[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 200510056187.8

[45] 授权公告日 2008 年 12 月 31 日

[11] 授权公告号 CN 100447782C

[22] 申请日 2005.3.22
[21] 申请号 200510056187.8
[30] 优先权
[32] 2004.3.22 [33] US [31] 10/805,805
[73] 专利权人 微软公司
地址 美国华盛顿州
[72] 发明人 M·S·马纳塞
[56] 参考文献
US6269362B1 2001.7.31
US6349296B1 2002.2.19
CN1461142A 2003.12.10
审查员 梁心新

[74] 专利代理机构 上海专利商标事务所有限公司
代理人 顾嘉运

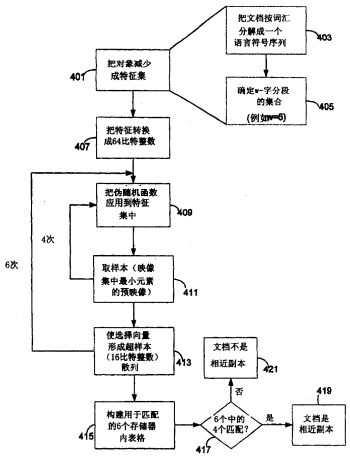
权利要求书 1 页 说明书 8 页 附图 4 页

[54] 发明名称

用于确定搜索引擎查询结果中相近副本条目的分组方法

[57] 摘要

本发明提供一种方法，它通过对现有方法的这样一种更改：降低对每个对象存储器的要求而避免错误的检测却与现有方法几乎一样好，来在对象集合中检测相似对象。该更改包括(i)把特征的 k 个样本组合成 s 个超样本，k 值从现有方法中使用的相应值减少而来；(ii)以 b 比特精确度记录每个超样本，b 值从现有方法中使用的相应值减少而来；以及(iii)需要 l 个匹配超样本以便作出两个对象足够相似的结论，l 值比现有方法中所需相应值大。本发明的一个应用与网络搜索引擎查询服务相关联以确定是相近副本文档的查询结果聚类。



1. 一种用于确定搜索引擎查询结果中相近副本条目的分组的方法，其特征在于，包括，对于进行比较的两个条目中的每一个：

把少于 14 个的特征样本组合成 6 个超样本之一；

以少于 64 比特精确度记录 6 个超样本中的每个超样本；以及

需要所述 6 个超样本中的多于 2 个匹配超样本。

2. 如权利要求 1 所述的方法，其特征在于，还包括在每个聚类中选择一单个文档报告。

3. 如权利要求 2 所述的方法，其特征在于，选择所述一个文档是通过排序函数。

4. 一种用于确定搜索引擎查询结果中相近副本条目的分组的方法，其特征在于，包括，对于进行比较的两个条目中的每一个：

把 4 个特征样本组合成 6 个超样本之一；

以 16 比特精确度记录每个超样本；以及

需要所述 6 个超样本中的 4 个匹配。

5. 如权利要求 4 所述的方法，其特征在于，还包括在每个聚类中选择一单个文档报告。

6. 如权利要求 5 所述的方法，其特征在于，选择所述一单个文档是通过排序函数。

7. 一种用于确定搜索引擎查询结果中相近副本条目的分组的方法，其特征在于，包括，对于进行比较的两个条目中的每一个：

把 4 个特征样本组合成 7 个超样本之一；

以 16 比特精确度记录每个超样本；以及

需要所述 7 个超样本中的 5 个匹配。

8. 如权利要求 7 所述的方法，其特征在于，还包括在每个聚类中选择一个文档报告。

9. 如权利要求 8 所述的方法，其特征在于，选择所述一单个文档是通过排序函数。

用于确定搜索引擎查询结果中相近副本条目的分组方法

技术领域

本发明涉及在大型集合中标识类似数据对象，尤其涉及在诸如万维网的极大文档全集中标识相近副本。

背景技术

大型文档集合通常包括许多彼此相同或者几乎相同的文档。确定两个数字编码的文档是否逐比特地都相同是容易的（例如使用散列技术）。然而，快速标识几乎或有效相同的文档则更具挑战性，并且在许多环境中是更为有用的任务。万维网是一个极大的文档集合。自诞生以来 Web 呈级数增长，目前 Web 索引约包括约 50 亿网页（据估计静态网页达 200 亿网页），其中很大一部分是副本或相近副本。诸如网络爬虫和搜索引擎的应用程序受益于检测相近副本的能力。例如，需要使这种应用程序忽略大部分的副本和相近副本，或者过滤查询结果从而把相似的文档组合在一起。

已经开发了“shingling”（搭迭法）或“shingleprinting” （搭迭拓印法）技术以解决在大型集合中找出相似对象的问题。这种技术的各个方面在以下专利文献中描述：Broder 等人在 1996 年 6 月 18 日提交的专利号为 5,909,677 的美国专利“用于确定文档相似性的方法”（“Method for Determining the Resemblance of Documents”）；Broder 在 1997 年 9 月 15 日提交的专利号为 5,974,481 的美国专利“用于估算指纹冲突概率的方法”（“Method for Estimating the Probability of Collisions of Fingerprints”）；Broder 等人在 1997 年 12 月 19 日提交的专利号为 6,269,362 的美国专利“用于通过比较生成摘要来监视网页的系统和方法”（“System and Method for Monitoring Web Pages by Comparing Generated Abstracts”），其中本申请的发明人是共同发明人；Broder 等人在 1998 年 3 月 26 日提交的专利号为 6,119,124 的美国专利“用于聚类较为相像数据对象的方法”（“Method for Clustering Closely Resembling Data Objects”），其中本申请的发明人是共同发明人；Broder 等人在 2000 年 8 月 21 日提交的专利号为 6,349,296 的美国专利“用于聚类较为相像数据对象的方法”（“Method for Clustering Closely Resembling Data Objects”），其中本申请的发明人

是共同发明人；以及 Manasse 等人在 2001 年 9 月 21 日提交 2003 年 3 月 27 日发表的专利申请号为 09/960,583 的美国专利申请“用于在生物统计数据库中确定相似身份的系统和方法”（“System and Method for Determining Likely Identity in a Biometric Database”），其中本申请的发明人是共同发明人。还可参见 Broder 在 1997 Proc. Compression & Complexity of Sequences 21-29 (IEEE 1998)中的文章“关于文档的相似性和包容性”（“On the Resemblance and Containment of Documents”）；Broder、Glassman、Manasse、和 Zweig, 在 Proc. 6th Intl. World Wide Web Conf. 391-404 (1997 年 7 月)中的“Web 的综合聚类”（“Syntactic Clustering of the Web,”）；Manasse 的“在大集合中快速找到相似物”（“Finding Similar Things Quickly in Large Collections”）<http://research.microsoft.com/research/sv/PageTurner/similarity.htm> (2004)。这些专利和非专利文献的每一篇都在此引入作为参考。

在“搭迭”方法中，文档被缩减为足以代表文档的特征集从而两个极相似的文档将共享大量特征。对于文本内容文档，已证明把重叠的连续 w 个字的分段（其“w 个沙砾”）集合提取为特征是有用的，其中 w 是固定的数字。设 D_1 和 D_2 为文档，且 F_1 和 F_2 为其相应的特征集，我们定义 D_1 和 D_2 的相似度为特征集的 Jaccard 系数（即，两个文档中共同特征的数量除以两个文档中特征的总数量）：

$$Sim(D_1, D_2) = \frac{|F_1 \cap F_2|}{|F_1 \cup F_2|}$$

这给出了 0 与 1 之间的数字；两个基本上相等的文档的相似度将是接近于 1 的数字，而大多数不相似文档对的相似度则是接近于零的数字。应当注意，用于在大型集合中检测有效相同条目的搭迭法技术不受文本全集的限制。可把搭迭法应用到任何类型的诸如录音或视频图像等的数据对象集合中，对于它们提取一个代表特征集是可能的。

从每个文档中提取的特征数量可能相当多（像文档中的字数一样多）。如果假设文档集本身很大（可能象 Web 情形中达数十亿），精确计算相似度值并执行成对比较是该集合大小的平方，那就过于昂贵。因此将把相似度加以近似以便把问题减少到可管理大小。

近似涉及以保留文档相似性的方式来采样每个文档的特征集。从原理而言，可使用从特征空间到比所有特征集更大的良好排序的集合的随机一一对应函数。根据良好排序性，在随机函数下有特征集映像的最小元素。最小元素的预先映像被取作选定的样本特征。因为所有函数的概率是相等的，这是可行的。任何集的元素都

同样可能被映射成最小元素，而且当从这两个集中选择时，最小元素均匀地在并集中选取。

必须改动前述方案以便它可在实践中实现。首先，为了统一地进行选取，使映像集成为有限的整数集是方便的。如果特征集是无界的，就难以得到对有限集的一一对应函数。使用优选的散列函数（最好是 Rabin 指纹）以使每个特征形成固定比特数的散列，可选择足够大的集合从而该集合中的冲突概率小到难以察觉。其次，不选择真正的随机函数，而是从较小的、易参数化的函数集中选择该函数，其中选定函数被证明能足够接近正确概率。通常，使用线性同余置换以及 Rabin 指纹的组合，即使它并未证明是正确的。

该技术提供了用于从每个特征组 F_i 中选择一个特征 f_i 的机制，从而 $\text{Prob}(f_i = f_j) = \text{Sim}(D_i, D_j)$。该选择机制提供了相似性的无偏点估计量。选择了某些数量的选择器 r。对于每个文档 D_i，使用 D_i 上每个选择器一次来计算 $f_i^1,...,f_i^r$。以某些预处理为代价，这使每个条目的数据存储降为一常数，并使集的比较降为与向量中的条目相匹配。

通过运行多个独立的选择机制，两个文档相似性比率的估算可通过对选择向量中的匹配进行计数来获取。如果 $p = \text{Sim}(D, E)$ 则 D 和 E 向量中的每个条目以概率 p 相匹配。在一行中匹配 k 个条目的概率为 p^k。通过对 k 个条目的非交叠游长进行散列以形成从足够大空间中选取的单个整数可以压缩向量，其中在散列值中冲突的概率可忽略而存储需要按系数 k 来减少。如果有 s 个长度为 k 的组（“超样本”），则有一个或多个超样本相匹配的概率为 $1-(1-p^k)^s$，而有两个或多个超样本相匹配的概率为 $1-(1-p^k)^s - s(1-p^k)^{s-1}$。

在有关 Alta Vista 搜索引擎的现有技术中，提取每个文档文本的归一化版本的“6 搭迭”（6-shingling）为特征集。特征被表示为 64 比特的整数。把使用线性同余置换的计数应用到每个 64 比特的整数中，产生一个新的 64 比特整数集，并选择该新集中最小值的预映像为样本。取 84 个样本，分成 6 个超样本（其中每个由 14 个样本组成）。因而，对于参数 k 和 s，使用的值为 k＝14 和 s＝6。之所以选择这些参数是因为相近副本文档的期望相似度阈值为.95。两个文档之间 14 个样本匹配的概率等于文档相似度的 14 次方，从而如果文档为相近副本，其概率为 $.95^{14}$，即约为 1/2。有了 6 组 14 个样本，6 组中至少 2 组匹配是可能的，6 组中少于 2 组匹配是不可能的。因而，为了确定文档可能是相近副本，要求 6 个超样本中的 2 个必须匹配。根据期望阈值，在标识相近副本条目中现有技术在实践中是有效的。

在现有技术中发现匹配过程可被简化为对每个条目的少量散列表查找。组中 k 个样本被压缩成 64 比特的整数。如下所述，每个超样本以 64 比特的精确度进行记录以避免非相似文档的意外巧合。然后把 s 个 64 比特整数的所有可能对插入散列表。如果例如 s=6，找出匹配至少两次运行的条目只需要 $\binom{s}{2}=15$ 查找，所以使用 15 个散列表。

现有技术集中在离线过程的使用上，因此节约主存储器不是关注重点。例如在前面段落中描述的散列优化很适于离线处理。然而，该技术对每个文档都存储的要求对于所有文档都在运行时执行的搜索引擎而言是不可接受的。在 50 亿文档的集合中，存储 6 个值就需要 240 兆比特的存储器覆盖区以及存储散列表的另外 520 兆比特。这在搜索执行中是不实用的，并限制了索引的构建。例如，搜索引擎可不在全文档集合中执行预处理过程，并可递增地建立其索引。搜索引擎希望的是在在线过程中确定要报告的哪些查询结果是相近副本，从而使报告缩减为每个相近副本聚类一个文档，其中使用排序函数来选择该文档，并且使反应最迅速的文档能被动态选择。

发明内容

根据某些实施例，本发明提供用于在这种对象集合中检测相似对象的方法。该方法包括对现有方法的这样一种更改：降低了对每个对象存储器的要求而避免错误的检测却与现有方法几乎一样好。该更改包括 (i) 把特征的 k 个样本合并成 s 个超样本，k 值从现有方法中使用的相应值减小；(ii) 以 b 比特精确度记录每个超样本，b 值从现有方法中使用的相应值减小；以及 (iii) 需要 l 个匹配超样本以便作出两个对象足够相似的结论，l 值比现有方法中所需相应值大。现有方法中的 l 值可以是例如 s、s-l、或者 s-2。

根据一实施例，k=4，从现有方法的 k=14 减小而来。根据一实施例，b=16，从现有方法的 b=64 减小而来，且 6 个超样本中的 4 个需要匹配，从现有方法 6 个中的 2 个增大而来。在另一实施例中，7 个超样本中的 5 个需要匹配，替换现有方法 6 个中的 2 个。在又一实施例中，所有超样本以及因而所有样本都需要匹配。

在本发明的某些实施例中，此方法结合网络搜索引擎查询服务使用，以确定相近副本文档的查询结果聚类。一旦使用该方法找到了这些聚类，例如根据排序排序函数从每个聚类中选择一个文档，并向查询提交者报告这个文档（以及相似文档的指针）而不是这个聚类。

可以预期，通过软件在计算机上的执行本发明可整个或部分地实现。

参阅附图阅读以下说明书，本发明的其它特征将变得显然。

附图说明

图 1 是显示使用搭选拓印法根据对若干不同参数，包括在本发明诸实施例中可用参数的选择来接受文档的概率的图表；

图 2 是错误接受率的图表（基于对图 1 所绘的值取对数）；

图 3 是显示在本发明实施例中由 web 搜索引擎查询服务所执行的步骤的流程图；

图 4 显示根据本发明实施例用于确定文档是否是相近副本的方法的示例性步骤。

具体实施方式

在以下说明书中，将描述本发明的某些实施例。为了进行解释，陈述特定的配置和细节以便提供对诸实施例的理解。然而，对本领域技术人员显而易见的是，没有这些特定细节也可实践本发明。此外，作为本发明固有部分并对本领域技术人员是基础的众所周知的特征，一般被略去或简化以便不混肴所描述的实施例。

本发明提供了一种标识大集合中相近副本条目的技术，该技术与现有技术方法相似但降低了每个文档的存储器要求。这使该技术在这些情景中能够实践并有用：不想或不能用离线处理、或者存在特定的存储器限制，如背景技术章节最后所描述的搜索引擎示例。

图 1 是显示当以不同参数值应用搭选拓印法时接受文档 101 的概率的图表，其中可能相似度 103 显示在 x 轴上。图例 105 中的标记指示可用的超样本总数量中需要匹配的超样本数量。Lk 标记指示每个超样本组合 k 个样本。最右边的曲线 107（标为“6 个中的 2 个，L14”）绘出了应用在现有技术中参数的结果，如上在背景技术章节中所述。该曲线较为接近地近似了阶跃函数，示出了现有方法在标识 95% 相似度的相近副本的可能性的优点。

如上在背景技术章节所述，在 web 搜索引擎查询结果中关于相近副本的现有技术中，需要其中 2 个匹配的 6 个超文本中的每一个，长度为 64 比特。对于组合所需的匹配需要 128 比特的精确度，以便确保防止由生日悖论引起的错误冲突，而根据生日悖论，在 n 个条目的集合中可能发生的冲突在$\sqrt{n}$个条目之后。如果仅有

64 比特精确度，在检查了 2^{32} 对条目的相似性之后，过程将易受生日悖论的影响。对于一个象整个 Web（如上所述数十亿的文档）一样大的文档集合中，实际上错误冲突将是必然的。因此，有了 64 比特的超样本，6 个中的 2 个匹配是必须的，且所有 6 个超样本的存储也是必须的。对于 50 亿个文档，显然这种存储要求在许多 web 搜索相关应用程序中是巨大并不实用的。

本发明体现了一种洞察力，即现有技术中鉴别的积累幂数以存储要求的大幅缩减来充分近似。将更少的样本组合为每个超样本，并需要超样本集中更大量的匹配。在某些实施例中，仅有一个或两个超样本需要匹配。在另一实施例中，所有超样本因而所有样本都需要匹配。在比特精确度上每个超样本都缩减，然而仍然避免生日悖论冲突。除了提供用于大集合的新的特定技术以外，本发明提供一种更改现有特定技术的方法，用于在特定的大集合中检测相似性以便获取存储器使用的节约。

在一实施例中，用于确定文档相近副本的技术结合处理用户查询结果的 web 搜索引擎查询服务使用，使得每个相近副本集中仅有一个向用户报告。图 3 显示用于这种查询服务的示例性过程。在步骤 301 处理该查询并产生结果集。在步骤 303 应用用于确定是相近副本的结果聚类的方法。在步骤 305，对于每个相近副本的聚类，应用排序排序函数以确定经估算结果。在步骤 307，向用户报告经过滤的查询结果。

在此实施例中，6 个超样本中的 4 个需要匹配而不是现有技术中 6 个中的 2 个。每个超样本长度仅有 16 比特。16 比特是足够的，因为有了 4 个必须的匹配，其中每一个超样本的比特数都为避免错误冲突的概率作贡献。此外，在此实施例中，通常不搜索整个 web；在特定时间仅有一查询的结果处于竞争中，更可能涉及的是数万或数十万、而不是数十亿的众多文档。因此，为了避免生日悖论的影响应该不需要 128 比特（除非搜索引擎用户请求一空查询）；64 比特可能就足够了。每个超样本组合压缩为 16 比特数字的 4 个样本，而不是压缩成 64 比特数字的 14 个样本，因为 6 个中（每个概率为 1/2）4 个匹配的概率相对较小。比特精确度的减小使存储器要求以约为 4 倍降低。进一步的降低是可能的，因为存储器上散列表为查询的返回文档集而构建，所以对于文档全集仅需要 6 个值而不是 15 个。

图 4 显示了根据以上所述本发明实施例用于确定文档是否是相近副本的方法的代表性步骤。在步骤 401，文档或其它对象被缩减为一特征集。对于诸如 Web 上可搜索文档的基于文本文档，可使用搭迭法。在步骤 403，把文档按词汇分解成

一个语言符号序列，其中忽略类似标点符号和大小写的特征。在步骤 405，确定连续的固定字大小的分段集。

在步骤 407，把特征转换成 64 比特的整数。在步骤 409，把伪随机函数应用到经转换的特征集中，而在步骤 411，通过取映像集最小元素的预映像选择样本。执行步骤 409 和 411 k=4 次。在步骤 413，使结果选择向量形成以 16 比特精确度记录的超样本散列。重复该子过程从而产生 6 个超样本。在步骤 415，构建用于匹配的 6 个存储器中表格。如果 6 个超样本的 4 个匹配（步骤 417），确定比较中的文档为相近副本（步骤 419），否则不是相近副本（步骤 421）。步骤 415 和 417 的有效实现更适合用存储 4 个超样本组合的 15 个散列表完成。

请求 s 中 l 个匹配的假阳性率是 $\frac{\binom{s}{l}}{2^{bl}}$，其中每个超样本的比特长度为 b。在 2^{33} 个文档的集合中假阳性的期望数量是：

$$2^{66-bl}\binom{s}{l}$$

对于 l=4，s=6 和 b=16，这导致整个集合中 60 个文档聚类被错误标识为副本；这仅在一次查询的结果集中有两个这样的聚类时才能看到。假设没有一次查询返回 2^{10} 以上的聚类，则在遇到一次错误冲突前可望处理约一万亿次查询。如果搜索引擎所支持的查询率假定约每年 2^{35} 次，则将发生的错误冲突将是每个世纪 2 到 3 次，假设查询服务一直全速运行并假设所有查询都有最大的结果集。

再看图 1，曲线 109（示为“6 个中的 4 个，L4 16 比特”）对应于上述实施例。“16 比特”标识每个超样本以 16 比特精确度记录。可以看到曲线 109 不象曲线 107 那样陡峭。对曲线 107 的一稍好近似由曲线 111 获得（示为“7 个中的 5 个，L4”）（标识 7 个中的 5 个匹配超样本，每个超样本组合 4 个样本）。在前面段落中所述的搜索引擎的可选实施例使用该 7 个中的 5 个方案。尽管未在图中画出，6 个中的 4 个，L5 也是一种好的近似。

参看图 2，所示的是错误接受率 201 的图表。基于取图 1 中所绘值的对数，所绘曲线在此反映了嵌入每个样本中的底层组合学和精确度。该图表一般显示当相似度下降，把一文档对错误标识为相近副本的可能性变得较小。“6 个中的 4 个，L4 16 比特”的曲线 203 的左端在 10^{-18} 变平（锤线 205），表示与 16 比特精确度相关联的对样本的限制。然而，对于上述实施例，具有可忽略的错误冲突概率是令人满意的。

其它变化在本发明的精神内。然而，尽管本发明可接受各种各样的变化和可选构建，其中某说明性实施例在附图中示出并已在说明进行了描述。然而，可以理

解并非旨在把本发明限制在所揭示的特定形式中，相反，旨在包括如在所附权利要求书中定义的本发明精神和范围内的所有变化、可选构建和等效实施方案。

本发明的优选实施例在此进行描述，包括发明者所知用于实践本发明的最佳模式。对于本领域技术人员而言，在阅读了前面的说明书之后，那些优选实施例的变化变得显然。发明者期望本领域技术人员能适当地采用这些变化，且发明者希望本发明能以不同于在此特别描述的其它方式来实践。因此，本发明包括适用于法律所允许的在所附权利要求书中引用的本发明的所有变化和等效方案。此外，本发明包含以上所述所有可能变化的元素的任何组合，除非以其它方式指明或明显与上下文相矛盾。

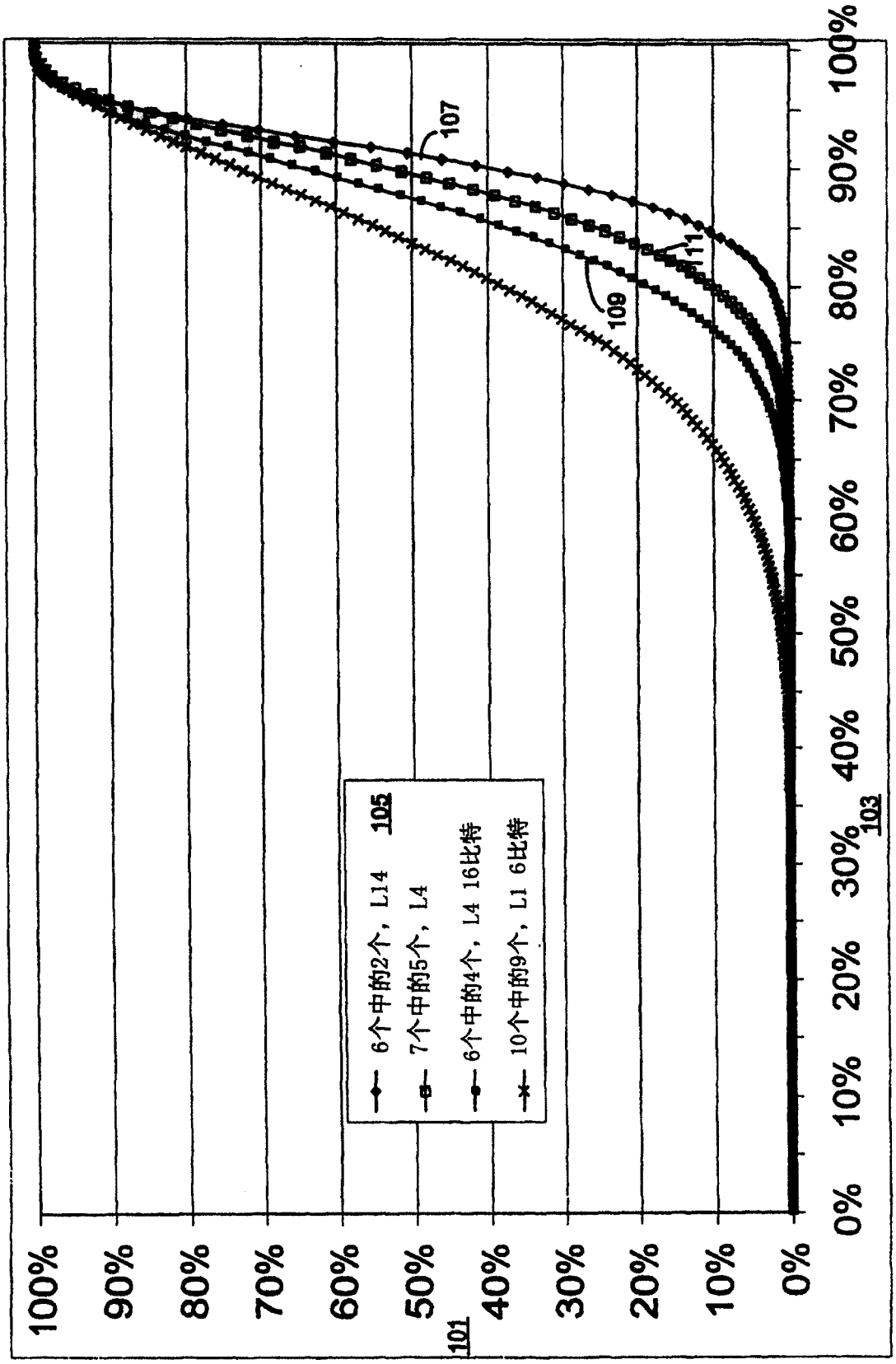

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
101
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
103
6个中的2个，L14
7个中的5个，L4
6个中的4个，L4 16比特
10个中的9个，L1 6比特
105
107
111
109

图 1

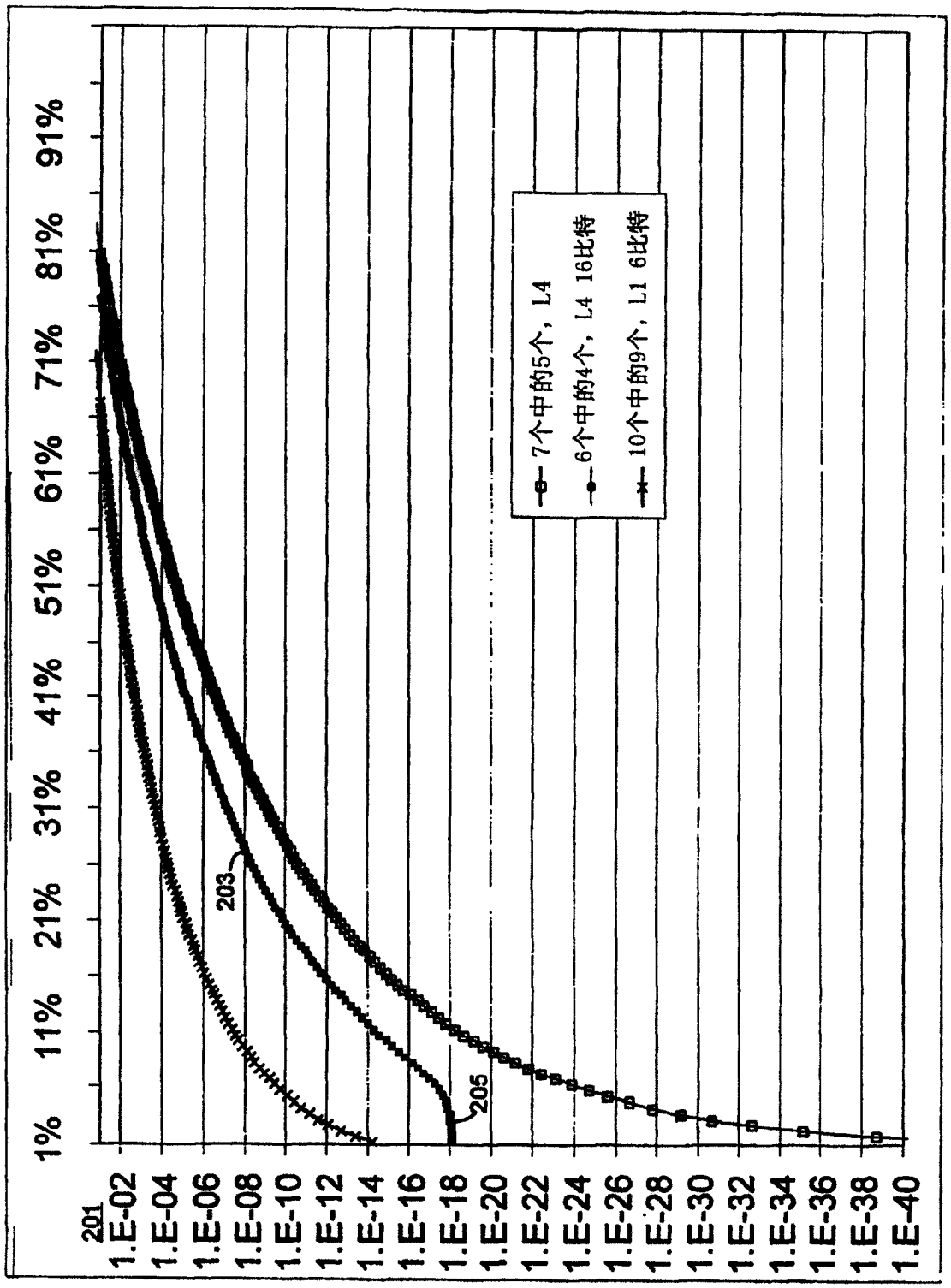

1% 11% 21% 31% 41% 51% 61% 71% 81% 91%
201
1.E-02
1.E-04
1.E-06
1.E-08
1.E-10
1.E-12
1.E-14
1.E-16
1.E-18
1.E-20
1.E-22
1.E-24
1.E-26
1.E-28
1.E-30
1.E-32
1.E-34
1.E-36
1.E-38
1.E-40
203
205
7个中的5个，L4
6个中的4个，L4 16比特
10个中的9个，L1 6比特

图 2

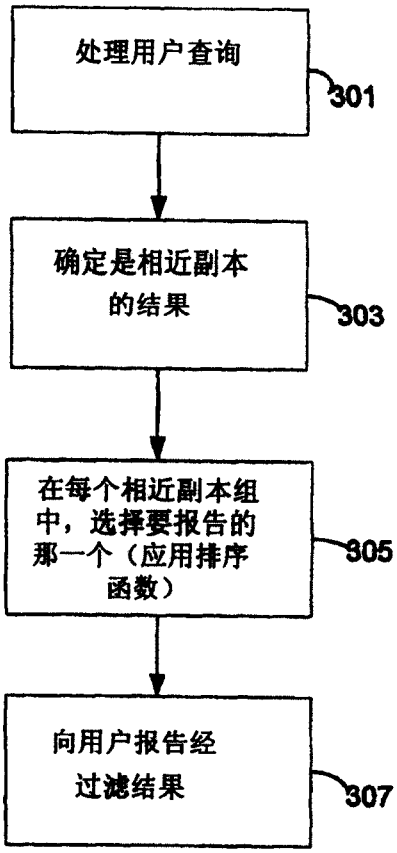

处理用户查询
301
确定是相近副本的结果
303
在每个相近副本组中，选择要报告的那一个（应用排序函数）
305
向用户报告经过滤结果
307

图 3

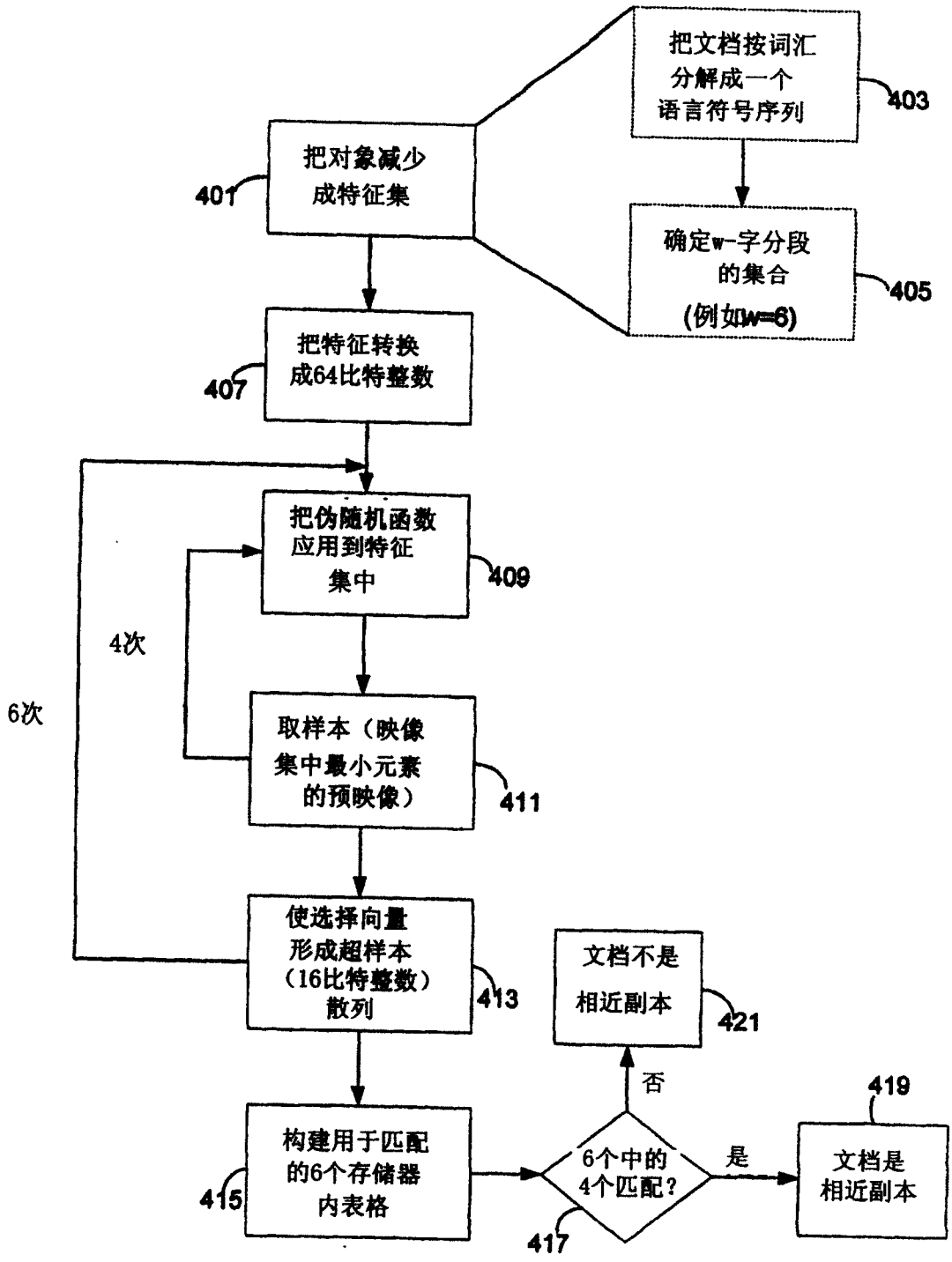

把文档按词汇
分解成一个
语言符号序列
403
把对象减少
成特征集
401
确定w-字分段
的集合
(例如w=6)
405
把特征转换
成64比特整数
407
把伪随机函数
应用到特征
集中
409
4次
6次
取样本（映像
集中最小元素
的预映像）
411
使选择向量
形成超样本
（16比特整数）
散列
413
文档不是
相近副本
421
否
419
构建用于匹配
的6个存储器
内表格
415
6个中的
4个匹配？
417
是
文档是
相近副本

图 4