(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 102226901 B
(45)授权公告日 2014.01.15

(21)申请号 201110200374.4
(22)申请日 2005.07.26
(30)优先权数据
10/900,041 2004.07.26 US
(62)分案原申请数据
200510085371.5 2005.07.26
(73)专利权人 咕果公司
地址 美国加利福尼亚州
(72)发明人 安娜·林恩·帕特森
(74)专利代理机构 北京律盟知识产权代理有限责任公司 11287
代理人 王允方
(51)Int.Cl.
G06F 17/30(2006.01)
(56)对比文件
US 6085186 A,2000.07.04, 说明书第14栏第55-65行,15栏第7-12、35-45、60-65行.
US 6363377 B1,2002.03.26, 全文.
US 2003/0078913 A1,2003.04.24, 全文.
Jones et al..Topic-based browsing within a digital library using keyphrases.《Proceedings of the ACM International Conference on Digital Libraries》.al,1999,114-121.
审查员 马晓宇

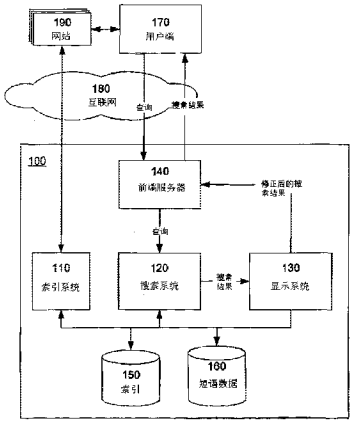
权利要求书1页 说明书23页 附图7页

(54)发明名称

信息检索系统中基于短语的搜索

(57)摘要

本发明涉及信息检索系统中基于短语的搜索。本发明涉及一种信息检索系统,其使用短语来编制索引、检索、组织并描述文献。识别预测文献中存在其它短语的短语。根据文献中所包括的短语来对文献编制索引。同时识别相关短语与扩展短语。识别并使用查询中的短语来检索文献并对文献分等级。同时使用短语来群集搜索结果中的文献、创建文献说明并从搜索结果与索引中去除重复文献。



CN 102226901 B

1. 一种用于响应查询而对包括在搜索结果中的文献分等级的方法，所述查询包含至少一个查询短语，所述方法包含：

针对所述搜索结果中的每一个文献：

存取所述查询的短语的相关短语位向量，其中所述相关短语位向量中的每一位指示所述查询短语的相关短语是否存在，其中所述相关短语位向量是在编制索引的时间预先计算的；

对用于指示所述查询短语的相关短语是否存在的每一位，将与所述位相关联的预定点数添加到所述文献的分数中；及

使用对应于所述文献的文献分数对所述搜索结果中的所述文献排序。

2. 如权利要求 1 所述的方法，其中，在相关短语相对于查询短语的信息增益超过预定阈值的情况下，所述查询短语具有预测相关短语的功能，其中所述信息增益是相关短语和所述查询短语的实际同时出现率和期望同时出现率的比率。

3. 如权利要求 2 所述的方法，当所述相关短语和所述查询短语在所述文献集中的所述实际同时出现率的度量超过所述相关短语和所述查询短语在所述文献集中的所述期望同时出现率时，所述预定阈值被超过，其中，所述相关短语和所述查询短语的所述期望同时出现率是在所述相关短语和所述查询短语是无关短语的情况下所述相关短语和所述查询短语分别在所述文献集中出现的百分比的乘积。

4. 如权利要求 1 所述的方法，其中：

所述相关短语位向量中的每一位与所述查询短语的相关短语相关联；且

对所述位排序，使所述位向量的最高有效位与相对于所述查询短语具有最大信息增益的相关短语相关联，并且最低有效位与相对于所述查询短语具有最小信息增益的相关短语相关联；及

与每一位相关联的所述预定点数量的范围是从与所述最高有效位相关联的最大点数到与最低有效位相关联的最低点数，

其中所述信息增益是相关短语和所述查询短语的实际同时出现率和期望同时出现率的比率。

5. 如权利要求 1 所述的方法，进一步包含存储所述搜索结果中的所述文献。

信息检索系统中基于短语的搜索

[0001] 分案申请的相关信息

[0002] 本申请为发明名称为“信息检索系统中基于短语的搜索”的原中国发明专利申请的分案申请。原申请的申请号为 200510085371.5 ;原申请的申请日为 2005 年 7 月 26 日 ;原发明专利申请案的优先权日为 2004 年 7 月 26 日。

[0003] 相关申请的交叉参考

[0004] 2004 年 7 月 26 日申请的申请案第 10/900,021 号“Phrase Identification in an Information Retrieval System”;

[0005] 2004 年 7 月 26 日申请的申请案第 10/900,055 号“Phrase-Based Indexing in an Information Retrieval System”;

[0006] 2004 年 7 月 26 日申请的申请案第 10/900,039 号“Phrase-Based Personalization of Searches in an Information Retrieval System”;

[0007] 2004 年 7 月 26 日申请的申请案第 10/900,259 号“Automatic Taxonomy Generation in Search Results Using Phrases”;

[0008] 2004 年 7 月 26 日申请的申请案第 10/900,075 号“Phrase-Based Generation of Document Descriptions”;及

[0009] 2004 年 7 月 26 日申请的申请案第 10/900,012 号“Phrase-Based Detection of Duplicate Documents in an Information Retrieval System”;

[0010] 所有这些申请案被共同拥有并且以引用的方式并入本文中。

技术领域

[0011] 本发明涉及一种用于对诸如互联网 (Internet) 的大规模语料库中的文献编制索引、搜索与分类的信息检索系统。

背景技术

[0012] 信息检索系统通常称作搜索引擎,如今它们是一种用于在诸如互联网的大规模、多样化并不断增长的语料库中寻找信息的基本工具。一般来说,搜索引擎创建索引以使文献 (或“页”) 与各文献中存在的个别字相关。响应一含有多个查询项的查询来检索文献,此通常是基于在文献中存在一定数量的查询项而实现的。根据诸如查询项出现的频率、主域、链接分析等其它统计度量来对检索到的文献分等级。然后,通常按分等级后的次序将检索到的文献呈现给用户,而不进行任何其他分组或强制分级。在某些状况下,仅呈现文献文本的选定部分以便使用户能够粗略了解所述文献的内容。

[0013] 查询项的直接“布尔 (Boolean)”匹配具有多个熟知的限制,并且尤其无法识别那些不具有查询项但具有相关字的文献。举例来说,在典型的布尔系统中,搜索“Australian Shepherds(澳大利亚牧羊犬)”时将不会返回不具有确切查询项的关于其它 herding dogs(牧羊犬)(例如,Border Collies(博得牧羊犬)) 的文献。反而,所述系统通常可能同时检索到关于 Australia(澳大利亚)(并且与 dogs(狗) 无关) 的文献与关于

“shepherds(牧羊犬)”的文献,并且将这些文献排在较高等级。

[0014] 这里的问题是传统的系统是根据个别项而不是概念来编制文献索引。概念通常以短语表示,如“Australian Shepherd(澳大利亚牧羊犬)”、“President of the United States(美国总统)”或者“Sundance Film Festival(圣丹斯电影节)”等。某些现有系统最多是就预定且非常有限的“已知”短语集合来编制文献索引,这些“已知”短语一般是由人工操作员选择的。因为察觉到识别由(比如)三个、四个或五个或更多个字组成的所有可能的短语需要计算与存储器,所以一般会避免对短语编制索引。举例来说,如果假定任意五个字可构成一个短语并且一个大的语料库将具有至少 200,000 个唯一项,那么将存在约 $3.2*10^{26}$ 个可能短语,此明显超出任何现有系统能够存储于存储器中的量或者其可另外编程操纵的量。另一个问题是短语不断输入并会超出其在词典中的用法,此比发明新的个别字频繁得多。新短语总是从诸如技术、艺术、世界事件与法律等来源中产生。其它短语将随时间降低使用。

[0015] 某些现有信息检索系统试图通过使用个别字同时出现的模式来提供概念检索。在这些系统中,搜索一个字,例如“President(总统)”,将同时检索到具有频繁地与“President(总统)”一起出现的其它字(如“White(白色)”及“House(房子)”)的文献。尽管这种方法可能产生具有在个别字水平上概念性地相关的文献的搜索结果,但其一般无法俘获在同时出现的短语之间存在的主题关系。

[0016] 因此,需要一种信息检索系统与方法,其能够全面地识别大规模语料库中的短语、根据短语编制文献索引、根据其短语搜索文献并将文献分等级、并提供关于所述文献的另外的群集与说明性信息。

发明内容

[0017] 本发明涉及一种信息检索系统与方法,其使用短语来对文献库中的文献编制索引、进行搜索、分等级及说明。所述系统适合于识别那些在文献库中具有足够频繁及/或独特用法的短语以指示其为“有效”或“好”短语。以此方式,可识别多字短语,例如由四个、五个或更多项组成的短语。这就避免了必须识别由给定数量的字的所有可能序列所产生的每个可能的短语并对其编制索引的问题。

[0018] 该系统还适合于根据短语预测文献中存在其它短语的能力来识别彼此相关的短语。更具体地说,利用使两个短语的实际同时出现率与这两个短语的期望同时出现率相关的预测度量。一种此类预测度量是信息增益,即实际同时出现率与期望同时出现率的比率。在预测度量超过一预定阈值时,两个短语相关。在那种状况下,第二短语相对于第一短语具有显著的信息增益。语义上,相关短语将是那些共同用来讨论或描述一给定主题或概念的短语,如“President of the United States(美国总统)”与“White House(白宫)”。对于一给定短语,相关短语可根据其相关性或有效性基于其各自的预测度量来定序。

[0019] 信息检索系统通过有效或好短语来对文献库中的文献编制索引。对于每一个短语,一个记入列表识别那些含有所述短语的文献。此外,对于一给定短语,使用第二列表、向量或其它结构来存储指示在含有所述给定短语的每一文献中还存在给定短语的哪些相关短语的数据。以此方式,所述系统不仅能够响应搜索查询而轻易地识别出哪些文献含有哪些短语,而且能够识别出哪些文献还含有与查询短语相关、并且因此更可能特定地关于查

询短语所表示的主题或概念的短语。

[0020] 使用短语与相关短语还创建并使用了相关短语的群集，其在语义上代表短语的有意义的分组。从在群集中的所有短语之间具有非常高的预测度量的相关短语来识别群集。群集可用来组织搜索结果，包括选择搜索结果中包括哪些文献及其次序，以及从搜索结果去除文献。

[0021] 信息检索系统还适合于在响应查询而搜索文献时使用短语。处理查询以便识别在查询中存在的任何短语，从而检索查询短语的相伴记入列表与相关短语信息。此外，在有些情况下，用户可以在搜索查询中输入不完整的短语，如“President of the(……总统)”。可以识别象这样的不完整短语并且用扩展短语来代替，如“President of the United States(美国总统)”。这有助于确保实际执行用户最有可能的搜索。

[0022] 系统也可使用相关短语信息来识别或选择搜索结果中包括哪些文献。对于一给定短语与一给定文献，相关短语信息指出在所述给定文献中存在所述给定短语的哪些相关短语。因此，对于一含有两个查询短语的查询来说，先处理第一查询短语的记入列表以识别含有第一查询短语的文献，接着处理相关短语信息以识别这些文献中哪些文献还含有第二查询短语。接着，将后面这些文献包括在搜索结果中。这就不需要系统接着单独处理第二查询短语的记入列表，由此提供更快的搜索时间。当然，此方法也可以扩展到查询中有任意数量的短语，从而能够显著节约计算与时间。

[0023] 系统还可适合于使用短语与相关短语信息来对一组搜索结果中的文献分等级。一给定短语的相关短语信息较佳以诸如位向量的格式存储，其表示每一相关短语相对于所述给定短语的有效性。举例来说，一个相关短语位向量对于给定短语的每一个相关短语均具有一个位，这些位根据相关短语的预测度量（例如，信息增益）来定序。相关短语位向量的最有效的位与具有最高预测度量的相关短语相关，并且最低有效位与具有最低预测度量的相关短语相关。以此方式，对于一给定文献与一给定短语，相关短语信息可用来对文献计分。位向量本身（作为一个值）的值可用作文献分数，以此方式，含有查询短语的高级相关短语的文献比具有低级相关短语的文献更可能在主题上与查询相关。位向量值也可用作更复杂的计分函数中的一个分量，并且还可以加权。接着，可以根据文献分数来对文献分等级。

[0024] 短语信息也可以用在信息检索系统中以使用户的搜索个性化。将用户模拟为一个从（例如）所述用户曾经访问过（例如，在屏幕上看、打印、存储等等）的文献所获得的短语集合。更特定地说，给定用户访问过的文献，则在用户模型或概况中就会包括在此文献中存在的相关短语。在随后的搜索期间，使用用户模型中的短语来过滤搜索查询的短语并对检索到的文献的文献分数加权。

[0025] 短语信息也可以用在信息检索系统中以创建（例如）包括在一组搜索结果中的文献的文献说明。给定一搜索查询，所述系统识别出查询中存在的短语以及其相关短语与其扩展短语。对于一给定文献，所述文献的每一个句子都具有一个在句子中存在多少个查询短语、相关短语与扩展短语的计数。可以通过这些计数（个别或组合）来对文献句子分等级，并且选择一定数量的最高等级的句子（例如，五个句子）来形成文献说明。当搜索结果中包括所述文献时，可以接着向用户呈现文献说明，使得相对于查询用户能够更好地了解所述文献。

[0026] 进一步改进这种产生文献说明的方法，以使系统能够提供反映用户兴趣所在的个性化说明。如上所述，用户模型存储了识别用户感兴趣的相关短语的信息。此用户模型与一列与查询短语相关的短语相交，以识别这两组共有的短语。然后，根据相关短语信息来对所述共有集合定序。接着，使用所得相关短语集合来根据每一文献中存在的这些相关短语的实例数来对文献的句子分等级。选择具有最高数量的共有相关短语的多个句子作为个性化文献说明。

[0027] 当对文献库编制索引（爬行）或当处理搜索查询时，信息检索系统也可以使用短语信息来识别并去除重复文献。对于一给定文献，所述文献的每一个句子都具有一个在句子中存在多少个相关短语的计数。可以通过此计数来对文献句子分等级，并且选择多个最高等级的句子（例如，五个句子）来形成文献说明。然后，将与文献相关的此说明存储（例如）为所述句子的字符串或散列。在编制索引期间，以相同方式处理新爬行的文献以产生文献说明。新的文献说明可与先前的文献说明匹配（例如，散列），并且如果发现匹配，那么这个新的文献就是一个重复文献。类似地，在准备搜索查询的结果期间，可以处理搜索结果集合中的文献以去除重复文献。

[0028] 本发明的系统与软件架构、计算机程序产品及计算机实施的方法与计算机产生的用户界面与呈现具有其它实施例。

[0029] 上文仅仅是基于短语的信息检索系统与方法的一些特征。信息检索领域的技术人员将了解，短语信息普遍性的灵活性使其能够在文献分析与处理的编制索引、文献注释、搜索、分等级与其它领域中广泛使用与应用。

附图说明

[0030] 图 1 是本发明的一个实施例的软件架构的方块图。

[0031] 图 2 说明一种用于识别文献中的短语的方法。

[0032] 图 3 说明一具有短语窗口与二级窗口的文献。

[0033] 图 4 说明一种用于识别相关短语的方法。

[0034] 图 5 说明一种对相关短语的文献编制索引的方法。

[0035] 图 6 说明一种基于短语检索文献的方法。

[0036] 图 7 说明用于显示搜索结果的显示系统的操作。

[0037] 图 8a 及图 8b 说明引用文献与被引用文献之间的关系。

[0038] 这些图式仅仅是为了说明的目的而描绘本发明的一较佳实施例。从以下讨论，所属技术领域的技术人员将容易地了解，在不偏离本文所述的本发明的原理下，可采用本文所述的结构与方法的替代实施例。

具体实施方式

[0039] I. 系统概述

[0040] 现在参看图 1，其展示了根据本发明的一个实施例的搜索系统 100 的一实施例的软件架构。在此实施例中，系统包括一索引系统 100、一搜索系统 120、一显示系统 130 与一前端服务器 140。

[0041] 索引系统 110 负责识别文献中的短语并根据其短语通过访问不同网站 190 与其它

文献库来对文献编制索引。前端服务器 140 从用户端 170 的用户接收查询，并且向搜索系统 120 提供那些查询。搜索系统 120 负责搜索与搜索查询相关的文献（搜索结果），包括识别搜索查询中的任何短语，接着使用出现的短语对搜索结果中的文献分等级以影响等级次序。搜索系统 120 向显示系统 130 提供搜索结果。显示系统 130 负责修正搜索结果（包括除去接近重复的文献和产生文献的主题说明），并将修正后的搜索结果返回给前端服务器 140，即将结果提供给用户端 170。系统 100 进一步包括一用于存储关于文献的索引信息的索引 150 与一用于存储短语与相关统计信息的短语数据存储 160。

[0042] 就本申请案而言，“文献”应理解为可以由搜索引擎编制索引并检索的任何类型的媒体，包括网页文献、图像、多媒体文件、文本文献、PDF 或其它图像格式的文件等等。一个文献可以具有一或多个页、分区、段或其他适合其内容与类型的组成部分。同等地，文献可以称为“页”，其常用来指互联网上的文献。使用通用术语“文献”并不意味对本发明的范畴进行任何限制。搜索系统 100 可对大的文献语料库进行操作，如互联网与万维网，但其同样可用于更有限的集合中，如用于图书馆或私营企业的文献库。在任一情形下应了解，文献一般分布在许多不同的计算机系统与站点中。于是，不丧失一般性，不管格式或位置（例如，哪个网站或数据库），将文献统称为语料库或文献库。每个文献都具有一个唯一识别所述文献的相伴识别符；所述识别符较佳为 URL，但也可以使用其它类型的识别符（例如，文献号）。在本揭示中，假定使用 URL 来识别文献。

[0043] II. 索引系统

[0044] 在一实施例中，索引系统 110 提供三个主要功能性操作：1) 识别短语与相关短语，2) 关于短语对文献编制索引，及 3) 产生并维持基于短语的分类。所属技术领域的技术人员将了解，在传统索引功能的支持下，索引系统 110 还将执行其它功能，因此本文不再进一步说明这些其它操作。索引系统 110 对短语数据的索引 150 与数据储存库 160 进行操作。下文进一步说明这些数据储存库。

[0045] 1. 短语识别

[0046] 索引系统 110 的短语识别操作识别文献库中的“好”与“坏”短语，这些短语有助于对文献编制索引并搜索。一方面，好短语是那些往往出现在文献库中超过某一百分比的文献中的短语，且 / 或表示为在所述文献中具有不同的外观，如由置标标签或其它形态、格式或语法标记来定界。好短语的另一方面是其能够预测其它好短语，而不仅仅是出现在词典中的字序列。举例来说，短语“President of the United States（美国总统）”是一个预测诸如“George Bush（乔治·布什）”与“Bill Clinton（比尔·克林顿）”等其它短语的短语。然而，诸如“fell down the stairs”或“top of the morning”、“out of the blue”的其它短语不具预测性，这是因为象这些的成语与习语往往与许多其它不同且无关的短语一起出现。因此，短语识别阶段确定哪些短语是好短语而哪些是坏短语（即，缺乏预测能力）。

[0047] 现在参看图 2，短语识别过程具有以下功能性阶段：

[0048] 200：收集可能且好的短语，以及所述短语的频率与同时出现的统计值；

[0049] 202：基于频率统计值将可能短语分为好短语或坏短语；

[0050] 204：基于从同时出现的统计值获得的预测性度量来精简好短语列表。

[0051] 现在将进一步详细地说明这些阶段的每个阶段。

[0052] 第一阶段 200 是这样一个过程，通过该过程，索引系统 110 爬行 (crawl) 文献库中的一组文献，随时间形成所述文献库的多个重复分区。每遍处理一个分区。每遍爬行的文献数可能变化，较佳为每个分区约 1,000,000 个文献。较佳仅处理每个分区中先前未爬行的文献，直到处理完所有文献，或满足某一其它终止准则。实际上，由于新文献不断地添加到文献库中，所以爬行不断继续。索引系统 110 对爬行后的每个文献采取下列步骤。

[0053] 以 n 的短语窗口长度遍历所述文献的各字，其中 n 是期望的最大短语长度。窗口的长度一般为至少 2 项，较佳为 4 或 5 项（字）。短语较佳包括短语窗口中的所有字，包括那些否则会被表征为结束字的字，如“a”、“the”等等。短语窗口可以由行尾、段落返回、置标标签或其他内容或格式变化的标志来终止。

[0054] 图 3 说明遍历期间文献 300 的一部分，其展示短语窗口 302 从字“stock”开始并向右扩展 5 个字。窗口 302 中的第一个字是候选短语 i，并且序列 i+1、i+2、i+3、i+4 与 i+5 中的每个短语同样为候选短语。因此，在此实例中，候选短语为：“stock”、“stock dogs”、“stock dogs for”、“stock dogs for the”、“stock dogs for the Basque”与“stock dogs for the Basque shepherds”。

[0055] 在每个短语窗口 302 中，依次检查每个候选短语以确定其是否已经存在于好短语列表 208 或可能短语列表 206 中。如果候选短语未出现在好短语列表 208 或可能短语列表 206 中，那就确定所述候选短语为“坏”短语并将其跳过。

[0056] 如果候选短语出现在好短语列表 208 中，如款目 g_j，那就更新短语 g_j 的索引 150 款目以包括所述文献（例如，其 URL 或其它文献识别符），以指示此候选短语 g_j 出现在当前文献中。短语 g_j 的索引 150 中的款目（或项）称作短语 g_j 的记入列表。记入列表包括其中出现短语的一列文献 d（通过其文献识别符，例如文献号或者 URL）。

[0057] 此外，如下文进一步解释，更新同时出现矩阵 212。在最初的第一遍中，好的与坏的列表都将为空，因此往往会将大多数短语添加到可能短语列表 206 中。

[0058] 如果候选短语没有出现在好短语列表 208 中，那就将其添加到可能短语列表 206 中，除非其中已经存在所述候短语。可能短语列表 206 上的每个款目 p 都具有三个相伴计数：

[0059] P(p)：存在可能短语的文献数；

[0060] S(p)：可能短语的所有实例数；及

[0061] M(p)：可能短语的引起注意的实例数。在可能短语与文献中的相邻内容的不同之处在于语法或格式标记，例如黑体或下划线或为超链接或引号中的锚文本时，可能短语的实例“引起注意”。这些（与其它）区别外观由各种 HTML 置标语言标签与语法标记来指示。当一个短语被放在好短语列表 208 中时，所述短语的这些统计值仍被保留。

[0062] 除了各列表外，还保留好短语的同时出现矩阵 212(G)。矩阵 G 具有 mxm 维，其中 m 是好短语的数量。矩阵中的每个款目 G(j,k) 代表一对好短语 (g_j, g_k)。同时出现矩阵 212 在逻辑上（但在物理上不一定）保留每对好短语 (g_j, g_k) 关于二级窗口 304 的三个独立计数，所述窗口 304 的中心位于当前字 i，并且扩展 +/-h 个字。在一实施例中，例如如图 3 所述，二级窗口 304 有 30 个字。因此，同时出现矩阵 212 保留：

[0063] R(j,k)：原始的同时出现计数，即短语 g_j 与短语 g_k 一起出现在二级窗口 304 中的次数；

[0064] D(j,k)：分离的引起注意的计数，即短语 g_j 或短语 g_k 作为特异文本出现在二级窗口中的次数；及

[0065] C(j,k)：连接的引起注意的计数，即短语 g_j 与短语 g_k 同时作为特异文本出现在二级窗口中的次数。使用连接的引起注意的计数尤其有利于避免短语（例如，版权通知）频繁出现在侧边栏、页脚或页眉中并因此实际上无法预测其它文本的情形。

[0066] 参看图 3 的实例，假定“stock dogs”以及短语“Australian Shepherd”与“Australian Shepard Club of America”都位于好短语列表 208 上。后两个短语出现在二级窗口 304 内当前短语“stock dogs”周围。然而，短语“Australian Shepherd Club of America”作为网站的超链接（由下划线指示）的锚文本出现。因此，所述对{“stock dogs”，“Australian Shepherd”}的原始同时出现计数递增，并且{“stock dogs”，“Australian Shepherd Club of America”}的原始同时出现计数和分离的引起注意的计数都递增，这是因为后者是作为特异文本出现的。

[0067] 对分区中的每个文献重复以序列窗口 302 与二级窗口 304 遍历每个文献的过程。

[0068] 在遍历完分区中的文献后，编制索引操作的下一阶段就是从可能短语列表 206 更新 202 好短语列表 208。如果可能短语列表 206 上的一个可能短语 p 的出现频率与出现所述短语的文献数指示其足够用作语义上有意义的短语，那就将所述短语移到好短语列表 208 中。

[0069] 在一实施例中，其测试如下。从可能短语列表 206 取一个可能短语 p 并且将其放在好短语列表 208 中，前提条件是：

[0070] a)P(p) ＞ 10 并且 S(p) ＞ 20（含有 p 的文献数大于 10，并且短语 p 的出现次数大于 20）；或者

[0071] b)M(p) ＞ 5（短语 p 的引起注意的实例数大于 5）。

[0072] 这些阈值与分区中的文献数成比例；例如，如果一个分区中爬行 2,000,000 个文献，那阈值大约加倍。当然，所属技术领域的技术人员将了解，这些阈值的具体值或测试其的逻辑可随需要而变化。

[0073] 如果短语 p 没有资格进入好短语列表 208，则检查其成为坏短语的资格。短语 p 是一个坏短语的条件是：

[0074] a) 含有短语的文献数 P(p) ＜ 2；并且

[0075] b) 短语的引起注意的实例数 M(p) ＝ 0。

[0076] 这些条件指示所述短语既不频繁，也不能用来指示有效内容，同样地，这些阈值可与分区中的文献数成比例。

[0077] 应注意，如上所述，除了多字短语外，好短语列表 208 自然将包括个别字作为短语。这是因为短语窗口 302 中的每个第一字总是一个候选短语，并且适当的实例计数将累积。因此，索引系统 110 可以自动地对个别字（即，具有单个字的短语）与多字短语编制索引。好短语列表 208 也将比基于 m 个短语的所有可能组合的理论最大值短很多。在典型实施例中，好短语列表 208 将包括约 $6.5x10^5$ 个短语。由于系统只需要明了可能短语和好短语，所以不需要存储坏短语列表。

[0078] 通过最后一遍检查文献库，由于大语料库中短语使用的预期分布，所以可能短语的列表将相对较短。因此，如果在第 10 遍（例如，10,000,000 个文献），一个短语第一次出

现，那么其在那次中是极不可能成为一个好短语的。其可能是刚开始使用的新短语，因此在随后爬行中变得越来越常见。在那种状况下，其相应计数将增大，并且可能最终满足成为一个好短语的阈值。

[0079] 编制索引操作的第三阶段是使用从同时出现矩阵 212 获得的预测性度量来精简 204 好短语列表 208。不经过精减，好短语列表 208 很可能包括许多尽管合理地出现在字典中但本身无法充分预测其它短语的存在或本身是更长短语的子序列的短语。除去这些较弱的好短语后更可能有力地获得好短语。为了识别好短语，使用一预测性度量，其表示给定一短语的存在，在文献中出现另一短语的可能性增加。在一实施例中，此完成如下。

[0080] 如上所述，同时出现矩阵 212 是存储与好短语相关联的数据的 mxm 矩阵。矩阵中的每行 j 代表好短语 g_i，并且每列 k 代表好短语 g_k。对于每个好短语 g_j，计算期望值 $E(g_j)$。期望值 E 是库中预期含有 g_j 的文献的百分比。例如，其计算为含有 g_j 的文献数与库中已爬行的文献总数 T 的比率 :P(j)/T。

[0081] 如上所述，当 g_j 每次出现在文献中时，即更新含有 g_j 的文献数。每次 g_j 的计数增加时或在此第三阶段期间，可更新 $E(g_j)$ 的值。

[0082] 接着，对于每个其它好短语 g_k（例如，矩阵的各列），确定 g_j 是否预测了 g_k。g_j 的预测性度量的确定如下：

[0083] i) 计算期望值 $E(g_k)$。如果 g_j 与 g_k 是无关短语，则其期望同时出现率 E(j，k) 为 $E(g_j)*E(g_k)$；

[0084] ii) 计算 g_j 与 g_k 的实际同时出现率 A(j，k)。即将原始同时出现计数 R(j，k) 除以文献总数 T；

[0085] iii) 据说当实际同时出现率 A(j,k) 超过期望同时出现率 E(j,k) 一临界量时，g_j 预测 g_k。

[0086] 在一实施例中，预测性度量为信息增益。因此，当在短语 g_j 面前另一短语 g_k 的信息增益 I 超过一阈值时，短语 g_j 预测短语 g_k。在一实施例中，此计算如下：

[0087] I(j，k) ＝ A(j，k)/E(j，k)。

[0088] 并且当满足下列条件时，好短语 g_j 预测好短语 g_k：

[0089] I(j，k) ＞信息增益阈值。

[0090] 在一实施例中，信息增益阈值为 1.5，但较佳在 1.1 与 1.7 之间。将阈值升高到超过 1.0 是为了减少两个原本无关的短语同时出现超过随机预测的可能性。

[0091] 如上所述，相对于给定行 j，对矩阵 G 的每列 k 重复信息增益的计算。在一行完成后，如果好短语 g_k 中无一短语的信息增益超过信息增益阈值，那这就意味着短语 g_j 无法预测任何其它好短语。在那种状况下，从好短语列表 208 除去 g_j，其基本上就变为坏短语。注意，不除去短语 g_j 的列 j，因为这个短语本身可由其它好短语来预测。

[0092] 当评估完同时出现矩阵 212 中的所有行后，结束这个步骤。

[0093] 该阶段的最后一个步骤是精简好短语列表 208 以除去不完整短语。一个不完整短语是一个仅预测其扩展短语并且从所述短语的最左侧（即，短语的开始处）开始的短语。短语 p 的“扩展短语”是一个以短语 p 开始的超序列。举例来说，短语“President of”预测“President of the United States”、“President of Mexico”、“President of AT&T”等等。由于所有后面这些短语都是以“President of”开始并且是其超序列，所以他们都是

“President of”的扩展短语。

[0094] 因此，保留在好短语列表 208 上的每个短语 g_j 都将基于前述信息增益阈值来预测一定量的其它短语。现在，对于每个短语 g_j，索引系统 110 执行其与其所预测的每个短语 g_k 的字符串匹配。字符串匹配测试每个预测短语 g_k 是否是短语 g_j 的扩展短语。如果所有预测短语 g_k 都是短语 g_j 的扩展短语，那么 g_j 就不完整，将其从好短语列表 208 中除去并添加到不完整短语列表 216 中。因此，如果存在至少一个不是 g_j 的扩展短语的短语 g_k，那 g_j 就是完整的，并且会保留在好短语列表 208 中。于是举例来说，当“President of the United”所预测的唯一其它短语是“President of the United States”并且这个预测短语是所述短语的扩展短语时，“President of the United”就是一个不完整短语。

[0095] 不完整短语列表 216 本身在实际搜索过程中非常有用。当接收到搜索查询时，可将其与不完整列表 216 比较。如果所述查询（或其一部分）与所述列表中的一个款目匹配，那搜索系统 120 就可以查找这个不完整短语的最可能的扩展短语（给定不完整短语，具有最高信息增益的扩展短语），并且向用户建议此短语或对扩展短语自动搜索。例如，如果搜索查询是“President of the United”，那搜索系统 120 可以自动向用户建议“President of the United States”作为搜索查询。

[0096] 在完成编制索引过程的最后一个阶段后，好短语列表 208 将含有在语料库中发现的大量好短语。这些好短语中的每一个短语都将预测至少一个不是其扩展短语的其它短语。即，每一个好短语都以足够的频率使用，并且独立代表语料库中所表示的有意义的概念或思想。与使用预定或人工选择的短语的现有系统不同，好短语列表反映了语料库中正在实际使用的短语。此外，由于新文献添加到文献库中使得周期性地重复上述爬行与编制索引过程，所以索引系统 110 在新短语进入词典时自动检测所述新短语。

[0097] 2. 识别相关短语与相关短语的群集

[0098] 参看图 4，相关短语识别过程包括以下功能性操作：

[0099] 400：识别具有高信息增益值的相关短语；

[0100] 402：识别相关短语的群集；

[0101] 404：存储群集位向量与群集号。

[0102] 现在详细描述这些操作中的每一个操作。

[0103] 首先回想，同时出现矩阵 212 含有好短语 g_j，其中每一个短语都预测至少一个具有大于信息增益阈值的信息增益的其它好短语 g_k。然后，为了识别 400 相关短语，对于每一对好短语 (g_j, g_k)，将信息增益与相关短语阈值（例如，100）进行比较。即，当

[0104] $I(g_j, g_k) > 100$ 时，

[0105] g_j 与 g_k 是相关短语。

[0106] 使用此高阈值来识别很好地超过统计期望率的好短语的同时出现。在统计上，其意指短语 g_j 与 g_k 同时出现率超过期望同时出现率的 100 倍。举例来说，给定文献中的短语“Monica Lewinsky”，如果短语“Bill Clinton”在相同文献中更可能出现率是其 100 倍，则短语“Bill Clinton”可能出现在任意随机选择的文献中。因为出现率是 100 ： 1，所以另一种表述方式是预测精确度为 99.999％。

[0107] 因此，将小于相关短语阈值的任何款目 (g_j, g_k) 调零，以指示短语 g_j，g_k 不相关。现在，同时出现矩阵 212 中任何剩余款目都指示所有相关短语。

[0108] 接着，通过信息增益值 $I(g_j, g_k)$ 来对同时出现矩阵 212 的各行 g_j 中的列 g_k 排序，使得首先列出具有最高信息增益的相关短语 g_k。因此，此排序为一给定短语 g_j 识别出按照信息增益哪些其它短语最可能相关。

[0109] 下一步骤是确定 402 哪些相关短语一起形成相关短语群集。群集是相关短语的集合，其中每个短语相对于至少一个其它短语而具有高信息增益。在一实施例中，群集的识别如下。

[0110] 在矩阵的每行 g_j 中，将存在一或多个与短语 g_j 相关的其它短语。这个集合就是相关短语集合 R_j，其中 $R = \{g_k\ g_1 \ldots g_m\}$。

[0111] 对于 R_j 中的每个相关短语 m，索引系统 110 确定 R 中的各其它相关短语是否也与 gj 相关。因此，如果 $I(g_k, g_l)$ 也非零，那 g_j、g_k 与 g_l 是群集的一部分。对 R 中的每一对 (g_l, g_m) 重复此群集测试。

[0112] 举例来说，假定好短语“Bill Clinton”与短语“President”、“Monica Lewinsky”相关，这是因为每一个这些短语相对于“Bill Clinton”的信息增益都超过相关短语阈值。另外，假定短语“Monica Lewinsky”与短语“purse designer”相关。这些短语于是形成集合 R。为确定群集，索引系统 110 通过确定这些短语的相应信息增益来评估每个短语相对于其它短语的信息增益。因此，索引系统 110 确定 R 中的所有对短语的信息增益 I(“President”，“Monica Lewinsky”)、I(“President”，“purse designer”) 等等。在此实例中，“Bill Clinton”、“President”与“Monica Lewinsky”形成一群集，“Bill Clinton”与“President”形成第二群集，并且“Monica Lewinsky”与“purse designer”形成第三群集，并且“Monica Lewinsky”、“Bill Clinton”与“purse designer”形成第四群集。这是因为尽管“Bill Clinton”没有足够的信息增益来预测“purse designer”，但“Monica Lewinsky”仍预测这两个短语。

[0113] 为记录 404 群集信息，向每一个群集指派一个唯一的群集号 (群集 ID)。然后，结合每一个好短语 g_j 一起记录此信息。

[0114] 在一实施例中，群集号是由群集位向量来确定，群集位向量还指示短语之间的正交关系。群集位向量是长度为 n 的位的序列，其中 n 是好短语列表 208 中的好短语的数量。对于一给定好短语 g_j，位位置对应于 g_j 的排序后的相关短语 R。如果 R 中的相关短语 g_k 与短语 g_j 在同一个群集中，则设定一个位。更一般来说，这意味着如果在 g_j 与 g_k 之间的任一方向上存在信息增益，则设定群集位向量中的相应位。

[0115] 于是，群集号就是所得位串的值。此实施例具有这样一个特性，即具有多向或单向信息增益的相关短语出现在相同群集中。

[0116] 如下是使用上述短语的群集位向量的一个实例 ：

[0117]

	Bill Clinton	President	Monica Lewinsky	purse designer	群集 ID
Bill Cliton	1	1	1	0	14
President	1	1	0	0	12
Monica Lewinsky	1	0	1	1	11
purse designer	0	0	1	1	3

[0118] 于是概述之，在此过程后，将为每一个好短语 g_j 识别一组相关短语 R，其按照信息增益 $I(g_j, g_k)$ 从高到低的次序排列。此外，对于每一个好短语 g_j，都将有一个群集位向量，其值是一个用于识别短语 g_j 所属的主要群集的群集号，且其正交值（对于每个位位置为 1 或 0）指示 R 中的相关短语中哪个短语与 g_j 处于共同群集中。因此，在上述实例中，“Bill Clinton”、“President”与“Monica Lewinsky”处于基于短语“Bill Clinton”的行中的位的值的群集 14 中。

[0119] 为存储此信息，可使用两种基本表示法。第一，如上所述，可将信息存储在同时出现矩阵 212 中，其中：

[0120] 款目 G[行 j，列 k] = (I(j，k)，群集号，群集位向量)。

[0121] 或者，可避免矩阵表示法，而将所有信息存储在好短语列表 208 中，其中每行代表一个好短语 g_j；

[0122] 短语行 j =列表 [短语 g_k，(I(j，k)，群集号，群集位向量)]。

[0123] 此方法提供了一种有用的群集组织法。首先，此方法不是一个严格并且通常任意界定的主题与概念的分级，而是认可相关短语所示的主题形成一个复杂的关系表，其中某些短语与许多其它短语相关，并且某些短语的范围更有限，并且其中各关系可能是相互的（每个短语预测其它短语）或单向的（一个短语预测其它短语，但反之则不可）。结果是可将群集表征成对每个好短语来说是“局部”的，于是某些群集将由于具有一或多个共同的相关短语而重叠。

[0124] 于是对于一个给定的好短语 g_j，相关短语按照信息增益的定序提供了一种用来命名短语群集的分类法：群集名是群集中具有最高信息增益的相关短语的名称。

[0125] 上述方法提供了一种用于识别出现在文献库中的有效短语的非常有力的方式以及这些相关短语在实际实施中一起用在自然“群集”中的方式。因此，对相关短语的此数据驱动群集避免了许多系统中常见的相关术语与概念的任何人工导向的“编辑”选择所固有的偏差。

[0126] 3. 以短语与相关短语对文献编制索引

[0127] 给定包括关于相关短语与群集的信息的好短语列表 208，索引系统 110 的下一个功能性操作是关于好短语与群集来对文献库中的文献编制索引，并将更新后的信息存储在索引 150 中。图 5 说明此过程，其中包括编制文献索引的下列功能性阶段：

[0128] 500：将文献记入在文献中所发现的好短语的记入列表中；

[0129] 502：更新相关短语与二级相关短语的实例计数与相关短语位向量；

[0130] 504：以相关短语信息来注释文献；

[0131] 506：根据记入列表大小来对索引款目重新定序。

[0132] 现在将更详细地描述这些阶段。

[0133] 如上所述，遍历或爬行一文献集合；此可以是相同或不同的文献集合。对于一给定文献 d，以上述方式从位置 i 开始，以长度为 n 的序列窗口 302 逐字遍历 500 文献。

[0134] 在一给定短语窗口 302 中，从位置 i 开始识别窗口中的所有好短语。每个好短语都表示为 g_i。因此，g1 是第一个好短语，g2 是第二个好短语，依此类推。

[0135] 对于每个好短语 g_i（实例 g1“President”与 g4“President of ATT”），将文献识别符（例如，URL）记入到索引 150 中的好短语 g_i 的记入列表中。此更新识别出，在此特定文献中出现好短语 g_i。

[0136] 在一实施例中，短语 g_j 的记入列表采用以下逻辑形式：

[0137] 短语 g_j：列表：（文献 d，[列表：相关短语计数][相关短语信息]）。

[0138] 对于每个短语 g_j，都有一个出现所述短语的文献 d 列表。对于每个文献，都有一个同样出现在文献 d 中的短语 g_j 的相关短语 R 的出现次数的计数列表。

[0139] 在一实施例中，相关短语信息是一个相关短语位向量。此位向量可表征为一个“双位”向量，这是因为对于每个相关短语 g_k，都有两个位位置：g_k-1 与 g_k-2。第一位位置存储一指示在文献 d 中是否存在相关短语 g_k 的标号（即，文献 d 中的 g_k 的计数大于 0）。第二位位置存储一指示在文献 d 中是否也存在短语 g_k 的相关短语 g_l 的标号。短语 g_j 的相关短语 g_k 的相关短语 g_l 在本文中称作“g_j 的二级相关短语”。所述计数与位位置对应于 R 中短语的规范次序（按照递减的信息增益的次序排列）。此排序次序产生这样一个效果，即使得 g_j 最高度预测的相关短语 g_k 与相关短语位向量的最有效位相关，而 g_j 最少预测的相关短语 g_l 与最低有效位相关。

[0140] 比较有用的是注意到对于一给定短语 g，就所有含有 g 的文献而言，相关短语位向量的长度以及相关短语与所述向量的个别位之间的缔合都相同。此实施例具有以下特性，即使系统容易地比较含有 g 的任何（或所有）文献的相关短语位向量，以观察哪些文献具有给定的相关短语。这有利于促进搜索过程响应搜索查询来识别文献。因此，给定文献将出现在许多不同短语的记入列表中，并且在每个此类记入列表中，所述文献的相关短语向量将专用于拥有所述记入列表的短语。这方面保持了相关短语位向量相对于个别短语与文献的局部性。

[0141] 因此，下一阶段 502 包括遍历文献中的当前索引位置的二级窗口 304（如前所述，+/-K 项（例如 30 项）的二级窗口），例如从 i-K 到 i+K。对于出现在二级窗口 304 中的 g_i 的每个相关短语 g_k，索引系统 110 相对于相关短语计数中的文献 d 来使 g_k 的计数递增。如果 g_i 稍后出现在文献中并且在稍后的二级窗口中再次发现相关短语，则再次递增计数。

[0142] 如上所述，基于计数来设定相关短语位映射中的相应第一位 g_k-1，如果 g_k 的计数 >0，则将位设置为 1，或如果所述计数等于 0，则将其设置为 0。

[0143] 接着，在索引 150 中查找相关短语 g_k，在 g_k 的记入列表中识别文献 d 的款目，然后检查 g_k 的任何相关短语的二级相关短语计数（或位），从而设定第二位 g_k-2。如果设定了任何这些二级相关短语计数 / 位，则此指示在文献 d 中还存在 g_j 的二级相关短语。

[0144] 当以此方式完全处理完文献 d 时，索引系统 110 将已经识别出：

[0145] i）文献 d 中的每个好短语 g_j；

[0146] ii）为每个好短语 g_j 识别出在文献 d 中存在其哪些相关短语 g_k；

[0147] iii) 为存在于文献 d 中的每个相关短语 g_k 识别出在文献 d 中还存在其哪些相关短语 g_1(g_j 的二级相关短语)。

[0148] a) 确定文献主题

[0149] 通过短语对文献编制索引并使用群集信息提供了索引系统 110 的另一个优点,即能够基于相关短语信息来确定文献的主题。

[0150] 假定对于一给定短语 g_j 与一给定文献 d,记入列表款目如下 :

[0151] g_j :文献 d :相关短语计数 := {3,4,3,0,0,2,1,1,0}

[0152] 相关短语位向量 := {11 11 10 00 00 10 10 10 01}

[0153] 其中,相关短语位向量展示为双位对。

[0154] 从相关短语位向量,我们可以确定文献 d 的一级与二级主题。一级主题由位对 (1,1) 指示,而二级主题由位对 (1,0) 指示。相关短语位对 (1,1) 指示文献 d 中同时存在所述位对的相关短语 g_k 以及二级相关短语 g_1。此可以解释为意味在撰写所述文献 d 时文献的作者一起使用了若干相关短语 g_j、g_k 与 g_1。位对 (1,0) 指示同时存在 g_j 与 g_k,但不存在 g_k 的任何其他二级相关短语,因此这是一个不那么有效的主题。

[0155] b) 改善分等级的文献注释

[0156] 索引系统 110 的另一方面是能够在编制索引过程中用使得随后搜索期间的分等级改善的信息注释 504 每个文献 d。注释过程 506 如下。

[0157] 文献库中的给定文献 d 可以具有一定数量的对其它文献的外链接。每个外链接(超链接)都包括锚文本与目标文献的文献识别符。为了解释,将正在处理的当前文献 d 称作 URL0,并且将文献 d 上的外链接的目标文献称作 URL1。为了稍候用于对搜索结果中的文献分等级,对于指向某些其它 URLi 的 URL0 中的每个链接,索引系统 110 创建所述链接相对于 URL0 的锚短语的外链接分数与所述锚短语相对于 URLi 的内链接分数。即,文献库中的每个链接都有一对分数,即一个外链接分数与一个内链接分数。这些分数的计算如下。

[0158] 在给定文献 URL0 上,索引系统 110 识别对另一文献 URL1 的每个外链接,其中锚文本 A 是在好短语列表 208 中的一个短语。图 8a 示意性地说明此关系,其中文献 URL0 中的锚文本“A”用于超链接 800 中。

[0159] 在短语 A 的记入列表中,将 URL0 作为短语 A 的外链接记入,并且将 URL1 作为短语 A 的内链接记入。对于 URL0,如上所述来完成相关短语位向量,以识别 URL0 中存在的 A 的相关短语与二级相关短语。将此相关短语位向量用作从 URL0 到含有锚短语 A 的 URL1 的链接的外链接分数。

[0160] 接着,如下确定内链接分数。对于对含有锚短语 A 的 URL1 的每个内链接,索引系统 110 扫描 URL1,并且确定在 URL1 的主体中是否出现短语 A。如果短语 A 不仅指向 URL1(通过 URL0 上的外链接),而且出现在 URL1 本身的内容中,那此就表明 URL1 可称作与短语 A 所代表的概念内部相关。图 8b 说明了此状况,其中短语 A 出现在 URL0(作为锚文本)与 URL1 的主体中。在此状况下,将 URL1 的短语 A 的相关短语位向量用作从 URL0 到含有短语 A 的 URL1 的链接的内链接分数。

[0161] 如果锚短语 A 没有出现在 URL1 的主体中 (如图 8a),则就采取不同的步骤来确定内链接分数。在此状况下,索引系统 110 创建用于短语 A 的 URL1 的相关短语位向量 (就好像在 URL1 中存在短语 A),其指示短语 A 的哪些相关短语出现在 URL1 中。接着,将此相关短

语位向量用作从 URL0 到 URL1 的链接的内链接分数。

[0162] 举例来说，假定在 URL0 与 URL1 中最初存在以下短语：

[0163]

	锚短语	相关短语位向量				
文献	Australian Shepherd	Aussie	blue merle	red merle	tricolor	agility training
URL0	1	1	0	0	0	0
URL1	1	0	1	1	1	0

[0164] （在上述与以下表中，未展示二级相关短语）。URL0 行是来自锚文本 A 的链接的外链接分数，并且 URL1 行是所述链接的内链接分数。这里，URL0 含有目标为 URL1 的锚短语“Australian Shepard”。在“Australian Shepard”的五个相关短语中，仅一个“Aussie”出现在 URL0 中。于是直观地，URL0 与 Australian Shepards 仅弱相关。通过比较，URL1 不仅具有存在于文献主体中的短语“Australian Shepherd”，而且还具有多个相关短语“blue merle”、“red merle”与“tricolor”。因此，由于锚短语“Australian Shepard”出现在 URL0 与 URL1 中，所以 URL0 的外链接分数与 URL1 的内链接分数是上述相应行。

[0165] 上述第二种状况是指 URL1 中没有出现锚短语 A 的情形。在那种状况下，索引系统 110 扫描 URL1 并确定在 URL1 中存在相关短语“Aussie”、“blue merle”、“red merle”、“tricolor”与“agility training”中的哪些短语，并因此产生一个相关短语位向量，例如：

[0166]

	锚短语	相关短语位向量				
文献	Australian Shepherd	Aussie	blue merle	red merle	tricolor	agility training
URL0	1	1	0	0	0	0
URL1	0	0	1	1	1	0

[0167] 这里，此表明 URL1 不含有锚短语“Australian Shepard”，但含有相关短语“blue merle”、“red merle”与“tricolor”。

[0168] 此方法有利于完全防止对网页（一类文献）进行某些类型的歪曲搜索结果的操纵。通过人工创建大量具有指向所要页的给定锚文本的页可以“轰击”使用有赖于指向给定文献的链接数的分等级算法来对所述文献分等级的搜索引擎。因此，当输入使用锚文本的搜索查询时，通常会返回所要页，即使实际上此页与锚文本几乎或完全没有关系。将相关位向量从目标文献 URL1 输入到文献 URL0 的短语 A 的相关短语位向量中消除了搜索系统对指向 URL1 以作为有效性的指示的 URL0 中或 URL1 中的短语 A 与锚文本短语之间的关系的依赖性。

[0169] 基于索引 150 中的每个短语在语料库中的出现频率，亦为各短语赋予一个短语号。短语越常见，其在索引中接收的短语号就越低。接着，索引系统 110 根据每个记入列表中的短语号所列出的文献数来对索引 150 中的所有记入列表降序排序 506，使得首先列出最频繁出现的短语。于是，可以使用短语号来查找特定短语。

[0170] III. 搜索系统

[0171] 搜索系统 120 用于接收查询并搜索与所述查询相关的文献，并且在搜索结果集合

中提供这些文献的列表（以及这些文献的链接）。图 6 说明搜索系统 120 的主要功能性操作：

[0172] 600：识别查询中的短语；

[0173] 602：检索与查询短语相关的文献；

[0174] 604：根据短语对搜索结果中的文献分等级。

[0175] 这些阶段中的每一阶段的细节如下。

[0176] 1. 识别查询及展开查询中的短语

[0177] 搜索系统 120 的第一阶段 600 是识别查询中存在的任何短语以便有效地搜索其索引。在这部分中使用下列术语：

[0178] q：所输入的并由搜索系统 120 接收的查询；

[0179] Qp：所述查询中存在的短语；

[0180] Qr：Qp 的相关短语；

[0181] Qe：Qp 的扩展短语；

[0182] Q：Qp 和 Qr 的并集。

[0183] 从用户端 190 接收查询 q，所述查询 q 具有至多某一最大数量的字符或字。

[0184] 搜索系统 120 使用大小为 N（例如 5）的短语窗口来遍历所述查询 q 的各项。所述短语窗口先从所述查询的第一项开始，然后向右扩展 N 项。然后，这个窗口向右移动 M-N 次，其中 M 是所述查询的项数。

[0185] 在每个窗口位置，窗口中都将存在 N 项（或更少项）。这些项构成一个可能的查询短语。在好短语列表 208 中查找可能短语，判断它是不是一个好短语。如果好短语列表 208 中有这个可能短语，那么给短语返回一个短语号；现在，这个可能短语就是一个候选短语。

[0186] 在测试完每个窗口中的所有可能短语以判断它们是否是好的候选短语后，搜索系统 120 将为查询中的对应短语赋予一组短语号。接着，将这些短语号排序（降序）。

[0187] 从作为第一候选短语的最高短语号开始，搜索系统 120 判断在排序后的列表中的固定数字距离内是否有另一个候选短语，即短语号之间的差值在（例如）20,000 的临界量内。如果有，那么选择查询中最左边的短语作为有效的查询短语 Qp。从候选短语列表中除去这个查询短语及其所有子短语，并且将所述列表重新排序并重复上述过程。这个过程的结果是一组有效查询短语 Qp。

[0188] 例如，假定搜索查询是“Hillary Rodham Clinton Bill on the Senate Floor（参议院议员希拉里·罗德翰·克林顿·比尔）”。搜索系统 120 会识别下列候选短语：“Hillary Rodham Clinton Bill on”、“Hillary Rodham Clinton Bill”及“Hillary Rodham Clinton”。删除前两个，而保持最后一个作为有效查询短语。接着，搜索系统 120 会识别“参议院议员比尔 (Bill on the Senate Floor)”和子短语“Bill on the Senate”、“Bill on the”、“Bill on”、“Bill”，并且会选择“Bill”作为有效查询短语 Qp。最后，搜索系统 120 会分解“on the Senate Floor”，并识别“Senate Floor”作为有效查询短语。

[0189] 然后，搜索系统 120 调整有效短语 Qp 的首字母大写。在分解查询时，搜索系统 120 识别每个有效短语中的潜在首字母大写。此可以通过利用已知的首字母大写表（例如，“united states”的首字母大写为“United States”）或利用以语法为基础的首字母大写算法来完成。此产生适当首字母大写的查询短语集合。

[0190] 接着，当该集合中同时存在短语及其子短语时，搜索系统 120 会第二遍检查首字母大写的短语，并且只选择那些短语的最左边字母将其变成大写。例如，对“president of the united states”的搜索的大写将为“President of the United States”。

[0191] 在下一阶段，搜索系统 120 识别 602 那些与查询短语 Q 相关的文献。搜索系统 120 接着检索查询短语 Q 的记入列表，并且使这些列表相交以判断哪些文献出现在查询短语的所有（或一些）记入列表上。如果查询中的短语 Q 中有一组扩展短语 Qe（下文中将进一步解释），那么搜索系统 120 首先形成所述扩展短语的记入列表的并集，然后使其与这些记入列表相交。如上所述，搜索系统 120 通过在不完整短语列表 216 中查找每个查询短语 Q 来识别扩展短语。

[0192] 相交的结果是一组与查询相关的文献。通过短语和相关短语编制文献索引，识别查询中的短语 Q，然后将查询展开到包括扩展短语，从而产生一组比传统的基于布尔的搜索系统更与查询相关的文献的选集，在传统的基于布尔的搜索系统中，只选择那些含有所述查询项的文献。

[0193] 在一实施例中，搜索系统 120 可以使用一优化机制来响应查询识别文献而不一定使查询短语 Q 的所有记入列表相交。由于索引 150 的结构，所以对于每一个短语 g_j，其相关短语 g_k 均知晓，并且在 g_k 的相关短语位向量中识别。因此，此信息可用于简化其中两个或两个以上查询短语是彼此相关的短语或具有共同的相关短语的相交过程。在那些情况下，相关短语位向量可直接存取然后用于接着检索相应文献。下文更全面地描述此方法。

[0194] 给定任意两个查询短语 Q1 和 Q2，会有三种可能的相关情形：

[0195] 1)Q2 是 Q1 的相关短语；

[0196] 2)Q2 不是 Q1 的相关短语，且其各自的相关短语 Qr1 和 Qr2 不相交（即，没有共同的相关短语）；及

[0197] 3)Q2 不是 Q1 的相关短语，但其各自的相关短语 Qr1 和 Qr2 相交。

[0198] 对于每一对查询短语，搜索系统 120 通过查找查询短语 Qp 的相关短语位向量来确定适当的情形。

[0199] 搜索系统 120 继续为查询短语 Q1 检索包括那些含有 Q1 的文献的记入列表，并为这些文献中的每个文献检索相关短语位向量。Q1 的相关短语位向量将指示短语 Q2（若有，则还包括剩余查询短语中的每个短语）是否是 Q1 的相关短语且是否存在于该文献中。

[0200] 如果第一种情况适用于 Q2，那么搜索系统 120 扫描 Q1 记入列表中的每个文献的相关短语位向量，以便判断其中是否设有 Q2 的位。如果 Q1 记入列表中的文献 d 没有设这个位，那么就意味 Q2 没有出现在那个文献中。因此，可以立即将这个文献排除在考虑之外。然后，可以对剩余文献计分。这还意味着搜索系统 120 无需处理 Q2 的记入列表来查看它还存在于哪些文献中，从而节省了计算时间。

[0201] 如果第二种情况适用于 Q2，那么这两个短语彼此无关。例如，查询“cheap bolt action rifle（便宜的手动枪栓步枪）”有两个短语“cheap”和“bolt action rifle”。这些短语无一相关，另外，这些短语中的每个短语的相关短语都不重叠；即“cheap”的相关短语有“low cost”、“inexpensive”、“discount”、“bargain basement”和“lousy”，而“bolt action rifle”的相关短语有“gun”、“22caliber”、“magazine fed”和“Armalite AR30M”，因此这些列表不相交。在此情况下，搜索系统 120 使 Q1 和 Q2 的记入列表正则 (regular)

相交以便获得文献用于计分。

[0202] 如果第三种情况适用，那么两个短语 Q1 和 Q2 虽然不相关，但它们具有至少一个共同的相关短语。例如，短语“bolt action rifle”和“22”都会有“gun”作为相关短语。在此情况下，搜索系统 120 检索这两个短语 Q1 和 Q2 的记入列表并且使这些列表相交以产生含有这两个短语的文献列表。

[0203] 然后，搜索系统 120 可以快速地对所得文献中的每个文献计分。首先，搜索系统 120 确定每个文献的分数调整值。分数调整值是由在一文献的相关短语位向量中对应于查询短语 Q1 和 Q2 的位置的位所形成的掩码。例如，假定 Q1 和 Q2 对应于文献 d 的相关短语位向量中的第三和第六个双位位置，并且第三个位置的位值是 (1,1) 且第六个位置的位值是 (1,0)，那么分数调整值就是位掩码“000011000010”。然后，使用分数调整值来屏蔽文献的相关短语位向量，接着将修正后的短语位向量载入分等级函数（如下所述）以便用于计算所述文献的体分数。

[0204] 2. 分等级

[0205] a) 基于所含短语对文献分等级

[0206] 搜索系统 120 提供分等级阶段 604，在此阶段，使用每个文献的相关短语位向量中的短语信息和查询短语的群集位向量来对搜索结果中的文献分等级。此方法是根据文献中所含有的短语或非正式的“体命中数”来分等级。

[0207] 如上所述，对于任一给定的短语 g_j，g_j 的记入列表中的每个文献 d 都有一个用于识别在文献 d 中存在哪些相关短语 g_k 和哪些二级相关短语 g_i 的相伴相关短语位向量。一给定文献中存在的相关短语和二级相关短语越多，在给定短语的文献相关短语位向量中就将设置越多的位。设置的位越多，相关短语位向量的数值就越大。

[0208] 因此，在一个实施例中，搜索系统 120 根据文献的相关短语位向量的值来对搜索结果中的文献排序。含有与查询短语 Q 最相关的短语的文献将具有最高值的相关短语位向量，并且这些文献将是搜索结果中的最高等级的文献。

[0209] 此方法之所以较理想是因为在语义上，这些文献在主题上与查询短语最相关。注意，此方法可以提供高度相关的文献，尽管这些文献不含高频率的输入查询项 q，这是因为相关短语信息不仅用于识别相关文献，而且接着对这些文献分等级。具有低频率的输入查询项的文献仍可具有查询项的大量相关短语，因此其可比具有高频率的输入查询项和短语但无相关短语的文献更相关。

[0210] 在第二实施例中，搜索系统 120 根据结果集合中每个文献所含有的查询短语 Q 的相关短语来对每个文献计分。此通过如下方式完成。

[0211] 给定每个查询短语 Q，将存在某一数量 N 的与所述查询短语相关的短语 Qr，其可在短语识别过程中识别。如上所述，根据相关查询短语 Qr 来自查询短语 Q 的信息增益来对相关查询短语 Qr 定序。然后，对这些相关短语指派点数，先为第一相关短语 Qr1（即，具有来自 Q 的最高信息增益的相关短语 Qr）指派 N 个点，然后为下一个相关短语 Qr2 指派 N-1 个点，然后为 Q3 指派 N-2 个点，依此类推，因此将最后一个相关短语 QrN 指派为 1 个点。

[0212] 然后，确定存在查询短语 Q 的哪些相关短语 Qr，并且为所述文献赋予指派给每个此等相关短语 Qr 的点数，从而对搜索结果中的每个文献计分。接着将所述文献按照从高到低的分数排序。

[0213] 作为另一改进，搜索系统 120 可以从结果集合中精选文献。在某些情况下，文献可能关于许多不同的主题；尤其对于较长文献而言更是如此。在许多情况下，相比于与许多不同主题相关的文献，用户更喜欢那些与查询中所表示的单个主题密切相关的文献。

[0214] 为了精选后一种文献，搜索系统 120 使用查询短语的群集位向量中的群集信息，并且除去其中具有多于临界数量的群集的任何文献。例如，搜索系统 120 可除去任何含有多于两个群集的文献。此群集阈值可预先确定，或由用户设定为一个搜索参数。

[0215] b) 基于锚短语对文献分等级

[0216] 除了基于查询短语 Q 的体命中数来对搜索结果中的文献分等级外，在一个实施例中，搜索系统 120 还基于以对其他文献的锚出现的查询短语 Q 和相关查询短语 Qr 来对文献分等级。在一个实施例中，搜索系统 120 计算每个文献的分数，所述分数是两个分数（即，体命中分数和锚命中分数）的函数（例如，线性组合）。

[0217] 例如，一给定文献的文献分数的计算可如下：

[0218] 分数＝.30*（体命中分数）+.70*（锚命中分数）。

[0219] .30 和 .70 的权值可根据需要调整。以如上所述的方式给定查询短语 Qp，则一文献的体命中分数就是所述文献的最高值的相关短语位向量的数值。或者，搜索系统 120 可以通过如下方式直接获得所述值：查找索引 150 中的每个查询短语 Q，从查询短语 Q 的记入列表访问文献，然后存取相关短语位向量。

[0220] 文献 d 的锚命中分数是查询短语 Q 的相关短语位向量的函数，其中 Q 是一引用文献 d 的文献中的锚项。当索引系统 110 为文献库中的文献编制索引时，其为每个短语保存文献列表，其中所述短语是一外链接中的锚文本，同时为每个文献保存来自其他文献的内链接（和相关联的锚文本）。一个文献的内链接是从其他文献（引用文献）到给定文献的引用（例如，超链接）。

[0221] 然后为了确定给定文献 d 的锚命中分数，搜索系统 120 用锚短语 Q 在以索引列出的引用文献 R 集合（i ＝ 1- 引用文献数）上迭代，然后对下列乘积求和：

[0222] Ri.Q. 相关短语位向量 *D.Q. 相关短语位向量。

[0223] 这里的乘积值是表示锚短语 Q 与文献 D 主题相关的程度的分数。这里将此分数称为“入站分数向量”。这个乘积有效地通过引用文献 R 中的锚短语的相关位向量来对当前文献 D 的相关位向量加权。如果引用文献 R 本身与查询短语 Q 相关（且因此具有较高值的相关短语位向量），那么此会增加当前文献 D 分数的有效性。然后，组合体命中分数和锚命中分数以便如上所述产生文献分数。

[0224] 接着，为每个引用文献 R，获得每个锚短语 Q 的相关短语位向量。这是对锚短语 Q 与文献 R 主题相关程度的度量。这里将该值称为“出站分数向量”。

[0225] 然后从索引 150，提取所有（引用文献，被引用文献）对的锚短语 Q。然后通过这些对的相关联的（出站分数向量，入站分数向量）值来对这些对排序。根据实施的不同，这些分量中的任一分量都可作为主排序关键字，而另一个分量可为次排序关键字。然后，将排序后的结果呈现给用户。根据出站分数分量对文献排序会使那些具有许多个与查询相关的短语作为锚命中的文献的等级最高，从而将这些文献表示为“专家”文献。根据入站文献分数排序会使那些因为锚项而经常被引用的文献的等级最高。

[0226] 3. 基于短语的搜索个性化

[0227] 搜索系统 120 的另一个方面是根据用户特定兴趣的模型来自定义搜索结果的分等级或使其个性化 606。以此方式，那些更可能与用户的兴趣相关的文献会排在搜索结果中的较高等级。搜索结果的个性化如下。

[0228] 作为预备，比较有用的是就查询和文献（这两项可用短语表示）定义用户兴趣（例如，用于模型）。对于一个输入搜索查询，一个查询是由查询短语 Q、Qr 的相关短语和查询短语 Qp 的扩展短语 Qe 表示。因此，这组术语和短语表示查询的含义。接着，用与页相关联的短语来表示文献的含义。如上所述，给定查询和文献，从所述文献索引所指的所有短语的体分数（相关位向量）确定所述文献的相关短语。最后，可以按照代表这些元素中的每一元素的短语将用户表示成一组查询与一组文献的并集。可以从用户在先前的搜索结果中选择的文献，或者通常通过浏览语料库（例如，访问互联网上的文献），使用监控用户动作及目的地的用户端工具，来确定所述集合中所包括的代表用户的特定文献。

[0229] 建构和使用用户模型以进行个性化分等级的过程如下。

[0230] 首先，为一给定用户，保存所访问过的最后 K 个查询和 P 个文献的列表，其中 K 和 P 较佳各为约 250。这些列表可以保存在用户帐号数据库中，其中用户是通过注册或通过浏览器 cookies 来辨识。对于一给定用户，这些列表在用户第一次提供查询时将是空的。

[0231] 接着，从用户接收查询 q。以如上所述的方式检索 q 的相关短语 Qr 以及扩展短语。此形成查询模型。

[0232] 在第一遍中（例如，若没有存储用户的任何查询信息），搜索系统 120 运作后只是返回搜索结果中与用户查询相关的文献，而不另外自定义分等级。

[0233] 用户端浏览器工具监控用户通过（例如）点击搜索结果中的文献链接访问了搜索结果中的哪些文献。用于作为选择哪些短语的基础的这些被访问文献将成为用户模型的一部分。对于每个此类被访问文献，搜索系统 120 检索所述文献的文献模型，其为与所述文献相关的一列短语。将每个与所述被访问文献相关的短语添加到用户模型中。

[0234] 接着，给定与一被访问文献相关的短语，可从每个短语的群集位向量确定与这些短语相关联的群集。对于每个群集，通过在含有群集号或如上所述的群集位向量表示的相关短语列表中查找短语来确定作为所述群集的组员的每个短语。然后，将这个群集号添加到用户模型中。此外，对于每个此类群集，保存一个计数器，并在每次将那个群集中的短语添加到用户模型中时，使计数器递增。如下所述，这些技术可用作权。因此，从用户通过存取而表示出兴趣的文献上所存在的群集中所包括的短语建立了用户模型。

[0235] 同样的通用方法可更精确地聚焦在俘获用户表明比仅仅访问文献更高等级的短语信息（对此，用户可能只是在需要是判断文献相关）。例如，将短语收集到用户模型中可能限于那些用户打印、保存、存储为喜爱或链接、电邮给另一用户、或在浏览器窗口中打开一段延长时间（例如，10 分钟）的文献。这些及其他动作都表明对文献的更高等级的兴趣。

[0236] 当从用户接收到另一查询时，检索相关查询短语 Qr。使这些相关查询短语 Qr 与用户模型中所列的短语相交，以便确定所述查询与用户模型中同时存在哪些短语。初始化所述查询的相关短语 Qr 的掩码位向量。如上所述，这个位向量是一个双位向量。对于同时存在于用户模型中的所述查询的每个相关短语 Qr，将此相关短语的两个位设定在掩码位向量中。因此，掩码位向量代表同时存在于查询与用户模型中的相关短语。

[0237] 然后，使用掩码位向量来通过使当前搜索结果集合中的每个文献的相关短语位向

量与所述掩码位向量进行与操作 (ANDing) 来屏蔽所述相关短语位向量。此达到通过掩码位向量调整体分数和锚命中分数的效果。然后，如前所述计算文献的体分数和锚分数并将其呈现给用户。此方法主要需要文献具有包括在用户模型中的查询短语以便排到较高等级。

[0238] 作为一个不会强加前述严格约束的替代实施例，掩码位向量可以排成数组，以便每个位都可用来对用户模型中的相关短语的群集计数加权。因此，每个群集计数都被乘以 0 或 1，从而有效地使计数为 0 或保持原计数。接着，就像使用权那样使用这些计数本身来乘正在计分的每个文献的相关短语。此方法的好处是仍允许适当地计分哪些没有查询短语作为相关短语的文献。

[0239] 最后，可将用户模型限于当前对话，其中对话是搜索中有效时期的时间间隔，在此对话后，转储用户模型。或者，一给定用户的用户模型可持续一段时间，然后使其权值下降或过期。

[0240] IV. 结果显示

[0241] 显示系统 130 从搜索系统 120 接收经过计分和排序的搜索结果，并且执行其他组织、注释和群集操作，然后将结果呈现给用户。这些操作有利于用户理解搜索结果的内容，去除重复结果，并且提供对搜索结果的更有代表性的取样。图 7 说明显示系统 120 的以下主要功能性操作 ：

[0242] 700 ：根据主题群集文献 ；

[0243] 702 ：产生文献说明 ；

[0244] 704 ：去除重复文献。

[0245] 这些操作中的每个操作都与输入搜索结果 701 和输出修正后的搜索结果 703 一起采用。如图 7 所示，这些操作的次序是独立的，且可根据一给定实施例的需要而改变，因此，可以流水线的方式而不是如图所示并行输送这些输入。

[0246] 1. 显示的动态分类产生

[0247] 对于一给定查询，通常会返回几百个、甚至可能几千个满足所述查询的文献。在许多情况下，某些文献虽然彼此内容不同，但其足够相关以形成一群有意义的相关文献，基本上就是一个群集。然而，大多数用户不会看搜索结果中前 30 或 40 个以外的文献。因此，如果前（例如）100 个文献来自三个群集，但接下来的 100 个文献代表另外的 4 个群集，那么在不经进一步的调整下，用户通常不会看后面这些文献，但事实上这些文献可能与用户查询十分相关，因为它们代表了各种与查询相关的不同主题。因此，这里需要为用户提供来自每个群集的样本文献，从而向用户展现来自搜索结果的不同文献的更宽的选集。显示系统 130 如下进行。

[0248] 如同系统 100 的其他方面，显示系统 130 利用搜索结果中每个文献 d 的相关短语位向量。更详细地说，对于每个查询短语 Q，且对于 Q 的记入列表中的每个文献 d，相关短语位向量指示文献中存在哪些相关短语 Qr。然后在搜索结果中的文献集合上，对于每个相关短语 Qr，通过合计对应于 Qr 的位位置的位值来确定表示多少文献含有相关短语 Qr 的计数。当对搜索结果求和及排序时，将指示最频繁出现的相关短语 Qr，其中的每个相关短语 Qr 都将是一文献群集。最频繁出现的相关短语是第一群集，取其相关短语 Qr 作为其名称，对于最高的三到五个群集依此类推。因此，识别了每个最高的群集，取短语 Qr 作为群集的名称

或标题。

[0249] 现在，可以各种方式将来个每个群集的文献呈现给用户。在一应用中，可显示固定数量的来自每个群集的文献，例如每个文献中计分在前 10 的文献。在另一应用中，可显示成比例数量的来自每个群集的文献。因此，如果搜索结果中有 100 个文献，其中 50 个来自群集 1，30 个来自群集 2，10 个来自群集 3，7 个来自群集 4，且 3 个来自群集 5，并且希望只显示 20 个文献，那么文献的选择如下 ：10 个文献来自群集 1，7 个文献来自群集 2，2 个文献来自群集 3，且 1 个文献来自群集 4。然后，在适当的群集名称作为标题下分组后，将各文献展示给用户。

[0250] 例如，假定搜索查询为“blue merle agility training(蓝色默尔敏捷训练)”，对此搜索系统 120 接收到 100 个文献。搜索系统 120 将已经识别“blue merle”和“agility training”作为查询短语。这些查询短语的相关短语为 ：

[0251] “blue merle”：“Australian Sphepherd”、“red merle”、“tricolor”、“aussie”；

[0252] “agility training”：“weave poles”、“teeter”、“tunnel”、“obstacle”、“border collie”。

[0253] 显示系统 130 然后为每个查询短语的每个上述相关短语确定表示含有所述短语的文献数的计数。例如，假定短语“weave poles”出现在 100 个文献中的 75 个文献中，“teeter”出现在 60 个文献中，“red merle”出现在 50 个文献中。那么，第一群集称为“weave poles”，且存在选定数量的来自该群集的文献 ；第二群集称为“teeter”，且同样存在选定数量 ；依此类推。对于一固定显示，可选择 10 个来自每个群集的文献。按比例显示将使用相对于总文献数成比例数量的来自每个群集的文献。

[0254] 2. 基于主题的文献说明

[0255] 显示系统 130 的第二个功能是创建 702 文献说明，所述文献说明可插入每个文献的搜索结果显示中。这些说明以每个文献中所存在的相关短语为基础，因此有助于用户以在内容上与查询相关的方式了解所述文献是关于什么内容。文献说明可以是一般性的，也可以是对用户个性化的。

[0256] a) 一般主题文献说明

[0257] 如上所述，给定一查询，搜索系统 120 先确定查询短语的相关查询短语 Qr 以及扩展短语，然后为查询识别相关文献。显示系统 130 访问搜索结果中的每个文献并执行下列操作。

[0258] 首先，显示系统 130 通过查询短语 Q、相关查询短语 Qr 和扩展短语 Qp 的实例数来对文献句子分等级，进而为文献的每个句子保存这三个方面的计数。

[0259] 然后，通过这些计数来对句子排序，其中第一排序关键字是查询短语 Q 的计数，第二排序关键字是相关查询短语 Qr 的计数，且最后一个排序关键字是扩展短语 Qp 的计数。

[0260] 最后，将排序后的前 N(例如 5) 个句子用作文献说明。可将这组句子格式化，并将其包括在修正后的搜索结果 703 中的文献显示中。对搜索结果中的一定数量的文献重复此过程，并且可以在每次用户请求下一页结果时按要求进行。

[0261] b) 个性化的基于主题的文献说明

[0262] 在提供搜索结果的个性化的实施例中，可同样使文献说明个性化以便反映用户模型中所表示的用户兴趣。显示系统 130 如下进行。

[0263] 首先，如上所述，显示系统通过使查询相关短语 Qr 与用户模型（其列出了出现在由用户访问过的文献中的短语）相交来确定与用户相关的相关短语。

[0264] 然后，显示系统 130 根据位向量本身的值来对这组用户相关短语 Ur 稳定的排序，将排序后的列表预先挂到查询相关短语 Qr 的列表上，并除去任何重复短语。稳定排序保留了同样等级的短语的现有次序。此产生与查询或用户相关的相关短语集合，称为集合 Qu。

[0265] 现在，以类似于上述一般文献说明方法的方式，显示系统 130 使用此有序短语列表作为对搜索结果中的每个文献中的句子分等级的基础。因此，对于一给定文献，显示系统 130 通过每个用户相关短语和查询相关短语 Qu 的实例数来对文献中的句子分等级，并且根据查询计数来对分等级后的句子排序，最后基于每个此类短语的扩展短语数排序。而在以前，排序关键字的次序是查询短语 Q、相关查询短语 Qr 和扩展短语 Qp，但这里的排序关键字的次序是从高到低等级的用户相关短语 Qr。

[0266] 再次地，对搜索结果中的文献重复此过程（按要求或预先）。于是对于每个此类文献，所得文献说明包括来自所述文献的 N 个最高等级的句子。此处，这些句子将是具有最高用户相关短语 Ur 数量的句子，因此代表文献中表示与用户最相关的概念和主题的关键句（至少根据用户模型中所俘获的信息）。

[0267] 3. 重复文献检测和去除

[0268] 在诸如互联网的大语料库中，其中在许多不同位置存在同一文献的多个实例或一文献的多个部分是十分常见的。例如，由一新闻局（例如，美联社 (Associated Press)）产生的一篇给定的新闻文章可能被复制在一打或一打以上的网站或各报纸上。响应搜索查询而包括所有这些重复文献只会使用户负担多余的信息，而不是有用地响应查询。因此，显示系统 130 提供另一个用于识别那些可能彼此重复或接近重复的文献而只在搜索结果中包括这些文献中的一个文献的能力 704。因此，用户接收到更多样化且更强大的结果集合，而不必浪费时间来看那些彼此重复的文献。显示系统 130 所提供的功能性如下。

[0269] 显示系统 130 处理搜索结果集合 701 中的每个文献。对于每个文献 d，显示系统 130 首先确定与所述文献相关联的相关短语 R 的列表。对于这些相关短语中的每个短语，显示系统 130 根据这些短语中的每个短语的出现频率来对文献中的句子分等级，然后选择 N 个（例如，5 到 10 个）最高等级的句子。然后将这组句子与所述文献结合存储。这样做的一个方法是连接这些选定的句子，然后利用散列表来存储文献识别符。

[0270] 接着，显示系统 130 将每个文献 d 的选定句子和搜索结果 701 中的其他文献的选定句子进行比较，如果这些选定句子匹配（在允许误差内），那么就认为所述文献重复，并将其中一个文献从搜索结果中去除。例如，显示系统 130 可以将连接后的句子弄散列，如果散列表中已经具有所述散列值的款目，那这就表明当前文献与不久前散列的文献重复。然后，显示系统 130 可以用所述文献中的一个文献的文献 ID 来更新此表。较佳地，显示系统 130 保持那个具有文献有效性的较高页等级或其他查询无关度量的文献。此外，显示系统 130 可以修正索引 150 以除去重复文献，使得它不会出现在将来的任何查询的搜索结果中。

[0271] 索引系统 110 可直接应用相同的重复文献去除方法。当爬行一文献时，执行上述文献说明方法以获得选定句子，然后将这些句子弄散列。如果散列表已填满，那再次地，新爬行的文献被视为是前一文献的重复文献。同样地，索引系统 110 可以接着包括那个具有较高页等级或其他查询无关度量的文献。

[0272] 上文就一个可能的实施例特别详细地描述了本发明。所属技术领域的技术人员将明白，可在其他实施例中实施本发明。首先，各组件的特定命名、术语的首字母大写、属性、数据结构或任何其他编程或结构方面都不是强制或重要的，实现本发明的机制或其特征可以具有不同的名称、格式或协议。另外，所述系统可以如上所述通过硬件和软件的组合或完全在硬件元件中来实现。而且，本文所描述的各系统组件之间的特定功能性划分仅仅是示范性的而不是强制性的；由单个系统组件执行的功能可以改为由多个组件执行，由多个组件执行的功能可以改为由单个组件执行。

[0273] 上述说明的一部分就信息操作的算法和符号表示介绍了本发明的特征。这些算法说明和表示是数据处理领域的技术人员所用的方法，因此最有效地将其工作内容转给了所属技术领域的其他技术人员。虽然在功能或逻辑上描述了这些操作，但应了解这些操作是由计算机程序实现的。此外，还证实有时可方便地将这些操作排列称为模块或其他功能名称，而不会丧失一般性。

[0274] 除非另外特定指出，否则由上述讨论显而易见，在整篇说明中，利用“处理”或“计算 (computing/calculating”或“确定”或“显示”等术语的论述是指计算机系统或类似电子计算装置的动作和过程，其操纵和转换计算机系统的存储器或寄存器或其他此类信息存储、传输或显示装置内表示为物理（电子）量的数据。

[0275] 本发明的某些方面包括本文所述的算法形式的过程步骤和指令。应注意，本发明的过程步骤和指令可体现在软件、固件或硬件中，当体现在软件中时，可将其下载以驻存在由实时网络操作系统使用的不同平台上并从这些平台操作。

[0276] 本发明还涉及一种用于执行本文所述的操作的设备。这种设备可以根据所需的目的特定地建造，或者其可包括一通用计算机，该计算机可以由一个存储在一可由所述计算机访问的计算机可读媒体上的计算机程序选择性地启动或重新配置。此类计算机程序可以存储在计算机可读存储媒体中，例如（但不限于）任何类型的磁盘（包括软盘）、光盘、CD-OM、磁光盘、只读存储器 (ROM)、随机存取存储器 (RAM)、EPROM、EEPROM、磁卡或光卡、特殊应用集成电路 (ASIC)，或任何类型的适合存储电子指令的媒体，且各自耦接至计算机系统总线。此外，本说明书中提到的计算机可以包括单个处理器，或者可以是采用多个处理器设计以便增加计算能力的架构。

[0277] 本文提出的算法和操作固有地与任何特定的计算机或其他设备无关。各种通用系统也可以与根据本文的教示的程序一起使用，或者可证实可以便利地建造更特殊的设备来设备来执行所需方法步骤。所属技术领域的技术人员将明白各种这些系统所需的结构以及等效变化。此外，并没有参照任何特定的编程语言来描述本发明。可知，可以使用各种编程语言来实现本文所述的本发明的教示，而且提到任何特定语言是为了揭示本发明的实现及最佳模式。

[0278] 本发明很适合众多拓扑学上的各种各样的计算机网络系统。在此领域，大网络的配置和管理包括存储装置和计算机，其在通信上耦合至诸如互联网的网络上的不同计算机和存储装置。

[0279] 最后应注意，本说明书中所用的语言主要是为了可读性和指导性的目的而选择的，也可以不选择这种语言来描绘或限定发明主题。因此，本发明的揭示内容只是想说明而不是限制本发明的范畴，本发明的范畴如权利要求书所述。

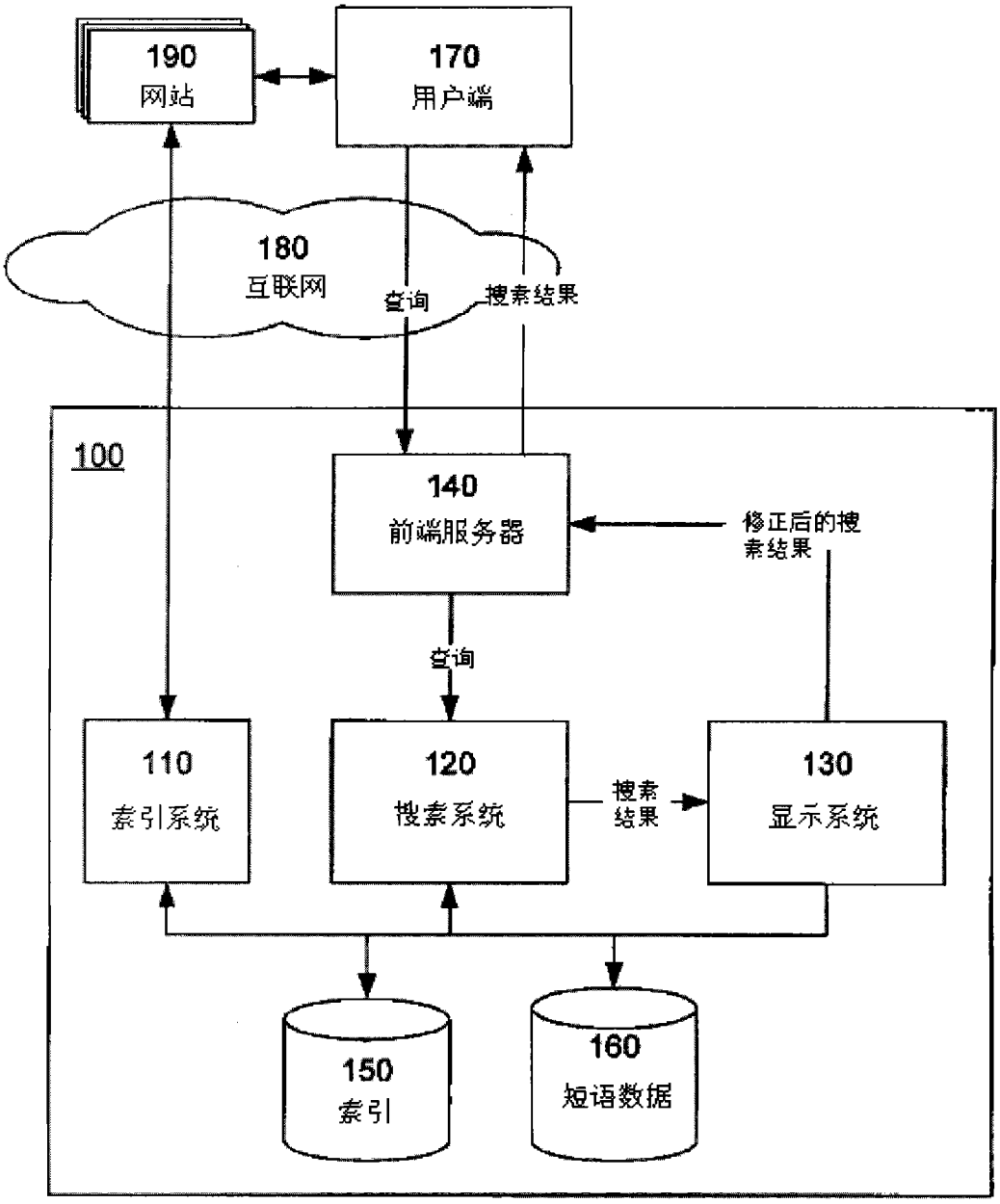

190
网站
170
用户端
180
互联网
查询
搜索结果
100
140
前端服务器
修正后的搜索结果
查询
110
索引系统
120
搜索系统
搜索结果
130
显示系统
150
索引
160
短语数据

图 1

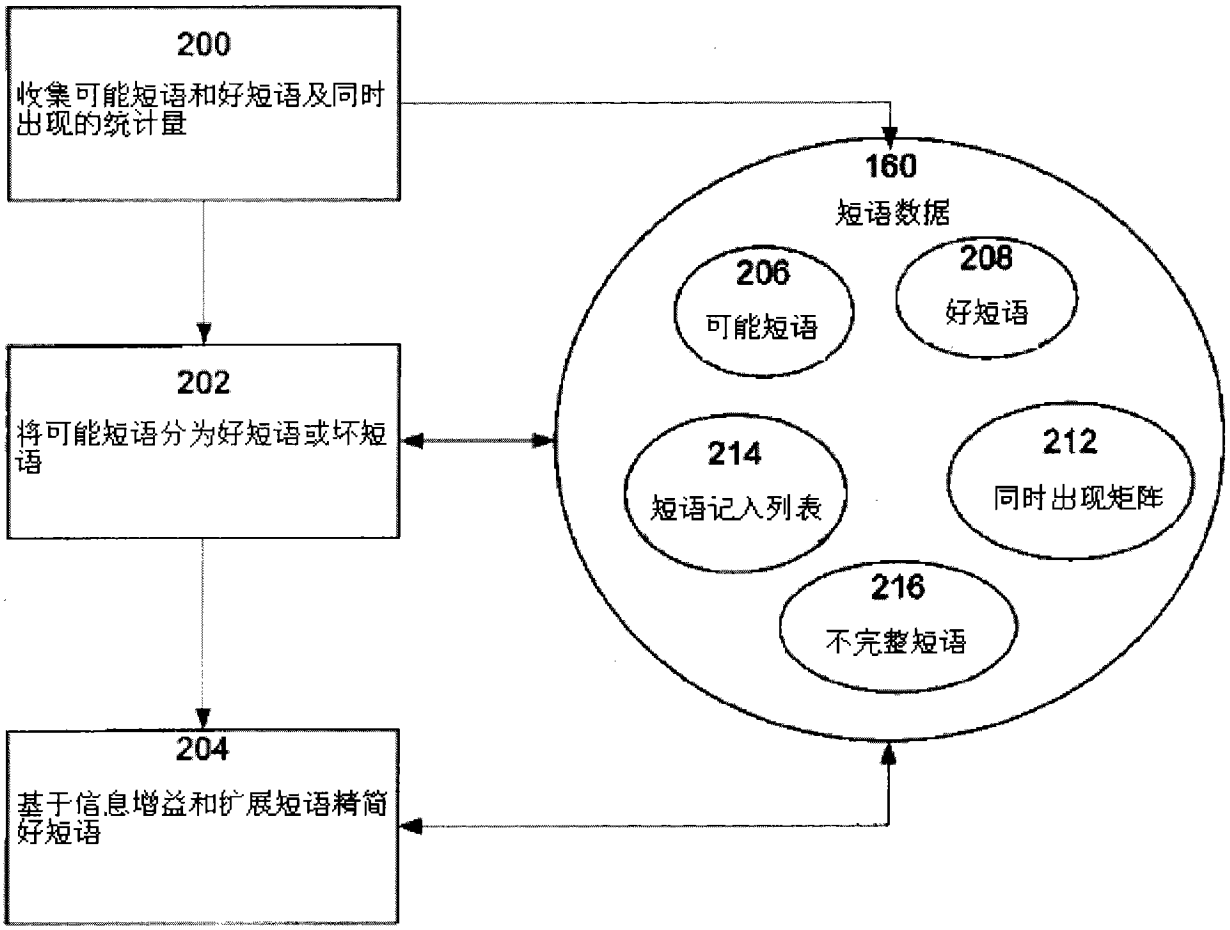

200
收集可能短语和好短语及同时出现的统计量
160
短语数据
206
可能短语
208
好短语
202
将可能短语分为好短语或坏短语
214
短语记入列表
212
同时出现矩阵
216
不完整短语
204
基于信息增益和扩展短语精简好短语

图 2

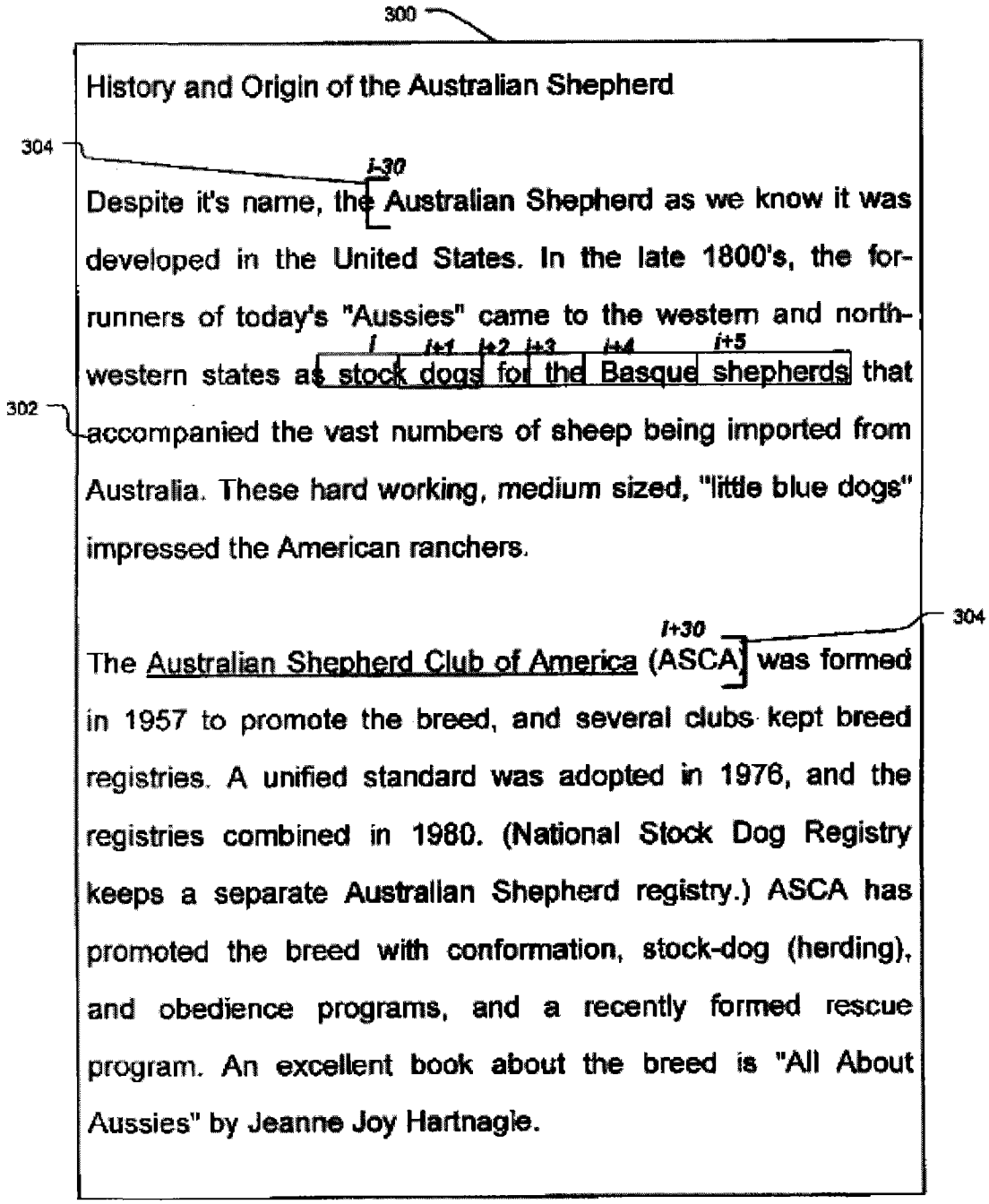

300
History and Origin of the Australian Shepherd
304
i-30
Despite it's name, the Australian Shepherd as we know it was developed in the United States. In the late 1800's, the forrunners of today's "Aussies" came to the western and northwestern states as stock dogs for the Basque shepherds that
i
i+1
i+2
i+3
i+4
i+5
302
accompanied the vast numbers of sheep being imported from Australia. These hard working, medium sized, "little blue dogs" impressed the American ranchers.
i+30
304
The Australian Shepherd Club of America (ASCA) was formed in 1957 to promote the breed, and several clubs kept breed registries. A unified standard was adopted in 1976, and the registries combined in 1980. (National Stock Dog Registry keeps a separate Australian Shepherd registry.) ASCA has promoted the breed with conformation, stock-dog (herding), and obedience programs, and a recently formed rescue program. An excellent book about the breed is "All About Aussies" by Jeanne Joy Hartnagle.

图 3

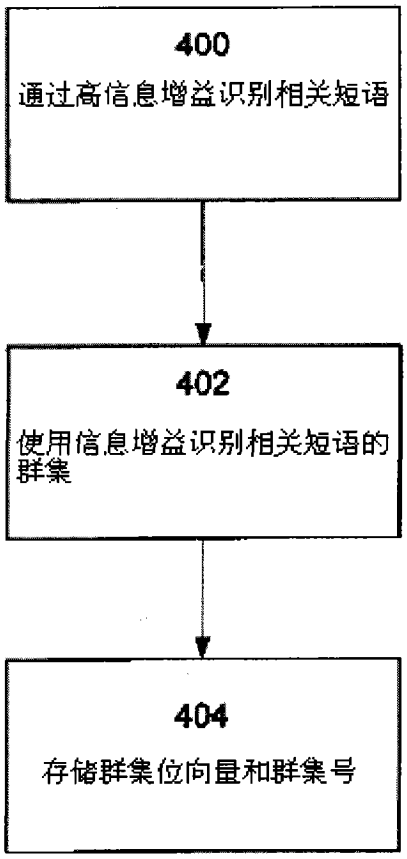

400
通过高信息增益识别相关短语
402
使用信息增益识别相关短语的群集
404
存储群集位向量和群集号

图 4

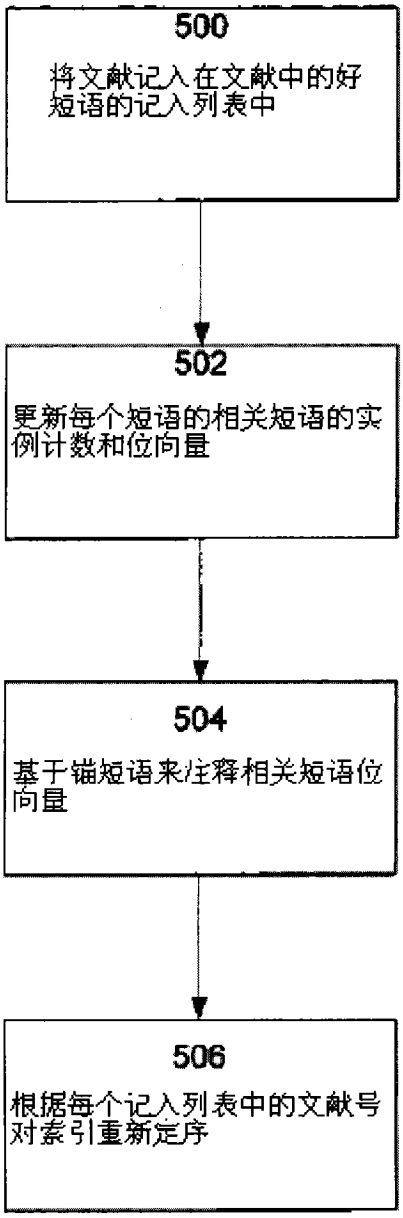

500
将文献记入在文献中的好短语的记入列表中
502
更新每个短语的相关短语的实例计数和位向量
504
基于锚短语来注释相关短语位向量
506
根据每个记入列表中的文献号对索引重新定序

图 5

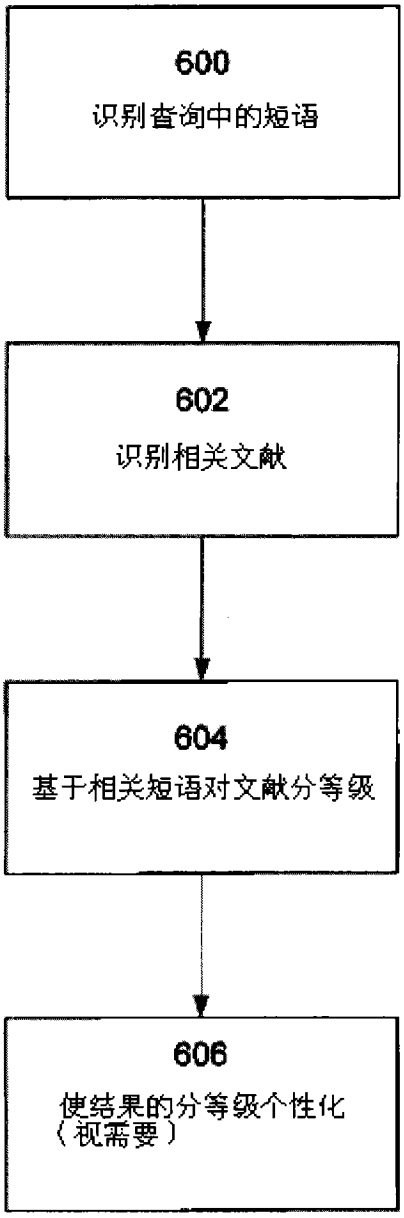

600
识别查询中的短语
602
识别相关文献
604
基于相关短语对文献分等级
606
使结果的分等级个性化
（视需要）

图 6

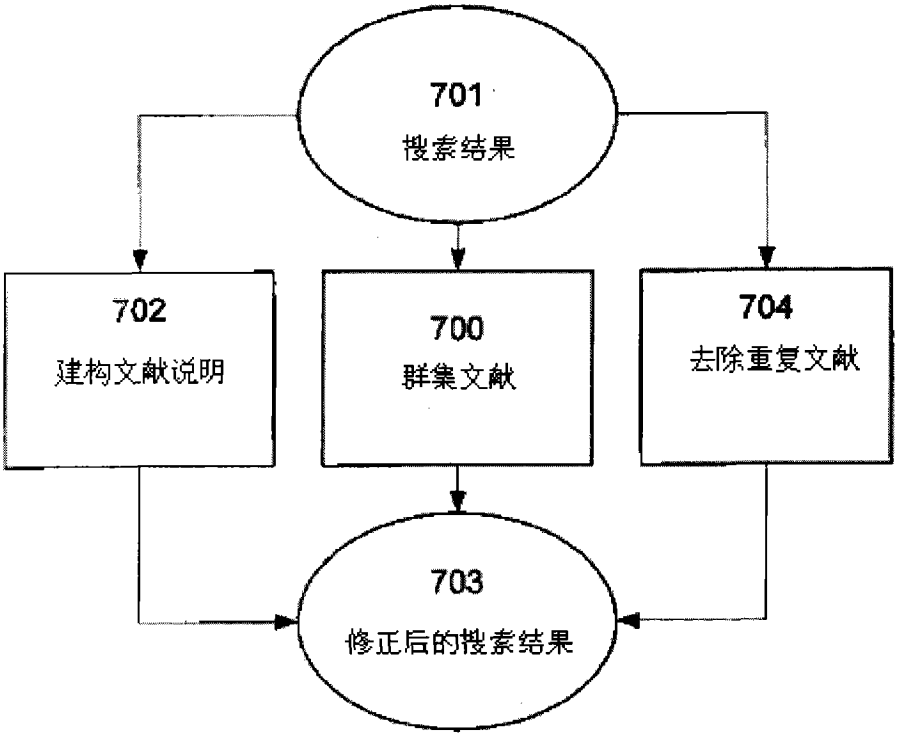

701
搜索结果
702
建构文献说明
700
群集文献
704
去除重复文献
703
修正后的搜索结果

图 7

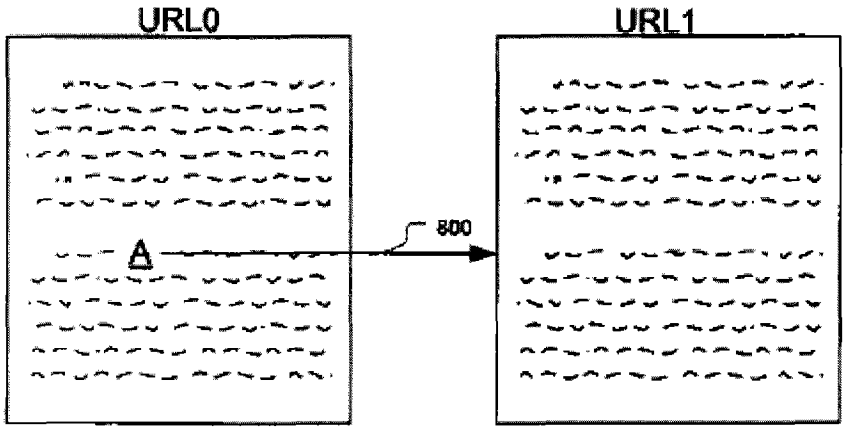

URL0
URL1
A
800

图 8a

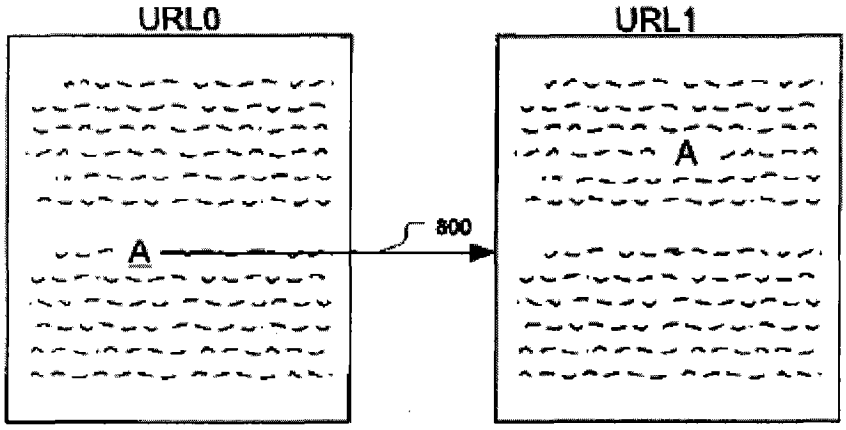


图 8b