的方法, 其特征在于, 文档的聚类被聚集在一起, 以形成文档要分配到其中的邻域。

21. 如权利要求 16 所述的方法, 其特征在于, 邻域被赋予随时间而改变的邻域相关性因子, 所述邻域相关性因子被用来确定邻域向谁发布。

22. 如权利要求 21 所述的方法, 其特征在于, 所述邻域相关性值还随着组成邻域的文档之文档相关性因子而变化。

23. 如权利要求 22 所述的方法, 其特征在于, 所述文档相关性因子取决于文档源的质量。

24. 如权利要求 22 所述的方法, 其特征在于, 所述文档相关性因子取决于文档源的位置和请求者的位置。

25. 如权利要求 21 所述的方法, 其特征在于, 所述邻域相关性因子随着分配给所述邻域的文档类别而变化。

26. 一种用于评估文档的系统, 其特征在于, 它包括:

a) 一预处理器, 用于从多个文档源之一接收文本文档、并通过根据在文本数据中权标出现的位置和频率抽取 K 个有最高相关性因子的权标来评估包含在每个接收的文档中的所述文本数据以确定基于一请求对文档进行随后发布的合适性; 所述预处理器根据权标-相关性对的内积将其共性大于阈值的诸文档归组, 以提供多个所述文档的聚类; 以及

b) 一 web 服务器, 可从所述预处理器访问聚类数据, 使一请求者可基于该请求者的请求和指定的一个或多个聚类的文档中所包含文本的归纳之间的比较获得聚类中包含的文档。

27. 如权利要求 26 所述的系统, 其特征在于, 所述预处理器可将聚类归组成一聚类邻域, 且其中, 邻域中的文档可为请求者所用。

28. 如权利要求 26 所述的系统,其特征在于,基于整个聚类的聚类相关性,文档聚类可从网络服务器的发布中移除。

29. 一种用于动态更新用于发布的信息的设备,其特征在于,包括:

a) 用于从接收的信息中提取一组表征所述接收的信息的表征性特征的装置,其中所述抽取包括根据在所述信息中权标出现的位置和频率提供相关性排序列表,该列表具有 K 个有最高相关性因子的权标;

b) 用于根据权标-相关性对的内积把有共同表征性特征的所述接收的信息归组成多个聚类的装置;

c) 用于使用在归组过程中获得的信息,基于客户的信息请求发布聚类中所有的信息的装置。

30. 如权利要求 29 所述的设备,其特征在于,所述接收的信息包括文本数据、图像数据或视频数据中的一种或多种的组合。

31. 如权利要求 29 所述的设备,其特征在于,所述接收的信息包括一给定类型的多个特征,且其中在特征被提取时所述多个特征按重要性进行分级。

32. 如权利要求 29 所述的设备,其特征在于,一聚类包括对聚类特征的归纳,并且还包括通过取对新接收的信息共同的特征与归纳所述聚类的特征的内积,对归纳新接收的信息的特征和归纳在聚类中的特征进行比较,以及当所述内积超过一阈值时将新接收的信息加入聚类。

33. 如权利要求 29 所述的设备,其特征在于,新接收的信息的 K 个最高特征与一聚类的 K 个最高特征相比较,以确定所述信息是否加入聚类中。

34. 如权利要求 29 所述的设备,其特征在于,每一特征都有所述特征得以衡量的一相关性因子,且包括在将所述接收的信息添加到一聚类中前,确定聚类和新接收的信息是否至少有 L 个具有非零相关性因子的共同特征。

35. 如权利要求 29 所述的设备,其特征在于,包括用于把有共同特征的聚类归组的装置,以便产生响应于客户请求可全部发布的一聚类邻域。

36. 如权利要求 35 所述的设备,其特征在于,邻域的相关性被用于确定是否向客户发布在邻域中文档。

37. 如权利要求 36 所述的设备,其特征在于,所述邻域的相关性随着文档在邻域中的时间长短而变化。

38. 如权利要求 36 所述的设备,其特征在于,所述邻域的相关性根据信息请求中所包含信息而变化。

39. 如权利要求 35 所述的设备,其特征在于,还包括用于维护一空邻域的装置,并且在开始接收所述信息时,将接收的信息添加到所述空邻域中。

40. 如权利要求 35 所述的设备,其特征在于,还包括用于维护一空邻域的装置,并且在因为所述邻域的重新构建而使邻域内容改变时,将接收的信息添加到所述空邻域中。

41. 如权利要求 35 所述的设备,其特征在于,还包括用于维护一空邻域的装置,并且当一邻域与接收信息变成不相关时,将接收信息添加到所述空邻域中。

动态内容聚类

(1)技术领域

本发明涉及为文档自动分类和再版所进行的文档分析。

(2)背景技术

希望从诸如因特网的数据源中找到相关的最新信息的用户不断地面对的是新内容的纷至沓来。通过把相似内容归组 (group together), 在这大量数据中从头至尾进行拣选的任务即可得到简化。

现有技术已被用来自动分拣基于 web 的原始文档。Lin 等所著的题为“从 web 文档中发现报道性内容块” (Discovering Informative Content Blocks from Web Documents) 的文章描述了从 web 文本的富有意义内容中自动移除冗余数据的过程。该文的目的是从网页上所显示的冗余的、重复的、和通常不受关注的数据中分拣出有意义数据。

一旦冗余数据从网页中被剥离, 网页的文字内容就可使用已知索引技术来进行分类了。然后, 被索引后的网页可通过诸如 Google、MSN 或 Yahoo 的现有网页搜索引擎来评估。Lin 等所著文章丢弃了被认为有冗余数据的网页不相关部分, 但并未改变被发现含有有意义信息的文本页面的索引或评估。

Watters 等著的题为“根据相似性对新闻文档分级” (Rating News Documents for Similarity) 的文章涉及新闻文档的个性化传送系统。该文讨论了基于特征短语的提取把新闻文档相关联的一套方法, 其中特征短语标识日期、地点、人物、和组织。从这些特征短语可创建新闻表示, 用来定义可对其进行比较和分级从而发现相关新闻报道的新闻对象。

在更大的搜索问题上下文中, 本发明提供了一种方法, 用户可由此快速浏览大量信息并通过仅显示概念上有显著性的内容来锁定他们所关注的事项。

(3)发明内容

本发明揭示了自动更新用于发布的信息的方法和系统。富有意义的内容从接收自多种来源的信息中被提取。该信息可包含文本、声音、图像和视频。所收到信息的一组表征性特征被确定。有相同表征性特征的信息被归组到一定数目的聚类之中。在归组步骤中获得的信息被用来确定如何基于客户信息要求来发布聚类中所包含的信息。此客户要求可基于查询或分配给该客户的客户提问单。

本发明的一种用途是用于 Newsbot 自动化新闻门户，其使用动态内容聚类，来不断地标识和改变全球范围的基于 web 的新闻网站所呈显的大量新闻故事。当文章进入 Newsbot 系统，如果它们是以先前涉及的故事为中心内容，就被分配到已先存在的聚类中去，否则它们就被分配到新的聚类中去。Newsbot 系统中聚类的状态每隔几分钟就被记录到用于建立各种 Newsbot 网页的目录文件中去。

根据本发明的一方面，提供了一种动态更新用于发布的信息的方法，包括：a)从接收的信息中抽取一套表征所述接收的信息的表征性特征，其中所述抽取包括根据在所述信息中权标出现的位置和频率提供相关性排列表，该列表具有 K 个有最高相关性因子的权标；b)根据权标-相关性对的内积将有共同表征性特征的接收的信息归组成多个聚类；以及 c)使用在归组步骤中的所获信息，基于客户的信息请求发布包含在一聚类中的信息。

根据本发明的另一方面，提供了一种用于评估文档的系统，包括：a) 一预处理器，用于从多个文档源之一接收文本文档、并通过根据在文本数据中权标出现的位置和频率抽取 K 个有最高相关性因子的权标来评估包含在每个接收的文档中的所述文本数据以确定基于一请求对文档进行随后发布的合适性；所述预处理器根据权标-相关性对的内积将其共性大于阈值的诸文档归组，以提供多个所述文档的聚类；以及 b) 一 web 服务器，可从所述预处理器访问聚类数据，使一请求者可基于该请求者的请求和指定的一个或多个聚类的文档中所包含文本的归纳之间的比较获得聚类中包含的文档。

根据本发明的再一方面，提供了一种用于动态更新用于发布的信息的设备，包括：a) 用于从接收的信息中提取一组表征所述接收的信息的表征

性特征的装置，其中所述抽取包括根据在所述信息中权标出现的位置和频率提供相关性排序列表，该列表具有 K 个有最高相关性因子的权标；b) 用于根据权标-相关性对的内积把有共同表征性特征的所述接收的信息归组成多个聚类的装置；c) 用于使用在归组过程中获得的信息，基于客户的信息请求发布聚类中所有的信息的装置。

参照附图阅读所附示例性实施例，这些和其他目标、优点和特征将能得到更好的理解。

(4)附图说明

图 1 是实现本发明示例性实施例组件的一种代表性计算机系统；以及图 2 是用来向请求者呈现数据的基于互联网的文档检索系统示意性表示；

图 3-5 是为了请求者的有效访问而对那些文档进行分类的过程中所执行的文档处理步骤流程图；以及

图 6A 和 6B 是关联在一起形成此类数据聚类的邻域的文档数据聚类示意图；以及

图 7 是交叠的聚类邻域的示意图。

(5)具体实施方式

实践本发明的示例系统

图 2 描绘的是基于用户的信息请求来评估文档并返回结果的代表性系统 1。该系统具有在带有助于从多个文档源 4 之一接收文档 3 的数据存储器的后端服务器 5 上实现的预处理器。后端服务器 5 对该文档进行评估用于随后的发布。图 1 描绘了可用于实现后端服务器的代表性计算机的架构。尽管本发明并不限于文本文档的评估，但在一示例性实施例中，所接收文档包含“XML”格式的文本。例如，当世界各地发生新闻时，新闻文档由美联社、路透社、纽约时报、CNN 等（代表性源 4）制作并被定期提供给系统 1。

示例性预处理器是后端服务器计算机 5 的一部分，该计算机运行诸如 Windows Server 软件等操作系统，并具有存有大量被评估和分类文档的存

存储器。每个接收到的 XML 文档中所包含的文本数据被评估或分类。在 web 服务器 7 中维护所接收文档的数据库。文档数据库的分类被更新，使得文档的输入请求（例如来自用户）可通过发布基于以下讨论的标准被认为是最适当的文档以最新的信息来响应。后端服务器 5 通过把有共同点的文档归组成被称作文档目录的众多文档聚类来完成评估或分类。

图 2 的 web 服务器 7 使用来自后端服务器 5 的聚类数据更新目录。网络服务器 7 把包含在聚类中的被网络服务器判断为最相关的文档提供给用户或客户 8。相关性的判断基于众多标准，其中部分标准可能包括特定客户提供给 web 服务器的信息。

计算机系统

图 1 描绘了一种示例性数据处理系统。诸如图 1 所示系统的数据处理系统可充当后端服务器 5 和网络服务器 7。该系统具有常规计算机 20 形式的通用计算设备，它包括一个或多个处理单元 21、系统存储器 22、和耦合各种系统组件包括将系统存储器耦合到处理单元 21 的系统总线 23。系统总线 23 可以是若干类型总线结构的任一种，包括存储器总线或存储器控制器、外围总线、和使用各种总线架构的任一种的本地总线。

系统存储器包括只读存储器(ROM) 24 和随机存取存储器(RAM) 25。包含帮助在计算机 20 元件间传送信息（如起动时）的基本例程的基本输入/输出系统 26（BIOS）存储在 ROM 24 中。

计算机 20 还包括读取和写入硬盘（未示出）的硬盘驱动器 27，读取和写入可移动磁盘 29 的磁盘驱动器 28，读取和写入可移动光盘 31，如 CD ROM 或其它光学介质的光盘驱动器 30。硬盘驱动器 27、磁盘驱动器 28、和光盘驱动器 30 分别由硬盘驱动器接口 32、磁盘驱动器接口 33 和光盘驱动器接口 34 连接到系统总线 23。驱动器及其相关联的计算机可读介质为计算机 20 提供计算机可读指令、数据结构、程序模块、和其它数据的非易失存储。尽管在此所述的该示例性环境采用了硬盘、可移动磁盘 29、和可移动光盘 31，但本领域技术人员应明白计算机可访问的可存储数据的其它类型计算机可读介质，如磁带、闪存卡、数字化视频光盘、Bernoulli 盘、随机存取存储器（RAM）、只读存储器（ROM）等等，也可被用于该示例

性操作环境中。

众多程序模块，包括操作系统 35、一个或多个应用程序 36、其它程序模块 37、和程序数据 38，可存储于硬盘、磁盘 29、光盘 31、ROM 24、或 RAM 25 中。用户可通过输入设备如键盘 40 和定位装置 42 向计算机 20 输入命令和信息。其它输入设备（未示出）可包括麦克风、游戏杆、游戏垫、卫星接收器、扫描仪等等。这些和其它输入设备通常通过与系统总线耦合的串行端口接口 46 连接到处理单元 21，但也可通过其它接口相连，如并行端口、游戏端口或通用串行总线（USB）。监视器 47 或其它类型显示设备也通过接口，如视频适配器 48 和系统总线 23 相连。除了显示器 742，个人计算机通常还包括其它外围输出装置（未示出），如扬声器和打印机。

计算机 20 可以在使用与一台或多台远程计算机，诸如远程计算机 49 的逻辑连接的网络化环境中运行。远程计算机 49 可以是另一台个人计算机、服务器、路由器、网络 PC、对等装置或其它普通网络节点，且通常包括上述与计算机 20 相关的许多或全部元件，尽管在图 1 中仅示出存储器存储设备 50。图 1 的逻辑连接包括局域网(LAN) 51 和广域网(WAN) 52。这样的网络化环境在办公室、企业范围计算机网络、企业内部互联网和因特网上是常见的。

当用于 LAN 网络化环境时，计算机 20 通过网络接口或适配器 53 与本地网 51 连接。当用于 WAN 网络化环境时，计算机 20 通常包括调制解调器 54 或其它用于在广域网 52 中建立通讯的设备。可以内置或外置的调制解调器 54 通过串行端口接口 46 连接到系统总线 23。在网络化环境中，与计算机 20 相关的所描述程序模块或其部分模块，可存储在远程存储设备中。可以理解的是，所示网络连接是示例性的，并且也可以使用在计算机间建立通讯连接的其它方法。

图 3-5 描绘了由后端服务器 5 对输入信息所执行的过程。该过程分成三个阶段：输入阶段 110、合并阶段 140、和分配阶段 160。在输入阶段 110 中，带有内容的数据进入该系统。在合并阶段，示例性系统通过把数据聚类将共同的数据归组，并将其存储在数据库中。在分配阶段，系统将内容分配给聚类邻域。这种分配也被维护在描述输入内容的一个或多个聚类及邻域的记录数据库中。

本发明三个阶段的职责由两个主要组件完成：分析组件、和聚类组件。分析组件负责把内容中所含数据缩减成为在该内容中出现的最重要 N 个权标 (token) 的相关性排序列表。分析在输入阶段 110 中进行。内容可以是文本、视频、音频等等。用于缩减的具体技术取决于内容的类型。聚类组件负责将内容分配到聚类中，并负责归组相似聚类。

输入阶段 110

考虑一种文本内容的分析组件。这种分析的一个应用是结合新闻分析域来使用，以向用户提供一个相关新闻文章清单。该分析组件在字—权标级上工作。

“权标”被认为是代表单一概念的一个或多个字。例如，“球”、“爆炸”、“航天飞机”都是指单一概念。一篇给定文章中的文本用以下方式被缩减成权标的相关性排序阵列：

- 注意每个权标在文本里出现的位置和频率。
- 如果权标出现在标题中，那么向该权标的字计数增加 T 出现次 (occurrence)，且该权标的起始位置设置为零。
- 权标 i 的相关性设置为 $e^{(-\alpha * P_{0i})} * N_i * R_i$ ，其中 α 为权标相似性的衰变率，它是与文本起始点距离的函数， P_{0i} 是权标 i 第一次出现在文章中的位置， N_i 是权标 i 在给定文章中的出现次数而 R_i 是权标 i 文档频率倒数的对数，其中频率可用该字出现在文档中的次数除以在代表性文集中所含的文档数。通常，用近期文章的集合作为代表性文集。
- 每条内容的最高 K 个权标—相关性对被发送到聚类引擎中。相关性之和被归一化为 1.0 (L1 归一)。

考虑以下短新闻故事；

Shuttle Disintegrates on Re-Entry

At 12:08 this afternoon, NASA announced that the space shuttle Discovery disintegrated as it was re-entering the earth's atmosphere. Witnesses in a remote area of Texas saw remnants of what was believed to be the doomed shuttle disintegrate in long white plumes that extended across the blue sky. Discovery's mission had been plagued with difficulties for the entire two weeks it had been in space. Just

yesterday, Commander Smith of Discovery was quoted by communications specialists at NASA's mission control as saying "this has been a tough ride and on behalf of my crew I want to thank you guys for all the help you have given us." Smith along with the seven other mission specialists are presumed dead and NASA has confirmed that it has contacted the families of all eight crew members. In the post 9/11 sensitivity to terrorist activity, NASA made it clear there was no evidence of tampering or attacks on the space craft. President Bush has already scheduled a news conference tomorrow evening (Wednesday 8:00PM EST) and it is speculated that by the time of the new conference NASA may have more information regarding the cause of this disaster.

考虑此故事中的字“shuttle”。在示例性系统中，字中字母的大小写被认为是不重要的，因此 Shuttle 和 shuttle 是一样的。注意，除了在文章正文中找到的正常频率次数外，术语 Shuttle 在标题中的出现增加了频率次数。假设 $T=2$ 。在此新闻故事中，术语 shuttle 出现六次（真实的四次和因为出现在标题中的两次）。另外，因为“shuttle”出现在标题中，其起始位置设置为零，使得 $e^{(-\alpha * P_{0i})}$ 等于 1。

当收到文档时，根据其来源它已被归类为“新闻”文档。其它类别为“运动”、“娱乐”、“旅游”等或这些类别下的小类别。运行于后端服务器上的预处理器软件维护着新闻类文档的数据库。该数据库具有当前在目录中的新闻文档中所包含的数以千计的字的文档频率数据。对于近来发生的新闻故事中某些字的频率可能起始时较低，而随时间流逝增长。对于新近收到的新闻故事，Shuttle 在所有目录中的新闻文档中的出现率低，从而其倒数的对数值高。基于这些因素，相当肯定的是此故事中 K 个最高相关性权标之一为字“shuttle”。

对于一给定新闻文章，这种分析的结果是一尺寸为 K 的（权标，相关性）对的向量，其中 K 是向量中这种对的数量。形式为 $\vec{X} = (W_i, X_i)$ 的向量形成。相关性最高的 K 个权标或字 W 分别对该向量有所贡献，而权标的值 X 由以上相关性公式确定。该向量的值被归一化，从而 N 项或 N 个权标的相关性因子值 X_i 的和为 1.0。基于粗略检查，最高 N 项中出现的其它权标为“NASA”、“Space”和“Discovery”。

预处理器的聚类组件采用了一种改进的“模糊 K 中数”聚类技术(fuzzy

K means clustering)。K 中数聚类是一种基于数据的相似性将数据归组的著名方法。K 中数聚类在例如 2000 年 1 月授权给 Fayyad 等人的 6,012,058 号美国专利中有描述。该专利在此引入作为参考。

聚类是在包括数据挖掘、统计学数据分析、和数据压缩在内的各种技术的重要应用领域。很多流行的聚类技术使用基本 K 中数法，其中数据聚类被初始化而且数据以硬方式添加到初始化聚类中，也就是说，每个数据项属于且仅属于一个聚类。根据该示例性实施例，一个信息内容项或文档可属于一个以上聚类。

根据示例性实施例，每个聚类由以下信息来表示：

- 以用于由 web 服务器发布的形式分配给聚类中的内容项（例如文本文档）。
- 代表该聚类的 K 个最高权标—相关性对（归一化使相关性因子之和为 1.0）。这些组对被称为聚类中数。
- 该聚类最高的 L 个所需权标。为了使一条内容，比如一个文本文档，能添加到该聚类中去，该内容必须在其 K 个最高权标—相关性对中包含所有的 L 个所需权标。L 是比 K 小的数字，且可以是对所有聚类而言相同的一个固定预设定值。在一示例性实施例中，L 的值被设定为取决于文档语言和类别的值。值 3 就曾在一系统中成功用于分类新闻文档。

当首次从自源 4 收到一个文档，该文档被评估并分配到一个现有聚类中。如果无法将其分配到一个聚类中，则它形成自己的包含一个文档的聚类。当给定内容的权标—相关性对进入聚类组件时，遵从图 3 所示过程 110：

- 检索当前内容可能会属于其的所有候选聚类，即其 L 个所需权标在当前内容中出现的所有聚类。
- 计算该内容和候选聚类之间的权标—相关性对的内积。内积也称为两个向量的点积。这一计算使后端服务器能作出内容是否加入聚类的决定（112）。
- 在上述内积超过给定阈值 T 时，内容添加到每个聚类中（114）。在一示例性系统中，如果内积超过一经验确定值，则做出添加的决定。此值基于在组或聚类中维持适当数量文档的同时，文档类别必须集中到什么程度来确定，且此值可动态变化。

- 当一条信息被添加到聚类中时，聚类中数通过对分配到该聚类中所有内容的所有权标—相关性对取和重新进行计算；选择 K 个最高权标—相关性对；将结果的和归一为 1.0。来自 L 个最高权标—相关性对的权标成为聚类的所需权标。
- 如果没有聚类匹配输入内容，则为该内容创建一个新的聚类。该单个文档聚类中的中数是 K 个权标—相关性对。来自 L 个最高权标—相关性对的权标成为该聚类的所需权标。

合并阶段 140

在周期性间隔上（在本发明的新闻收集实施例中为每隔几分钟），在所谓发布周期期间，对聚类进行检查（图 4）以了解是否应将先前独立的聚类归组成一个邻域（neighborhood）。这个过程在“合并阶段”140 期间发生。用于发现输入内容聚类的相同过程在此阶段使用，即聚类中数的内积被确定。但是，在找到匹配时所发生的就不相同了。

当两个或多个聚类被合并时，聚类中的中数（归一化的权标/相关性对）没有改变。然而，导致合并的聚类被选为父聚类。这个“多个聚类的聚类”称为邻域。一个邻域可包含一个或多个聚类。

如果聚类 B 合并到聚类 A 中，即聚类 B 现在以聚类 A 为父，则所有以聚类 B 为父的聚类现在以聚类 A 为父，并属于 A 的同一邻域。

当聚类成员改变，即新聚类加入邻域，或老邻域过期，则原来分配到该邻域的内容（例如文本文档）被分配到空邻域（在内容首次进入聚类系统时它也被分配到空邻域）。在上例中，当聚类 B 合并到聚类 A 中，B 和 A 的所有内容被分配到空邻域。类似地，如果由于时间流逝邻域过期，该邻域所有聚类的内容被分配到空邻域中。邻域过期意味着其相关性下降到低于阈值，如下所述。

考虑如图 6A 所示的三个聚类 142, 144, 146。这三个聚类包含具有相似内容的文档。但是，这三个文档经其中数内积衡量并未相似到可以称为一个邻域。换一种说法，聚类的内积没有超过所确立的将聚类归组的阈值。这一阈值通常与为把文档分类到聚类而确立的阈值相同。现假定聚类 148 被创建，且彼此相似的文档被添加到该聚类中。在图 4 流程图的执行过程

中，聚类 148 被选择（141），且该过程检查邻居 142, 144, 146（及更多远距离聚类，注意，当其内积较小时聚类被视作远距离）（143），确定这四个聚类 142, 144, 146, 148 足够相似以保证邻域 150 的创建（图 6B）。如在图 4 流程图中所见，其评估导致邻域建立的聚类 148 被选作邻域 150 的父聚类。

分配阶段 160

尽管内容在输入阶段被分配到多个聚类，该内容（文档）还未被分配到一邻域中。

在每个发布周期中，有一个分配阶段 160。该分配阶段确定一条内容应从属于哪个邻域。回想一下，在输入阶段时，一条内容可被分配到多个聚类中去。考虑图 7 的情况。文档 165 被分配到聚类 148 及第二个邻域中的第二个聚类 170。因此，有时就是这种情形，在合并合并阶段 140 中将内容或文档分配到其中的聚类分别属于独立邻域，诸如邻域 150、175。因此在发布给用户时，有必要选择该内容属于哪个邻域。

在分配阶段 160 中，所有分配给空邻域的内容被选中。对于每一条内容，计算该内容和该内容被分配的每个聚类的内积。有最高内容一聚类内积的聚类所属的邻域是该内容被分配的邻域（见图 5）。回想一下，有两种类型的聚类，一种类型是内容可以归属的低级聚类。实际上内容可属于若干个这种低级聚类。第二种类型是作为“多个聚类的聚类”的邻域。一条内容仅可属于一个邻域。当内容被带入系统、或当邻域过期、或当邻域被拆分或组合，邻域的成员被设置为空。这样，系统知道哪些内容需要被重新分配，即先前从属于已改变的邻域之一的内容需要进行重新分配。系统使用以上过程来了解该内容从属于哪个聚类，以确定该内容应重新分配到哪个邻域。在过期邻域情形中，重新分配是必须的。在被改变邻域的情形中，对邻域的改变可导致在一个邻域中的内容要重新进行分配。

用户或客户 8 希望以文档的邻域形式从 web 服务器 7 得到信息。在典型实例中，文档的这些邻域显示在网络浏览器上，诸如互联网浏览器（IE）或众多其它适于在通过公司内部互联网或诸如因特网的广域网而网络化的计算机上作文档显示的众多浏览器之一。在图 2 中，网络服务器 7 和客

户 8 之间的信息交换通过请求进行。服务器通过发布最相关邻域中所包含的众多文档来对这个请求作出响应。

客户 8 可提供给服务器 7 关于该特定客户的翔实信息。例如，如果该客户用其 .Net 密码登录他或她的计算机。从客户 8 处可得到的信息包括客户过去的浏览器行为。更具体地，所提供的信息是涉及不同类型文档的各种链接的鼠标点击形式的过去行为。因此，如果客户是体育迷，文档的类别可能都涉及“棒球”。如果客户是投资家，文档类别绝大多数是“股市新闻”相关类的。如果用户未以网络服务器借以标识具体过去行为的方式登录他或她的计算机，则客户请求即以一般方式对待，且被赋予所有这种一般用户的过去行为。

其它信息也可自动提供给网络服务器 7。请求的源可被编码为客户唯一互联网地址的一部分。如果这样，客户的独特语言和国别也可作为与该请求一并提供的信息的一部分。该请求包括具体查询也是可能的。因而，该请求可包括一的直接指示：客户对近期涉及航天飞机事件的所有文章感兴趣。这个请求可能会导致网络服务器向用户发布以上所摘录的那则样本新闻。基于网络服务器 7 可得到的信息，服务器 7 通过发布被判断为与该请求最为相关的邻域来响应该请求。网络服务器维护着每个文档邻域的相关性因子。

邻域的相关性会随着时间的改变。旧新闻对搜寻有关当前事件信息的人们而言，不会引起多少兴趣。然而，旧文档会与对涉及某一事件、地方或人物的历史或事件年表感兴趣的人特别相关。为确定邻域的相关数或相关因子，必须一开始就考虑文章或文档的相关性。

文章的相关性 A_r 如下：

$$A_r = (A \cdot N) * e^{(-t/t_a)} * ICB_a * PR_a$$

A 是代表文章 A 关键字的向量，其中每一元素对应于一个字，元素的大小对应于字的相关性。

N 是代表该文章父邻域的向量。此向量基于组成该邻域的所有文档的中数。A 和 N 间的内积（写作 $A \cdot N$ ）是 A 和 N 间的相似性度量。

t_a 是文章衰变率，为时间 t 的函数。

ICB_a 是奖励从代表性市场向用户显示的文章的因子。例如，如果文章是向英国的用户或客户 8 发布，而文章的发布者或源 4 也来自英国，那么 ICB_a 的值将设为大值，即 10.0，否则， ICB_a 通常取 1.0。

PR_a 是奖励来自重要源的文章的因子。该值代表发布者的值。例如，来自 BBC 的文章的点击率比随机选择的发布者平均点击率高两倍。因此，来自 BBC 的文章的 PR_a 值将为 2.0。

文章的相关性在以下所示用来确定文档邻域相关性的计算中使用。

进入系统的某些文章因为缺乏中心而从未获发布。这些文档可能是关于多种话题的。文章的中心由文章中最相关关键字的相关性值给出。如果此值过低，它表示对文章是围绕单个主题的假设可能是假的。因此，邻域不会考虑包含该文章。当前，使用的值为 0.1。接近 0.05 的值表示根本不存在必需的中心，其它关键词关联性也是 0.05，因此该文章没有中心。

另外，两天后，因为缺乏中心而未被分配到邻域中的文章被丢弃。它将永远不会被发布。

邻域相关性 N_r 如下：

$$N_r = \sum A_r * e^{-l_a} * S_{r_n}$$

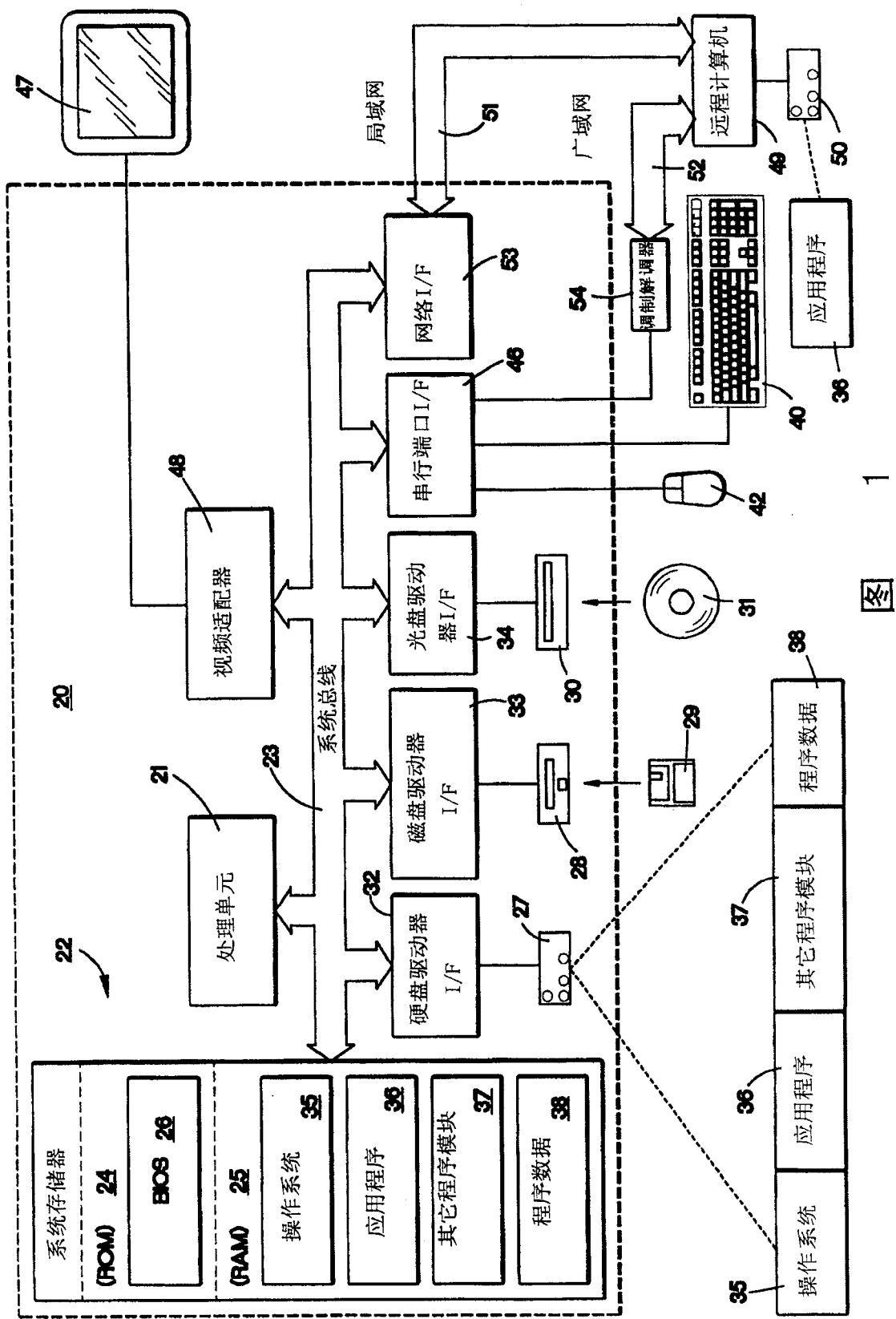
A_r 的和是基于以邻域 N 为父的所有文章求和。

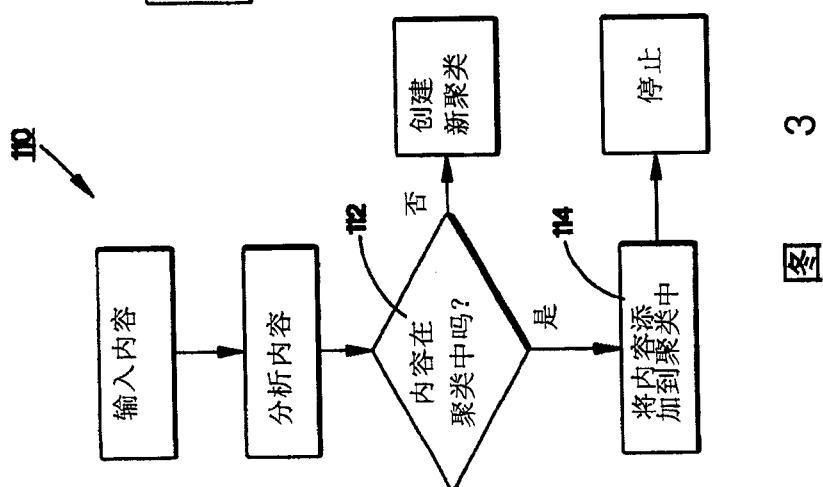
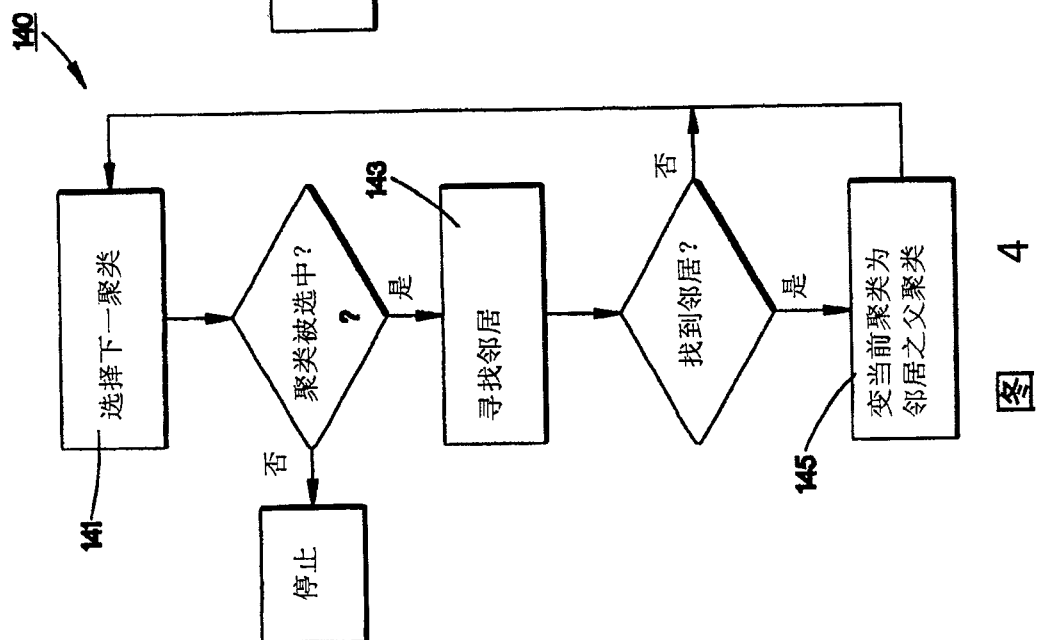
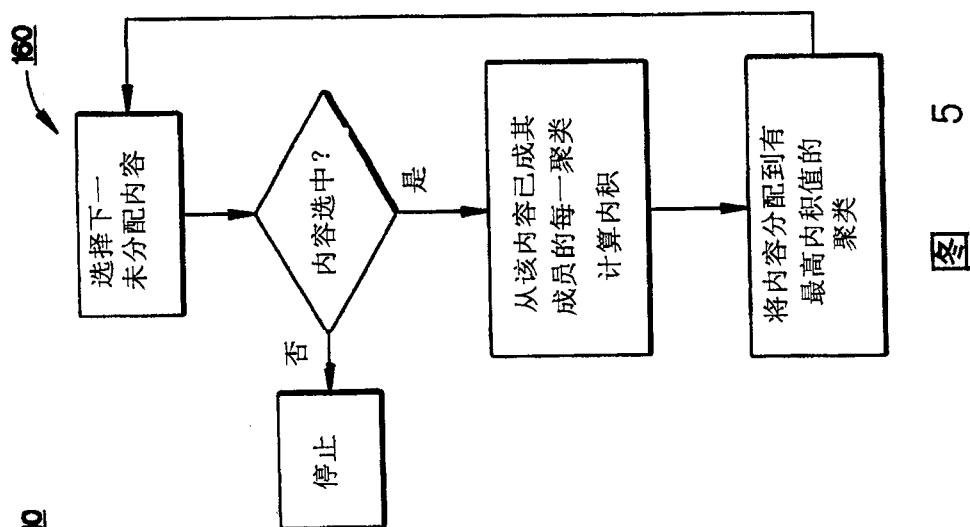
l_a 是邻域衰变率，为时间 t 的函数。

S_{r_n} 是从其它类别文档对邻域加权的因子。例如，体育新闻的分级可有 1.0 的值，而世界新闻的分级可能有 10.0 的值。 S_{r_n} 因子可以是邻域默认的，或是基于经特定客户过去的行为而确定的特定用户或客户兴趣而赋值的。

一旦邻域的相关性随时间流逝降到阈值以下，或取决于网络服务器接收的用户请求，该邻域不再通过相关性测试，并不再因回应请求而返回。

尽管本发明进行了一定程度的详细描述，本发明旨在包括所有在所附权利要求的精神或范围内的更改和变化。





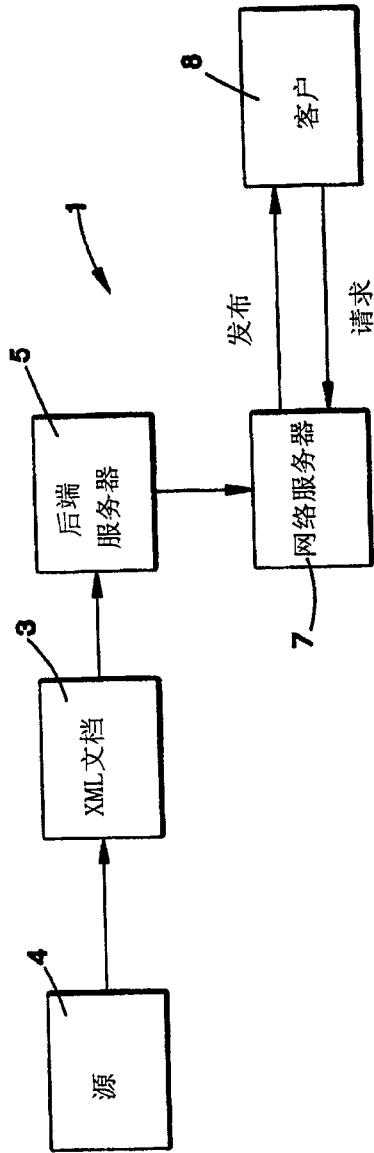


图 2

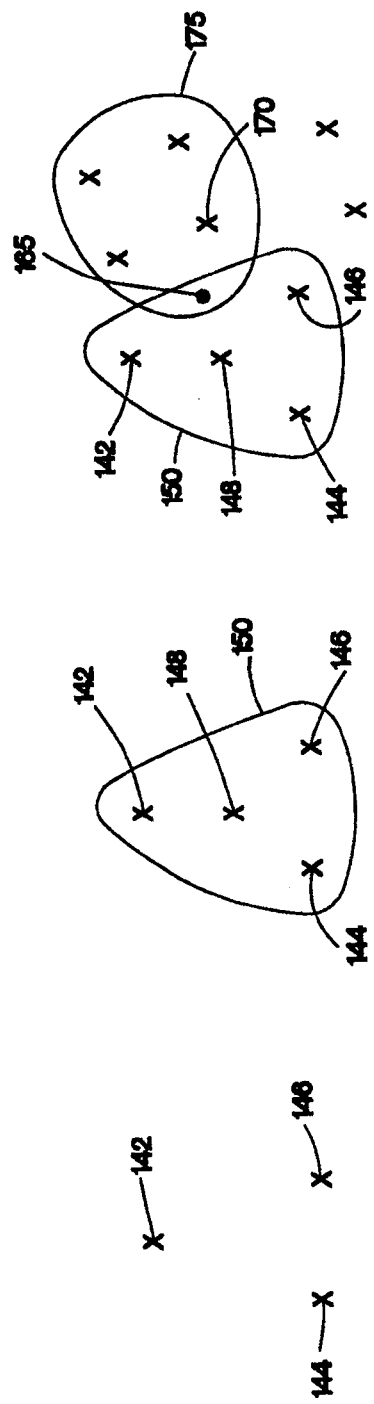


图 6A

图 6B

图 7