(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 101681251 B
(45)授权公告日 2019.05.10

(21)申请号 200880015001.6
(22)申请日 2008.03.14
(65)同一申请的已公布的文献号
申请公布号 CN 101681251 A
(43)申请公布日 2010.03.24
(30)优先权数据
11/728,748 2007.03.27 US
(85)PCT国际申请进入国家阶段日
2009.11.06
(86)PCT国际申请的申请数据
PCT/US2008/056934 2008.03.14
(87)PCT国际申请的公布数据
WO2008/118655 EN 2008.10.02
(73)专利权人 奥多比公司
地址 美国加利福尼亚州
(72)发明人 W·常 N·格哈姆拉维
(74)专利代理机构 北京市金杜律师事务所 11256
代理人 吴立明 姜彦
(51)Int.Cl.
G06F 7/00(2006.01)
(56)对比文件
CN 1910654 A,2007.02.07,
US 7478089 A,2009.01.13,
US 5384703 A,1995.01.24,
US 2007/0016848 A1,2007.01.18,
US 2004/0064438 A1,2004.04.01,
US 2006/0212469 A1,2006.09.21,
审查员 庄湧

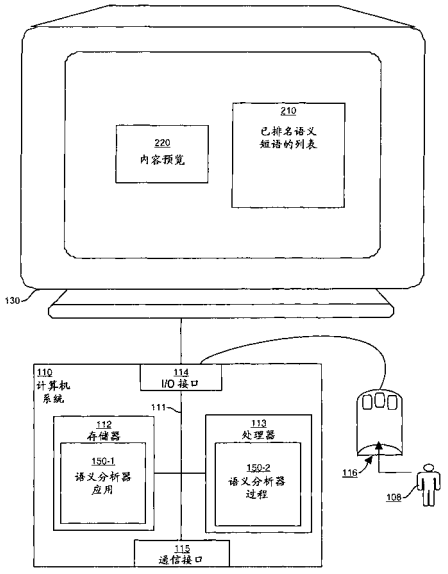
权利要求书5页 说明书8页 附图6页

(54)发明名称

从文档到排名短语的语义分析

(57)摘要

一种为语义分析器提供的方法、装置和计算机产品，用于产生和排名语义短语以反映它们与文档主题和话题的关联性。在语义分析器执行文本提取之前，文本和文档可以与任意预先选择的关键字没有关系。语义分析器从文档中提取文本，并且对提取出的文本执行语义分析。作为该语义分析的结果，语义分析器提供多个已排名的语义短语，并且将语义短语作为语义关键字与该文档相关联。语义短语定义出与该文档一起被呈现的内容，该内容是广告、到远程信息资源的链接和第二文档。



CN 101681251 B

1.一种由计算机实施的方法,包括:

从文档中提取文本;

对从所述文档提取出的所述文本执行语义分析,其中执行语义分析包括:针对从所述文档中提取的至少一个词组,

(i)为从所述文档中提取的每个单词标识单词频率,所述单词频率表示相应的单词多频繁地出现在所述文档中;

(ii)标识所述词组在所述文档中的出现的至少一个文档位置;

(iii)确定指示所述文档的内容的至少一个主题的至少一个语义短语,所述至少一个主题是基于所述词组的所述出现的所述至少一个文档位置并且关于所述相应词组中所使用的至少一个单词的单词频率的;

作为所述语义分析的结果,提供多个已排名的语义短语;和

将至少一个语义短语与所述文档相关联,所述至少一个语义短语定义出将与所述文档一起被呈现的内容;

其中确定所述至少一个语义短语包括:

标识整个所述文档中第一词组的放置模式;以及

基于所述第一词组的所述放置模式并且关于所述第一词组中每个单词的单词频率,确定所述至少一个语义短语,其中整个所述文档中至少出现一次的至少一个第二词组使用所述第一词组中发现的特定单词,其中针对所述特定单词的单词频率反映所述第一词组和所述第二词组对所述特定单词的使用;

其中标识整个所述文档中第一词组的放置模式包括:

标识由所述文档的内容体现的文章的类型;

获取与所述文档的文章类型相对应的期望的文章结构;

检测所述第一词组的第一相邻出现对的文档位置之间的第一距离;

检测所述第一词组的第二相邻出现对的文档位置之间的第二距离;以及

基于关于所述期望的文章结构的所述第一距离和所述第二距离,标识所述放置模式。

2.根据权利要求1所述的由计算机实施的方法,其中所述至少一个语义短语定义出将与所述文档一起被呈现的内容包括:定义出包括广告、到远程信息资源的链接和第二文档的组中的至少一个。

3.一种由计算机实施的方法,包括:

从文档中提取文本,其中从所述文档中提取所述文本包括从所述文档中提取纯文本,所述文本和所述文档缺乏到一个或多个预选择的关键词的关系;

对从所述文档中提取的所述文本执行语义分析;

将至少一个语义短语与所述文档相关联,所述至少一个语义短语定义出将与所述文档一同呈现的内容,其中从所述文档提取纯文本还包括:

在提取的所述纯文本中标识至少一个标志,所述标志表示文档中的文本与字符的串;

为表示所述标志在所述文档中出现的总次数的标志值列表;

在提取的所述纯文本中标识至少一个短语,所述短语包含标志的连续分组;

为短语x_j的短语值n列表,所述短语值n表示所述短语在所述文档中出现的总次数;和

为所述短语的至少一个提及偏置列表,所述提及偏置offset(m_i)表示所述短语x_j在一

组所述短语的n次出现中的单个出现m_i在所述文档中的位置，其中n等于或大于1；

其中将所述至少一个语义短语关联到所述文档中包括：将所述至少一个语义短语插入到所述文档的元数据部分中；

其中将所述至少一个语义短语插入到所述文档的所述元数据部分中包括将所述至少一个语义短语插入到所述文档的可扩展元数据平台XMP部分中；

其中由计算机实施的所述方法还包括：向所述文档指派策略，所述策略支持所述文档将与所述文档相关联的所述至少一个语义短语利用作为语义关键词，以便于在所述文档被呈现时显示所述内容，所述策略还包括支持所述文档略过与所述文档相关联的所述至少一个语义短语，以便于在所述文档被呈现时不显示所述内容。

4.根据权利要求3所述的由计算机实施的方法，进一步包括计算至少一个短语统计量，所述至少一个短语统计量从包含下列的组中选择：

作为所述短语中标志的标志值的函数，来计算短语的标志频率$tf(x_j)$，所述函数包括所述短语中标志的标志值的平均值和中间值中的至少一个；

计算所述短语的中间偏置$moffset(x_j)$；和

计算所述短语的偏置标准差$soffset(x_j)$。

5.根据权利要求4所述的由计算机实施的方法，进一步包括：当所述文档是讨论至少一个中心主题的长文章时，计算文章分值$ascore(x_j)$，其中所述文章分值由以下等式所定义：

$$ascore(x_j)=\frac{n(soffset(x_j))\sqrt{tf(x_j)}}{\sqrt{moffset(x_j)}}。$$

6.根据权利要求4所述的由计算机实施的方法，进一步包括：当所述文档是科技文章时，计算科技文章分值$tscore(x_j)$，所述科技文章分值包括：

计算所述短语的两个单个出现的至少一个差值r；

计算中间差距$r(x_j)$；和

计算差距标准差$rsdiff(x_j)$，

其中科技文章分值由以下等式所定义：

$$tscore(x_j)=\frac{n}{rsdiff(x_j)}。$$

7.根据权利要求4所述的由计算机实施的方法，进一步包含：当所述文档是信件时，计算标准差信件分值$sdlscore(x_j)$和计算微频信件分值$mflscore(x_j)$，其中所述标准差信件分值由以下等式所定义：

$sdlscore(x_j)=n(soffset(x_j))$，以及

所述微频信件分值由以下等式所定义：

$$mflscore(x_j)=n\sqrt{tf(x_j)}(soffset(x_j))。$$

8.根据权利要求4所述的由计算机实施的方法，其中作为所述语义分析的结果，提供多个已排名的语义短语包括：

提供所述多个排名的语义短语的列表，所述列表根据一个或多个短语统计量对所述语义短语排名；

从所述列表中选择一个或多个语义短语，以创建由选择的一个或多个语义短语中每一

个所定义的内容的预览;所述预览显示由选择的一个或多个语义短语中每一个所定义的内容的聚集;和

从所述列表中指定一个或多个语义短语,以作为一个或多个语义关键字与所述文档相关联。

9.一种非暂态计算机可读存储介质,所述非暂态计算机可读存储介质包括编码于其上的可执行指令,所述可执行指令在计算化设备上可操作以执行处理,所述可执行指令包括:

用于从文档中提取文本的指令,其中用于从所述文档中提取文本的指令包括:用于从所述文档中提取纯文本的指令,所述文本和所述文档缺乏到一个或多个预选择的关键词的关系;

用于对从所述文档提取出的所述文本执行语义分析的指令;

用于作为所述语义分析的结果,提供多个已排名的语义短语的指令;和

用于将至少一个语义短语与所述文档相关联的指令,所述至少一个语义短语定义出与所述文档一起被呈现的内容;

其中用于提取纯文本的所述指令包括:

用于在提取的所述纯文本中标识至少一个标志的指令,所述标志表示所述文档中的文本与字符的串;

用于为表示所述标志在所述文档中出现的总次数的标志值列表的指令;

用于在提取的所述纯文本中标识至少一个短语的指令,所述短语包含标志的连续分组;

用于为短语x_j的短语值n列表的指令,所述短语值n表示所述短语在所述文档中出现的总次数;以及

用于为所述短语的至少一个提及偏置列表的指令,所述提及偏置$offset(m_i)$表示短语x_j在一组所述短语的n次出现中的单个出现m_i在所述文档中的位置,其中n等于或大于1;

其中用于将所述至少一个语义短语关联到所述文档中的指令包括:用于将所述至少一个语义短语插入到所述文档的元数据部分中的指令;

其中用于将所述至少一个语义短语插入到所述文档的所述元数据部分中的指令包括用于将所述至少一个语义短语插入到所述文档的可扩展元数据平台XMP部分中的指令;

用于向所述文档指派策略的指令,所述策略支持所述文档将与所述文档相关联的所述至少一个语义短语利用作为语义短语,以便于在所述文档被呈现时显示所述内容,所述策略还支持所述文档略过与所述文档相关联的所述至少一个语义短语,以便于在所述文档被呈现时不显示所述内容。

10.根据权利要求9所述的非暂态计算机可读存储介质,其中所述至少一个语义短语定义出与所述文档一起被呈现的内容包括:用于定义出包括广告、到远程信息资源的链接和第二文档的组中的至少一个的指令。

11.根据权利要求9所述的非暂态计算机可读存储介质,进一步包括:用于计算至少一个短语统计量的指令,所述用于计算至少一个短语统计量的指令从包括下列的组中选择:

用于作为短语中标志的标志值的函数,来计算短语的标志频率$tf(x_j)$的指令,所述函数包括短语中标志的标志值的平均值和中间值中的至少一个;

用于计算所述短语的中间偏置$moffset(x_j)$的指令;和

用于计算所述短语的偏置标准差soffset(x_j)的指令。

12.根据权利要求11所述的非暂态计算机可读存储介质，进一步包括：用于当所述文档是讨论至少一个中心主题的长文章时计算文章分值ascore(x_j)的指令，其中所述文章分值由以下等式所定义：

$$ascore(x_j)=\frac{n(soffset(x_j))\sqrt{tf(x_j)}}{\sqrt{moffset(x_j)}}。$$

13.根据权利要求11所述的非暂态计算机可读存储介质，进一步包括：用于当所述文档是科技文章时计算科技文章分值tscore(x_j)的指令，所述用于计算所述科技文章分值的装置包括：

用于计算所述短语的两个单个出现的至少一个差值r的指令；

用于计算中间差距r(x_j)的指令；和

用于计算差距标准差rsdiff(x_j)的指令，

其中所述科技文章分值由以下等式所定义：

$$tscore(x_j)=\frac{n}{rsdiff(x_j)}。$$

14.根据权利要求11所述的非暂态计算机可读存储介质，进一步包括：用于当所述文档是信件时，计算标准差信件分值sdlscore(x_j)的指令和用于计算微频信件分值mflscore(x_j)的指令，其中所述标准差信件分值由以下等式所定义：

sdlscore(x_j)＝n(soffset(x_j))，以及

所述微频信件分值由以下等式所定义：

$$mflscore(x_j)=n\sqrt{tf(x_j)}(soffset(x_j))。$$

15.根据权利要求9所述的非暂态计算机可读存储介质，其中用于作为所述语义分析的结果，来提供多个已排名的语义短语的指令包括：

用于提供所述多个已排名的语义短语的列表的指令，所述列表根据一个或多个短语统计量对所述语义短语排名；

用于从所述列表中选择一个或多个语义短语以创建由选择的一个或多个语义短语中每一个所定义的内容的预览的指令；所述预览显示由选择的一个或多个语义短语中每一个所定义的内容的聚集；和

用于从所述列表中指定一个或多个语义短语以作为一个或多个语义关键字与所述文档相关联的指令。

16.一种计算机系统，包括：

处理器；

存储器单元，其存储与所述处理器执行的应用相关联的指令；和

互连单元，其耦合所述处理器和存储器单元，使所述计算机系统能够执行所述应用而且执行操作：

从文档中提取文本，其中用于从所述文档中提取文本的指令包括：用于从所述文档中提取纯文本的指令，所述文本和所述文档缺乏到一个或多个预选择的关键词的关系；

对从所述文档提取出的文本执行语义分析；

作为所述语义分析的结果，来提供多个已排名的语义短语；和

将至少一个语义短语与所述文档相关联，所述至少一个语义短语定义出与所述文档一起被呈现的内容，其中从所述文档提取纯文本还包括：

在提取的所述纯文本中标识至少一个标志，所述标志表示文档中的文本与字符的串；

为表示所述标志在所述文档中出现的总次数的标志值列表；

在提取的所述纯文本中标识至少一个短语，所述短语包含标志的连续分组；

为短语x_j的短语值n列表，所述短语值n表示所述短语在所述文档中出现的总次数；和

为所述短语的至少一个提及偏置列表，所述提及偏置offset(m_i)表示短语x_j在一组所述短语的n次出现中的单个出现m_i在所述文档中的位置，其中n等于或大于1；

计算至少一个短语统计量，所述至少一个短语统计量从包含下项的组中选择：

作为所述短语中的所述标志的所述标志值的函数，来计算短语的标志频率tf(x_j)，所述函数包括所述短语中所述标志的所述标志值的平均值和中间值中的至少一个；

计算所述短语的中间偏置moffset(x_j)；和

计算所述短语的偏置标准差soffset(x_j)；

当所述文档是讨论至少一个中心主题的长文章时，计算文章分值ascore(x_j)，其中所述文章分值由以下等式所定义：

$$ascore(x_j)=\frac{n(soffset(x_j))\sqrt{tf(x_j)}}{\sqrt{moffset(x_j)}};$$

当所述文档是科技文章时，计算科技文章分值tscore(x_j)，所述科技文章分值包括：

计算所述短语的两个单个出现之间的至少一个差值r；

计算中间差距r(x_j)；以及

计算差距标准差rsdiff(x_j)，其中所述科技文章分值由以下等式所定义：

$$tscore(x_j)=\frac{n}{rsdiff(x_j)};$$

当所述文档是信件时，计算标准差信件分值和计算微频信件分值，其中所述标准差信件分值由以下等式所定义：

sdlscore(x_j)＝n(soffset(x_j))，以及

所述微频信件分值由以下等式所定义：

$$mflscore(x_j)=n\sqrt{tf(x_j)}(soffset(x_j))。$$

从文档到排名短语的语义分析

背景技术

[0001] 用于向潜在消费者呈现广告的传统技术提供了在其中呈现出这些广告的多个媒介。目前，与杂志出版商在他们的杂志中出售广告空间相类似，人们可以通过出售文档中的空间，而在他们的非网站内容文档中布置广告。广告也可以例如经由广告横幅而显示在网站上。另外，广告可以经由赞助广告而显示在搜索引擎上。在目标广告中，广告商通过选择关键字或关键字词组，并通过与也期望这些关键字或关键字词组相关的网站上出现他们广告的其它广告商竞争而为广告付费。

[0002] 当最终用户登录包含广告的网站时，广告(广告商已经为关键字或关键字词组出价)被显示。广告的显示称为“印象(impression)”。广告商并不为印象付费。相反，广告商为他们的目标消费者最有可能在为找到他们的产品类型或服务类型而在搜索栏中输入的“关键字”而出价。当最终用户选择(即，“点击”)广告时，广告商因该选择而被计费。无论广告商为导致广告显示(即，印象)的关键字或关键字词组出价多少，广告商都被计费。最终用户每点击一次该广告，广告商都由于该选择被计费。例如，如果广告商出售蓝点(blue dot)，则广告商将为关键字“蓝点”出价，并期望用户在搜索栏中键入这些字、浏览他们的广告、点击这个广告并且最终购买。这些广告称为“赞助链接”或“赞助广告”，并出现在搜索引擎的网页上邻近搜索结果处或在搜索结果之上。广告商只是当最终用户选择(即，“点击”)广告时才为广告付费，这是已知的“按每次点击支付”。每当最终用户选择(即，“点击”)出现在网站所有者的网站上的广告时，该网站所有者也会有少量的收入。

发明内容

[0003] 传统的计算机系统有许多缺陷。例如，这些传统系统(例如，传统广告内容服务系统)需要手动输入被用于确定广告关联性的关键字。具体地，搜索引擎通常依赖诸如关键字元标签的网站管理者提供的信息。这里，网站管理者可以向期望的搜索引擎提交网页或URI(统一资源标识符)，该搜索引擎会发出蜘蛛(spider)以“爬取”该页面、从中提取到其它页面的链接以及在该网页中发现的索引信息。搜索引擎蜘蛛下载该页面，并且在搜索引擎自己的服务器时存储该页面，其中，在搜索引擎自己的服务器中有称为索引器的第二程序，其提取关于该页面的各种信息。由于网站管理者通常通过加入不相关的关键字以虚假增加他们网站的页面印象以及增加他们的广告收入，而经常滥用提交的元标签，所以基于元数据来为页面编制索引并不很可靠。由于这会导致为不相关的搜索创建条件以及创建具有不准确搜索结果排名的搜索，所以网站管理者负面地影响搜索引擎的声誉。

[0004] 传统系统的另外缺陷在于针对从广告商群体(pool)收集的广告关键字来匹配用户查询(或用户创建的电子邮件文本)。为了实施这样的技术，传统搜索引擎依赖跨越巨量内容搜索基础架构的数据挖掘。维持这样的基础架构花费高昂，而且久而久之，随着存储的数据量的显著增加而变得复杂。由于还必须创建和存储为了高效挖掘数据而需要与预定广告关键字的复杂联系和数据关系，并且还要保持其容易访问且响应于用户查询及时被处理，这点尤其可以体会。

[0005] 这里讨论的技术显著克服了诸如上面已讨论那些之类的传统应用和现有技术已知的其它技术的缺陷。正如进一步被讨论，这里的某些特定实施方式是针对语义分析器的。这里描述的语义分析器的一个或多个实施方式与传统系统相比，允许自动分析文档以识别关键字。

[0006] 例如，用户具有需要关键字的大文档，以确保与文件内容相关的广告将与该文件一起被呈现。用户不是该文档的作者，而且仅具有关于该文档内容的相关主题的浅显知识。虽然如此，该用户期望还是想要选择能提供恰当反映该文档的主题与内容的广告的关键字。在一个实施方式中，该语义分析器能够自动向用户提供这样的关键字。

[0007] 与用户浪费时间阅读整篇文档不同，语义分析器自动分析文档中的短语，而且评估文档内容与短语的关联性。语义分析器提取短语(短语是包括一个或多个并列字、标点、数字和/或缩写的组)的所有出现(即，提及)、对提取的短语执行统计功能(或计算)，而且基于与文档内容的主题的相关程度来产生相异语义短语的统计排名类。

[0008] 基于反馈广告的优良程度，语义分析器允许用户审核和编辑语义短语的列表。一旦用户预览出语义短语反馈广告的优良程度，语义分析器便允许用户将期望的语义短语作为语义关键字关联回到文档中。

[0009] 因此，语义分析器向用户提供了这样的优势，即，与用户对文档主题对象的理解无关，而是从基于从文档中提取的文本的列表已排名语义短语中确定关键字。而且，允许该用户当与文档一起呈现广告时，预览具体某个语义短语如何最终执行。

[0010] 特别地，语义分析器对语义短语进行排名，以反映它们与文档的主题和话题的关系。在语义分析器执行文本提取之前，文本和文档可以与任意预先选择的关键字没有关系。语义分析器从文档中提取文本，并且对提取出的文本执行语义分析。语义分析器提供多个已排名的语义短语作为该语义分析的结果，和将语义短语作为语义关键字与该文档相关联。语义短语定义出与该文档一起被呈现的内容，该内容是广告、到远程信息资源的链接或第二文档。

[0011] 这里披露的其它实施方式包括任意类型的计算化设备、工作站、手持或膝上型计算机，或配置有软件和/或电路(例如，处理器)以处理这里披露的方法的全部或任意操作的其它设备。换句话说，被编程或配置以如这里解释的来进行操作的诸如计算机或数据通信设备或任意类型处理器的计算化设备，视为这里披露的实施方式。这里披露的其它实施方式包括用于执行上面概括和下面详细披露的步骤和操作的软件程序。一种这样的实施方式包括具有编码其上的计算机程序逻辑的计算机可读介质的计算机程序产品，当在具有存储器和处理器耦合的计算化设备中执行时，为处理器编程以执行这里披露的操作。典型地，这些布置被提供为：布置或编码在诸如光学介质(例如，CD-ROM)、软盘或硬盘、或一或多个ROM或RAM或PROM芯片中诸如固件或微代码的其它介质的计算机可读介质上的软件、代码和/或其它数据(例如，数据结构)，或作为专用集成电路(ASIC)。软件或固件或其它这种配置可以安装到计算化设备中，以引起该计算化设备执行作为在这里披露的实施方式所解释的技术。

[0012] 可以理解，这里披露的系统可以仅实施为软件程序，软件和硬件，或仅硬件。这里披露的实施方式，可以在数据通信设备和这些设备的其它计算化设备和软件系统中使用，诸如由美国加利福尼亚圣何塞的Adobe系统集成公司制造的那些，此后在这里称为“Adobe”

和“Adobe系统”。

附图说明

[0013] 根据下面对配置用于提供反映文档主题和话题的语义短语排名列表的语义分析器的方法和装置的实施方式的更加特定描述，如在各个不同视图中用相同参考符号表示相同部分的附图中所图示，本发明的前述以及其它目标、特点和优点将变得明显。附图侧重于图示出与本发明一致的方法和装置的实施方式、原则和概念，并没有必要严格依照比例。

[0014] 图1为配置有根据这里的实施方式的语义分析器的计算机系统的框图。

[0015] 图2为配置有根据这里的实施方式的语义分析器的计算机系统的框图。

[0016] 图3为由语义分析器执行的处理步骤的流程图。

[0017] 图4为根据这里的实施方式，由语义分析器执行对提取文本的语义处理的处理步骤的流程图。

[0018] 图5为根据这里的实施方式，由语义分析器执行对提取文本的统计处理的处理步骤的流程图。

[0019] 图6为根据这里的实施方式，由语义分析器执行提供语义短语的排名列表以及内容预览的处理步骤的流程图。

具体实施方式

[0020] 根据这里的实施方式，语义分析器被配置以提供反映本文档的主题和话题的语义短语的排名列表。这样的已排名语义短语可以由用户选择，以作为文档的关键字。具体地，在语义分析器执行文本提取之前，文本和文档可以与任意预先选择的关键字没有关系。语义分析器从文档中提取文本，并且对提取出的文本执行语义分析。以作为语义分析的结果，语义分析器提供多个已排名语义短语，并且将语义短语作为语义关键字与该文档相关联。语义短语定义出与文档一起呈现的内容，该内容是广告、到远程信息资源的链接或第二文档。

[0021] 可以理解，标志是包含文档中单词(误拼写的或虚构的)，以及来自标点、缩写和数字等字符的任意串。短语可以是标志的连续组，诸如两或三个单词。另外，提及是短语在文档中的实例，或者换句话说，短语在文档中的单次出现。关键字是具有特别意义的单词或概念，诸如被分离出或指定为具有特别意义的短语。语义分析器根据由语义分析器的计算所推导的关联分数，对在文档中发现的短语进行排名。语义分析器接着向用户提供机会以便预览已排名短语，以及选择哪个已排名短语作为文档的关键字。

[0022] 现在转到图1，框图示出根据这里的实施方式，实施、运行、翻译、操作或以其他方式执行语义分析器应用150-1和/或语义分析器过程150-2(例如，由用户108控制的应用150-1的执行版本)的计算机系统110的架构实例。计算机系统110可以是任意类型的计算化设备，诸如个人计算机、工作站、便携式计算设备、控制台、膝上型计算机、网络终端或类似。

[0023] 如在当前实例中所示，计算机系统110包括诸如数据总线、主板或耦合存储器系统112、处理器113、输入/输出接口114和显示器130的其它电路的互连结构111。输入设备116(例如，诸如键盘、鼠标、触摸板等的一个或多个用户/开发者控制设备)通过输入/输出(I/O)接口114耦合到计算机系统110和处理器113。

[0024] 存储器系统112可以是任意类型的计算机可读介质，以及在这个实例中，其编码有支持这里将进一步解释的功能操作的产生、显示和实施的自封闭时间线修剪器应用150-1。例如，可以在显示器130提供已排名语义短语210的列表。而且，可以响应于用户108从已排名语义短语210的列表中选择一个或多个语义短语，提供内容预览220。

[0025] 在计算机系统110的操作中，处理器113经由互连111访问存储器系统112，以发起、运行、执行、翻译或以其他方式执行语义分析器应用150-1的逻辑指令。以这种方式执行语义分析器应用150-1会产生语义分析器过程150-2。换句话说，语义分析器过程150-2表示语义分析器应用150-1的于运行时在计算化设备110中的处理器113内部或其上执行或实施的一个或多个部分或运行时实例(或整个应用150-1)。

[0026] 关于图2，示出了配置有根据这里的实施方式的语义分析器150的计算机系统110的方框图。语义分析器150包括文本提取器150-1、语义处理器150-2、短语存储150-3、统计量处理器150-4，以及报告组件150-5。另外，该计算机系统100包括显示器130，以提供针对文档200的语义短语的已排名列表210的视图以及由所选择的语义短语定义的内容的预览220。

[0027] 一旦将文档200提交到计算机系统110，会由语义分析器150处理。开始，文本提取器150-1使用例如标准开源文本提取从文档200提取纯文本。可以理解，纯文本可以表示ASCII格式的文本数据。纯文本可以是每个机器或计算器上的几乎所有应用支持的便携式格式，不包含任意的格式命令。接着，纯文本被提交到语义处理器150-2。

[0028] 语义处理器150-2可以包括一个或多个短语、提及和主题提取器。语义处理器150-2另外还可以包括具有结束词消除的标志化器(tokenizer)，以从文档文本提取一序列标志并且移除诸如介词、限定词和逻辑算子的不重要单词(例如，不重要文本，不重要文本串)。语义处理器150-2识别标志、短语；计数短语和标志的单次出现(即，提及)；以及记录在文档中针对每个短语的每个提及开始的位置(即，提及偏置)。

[0029] 在短语存储150-3中，所有提取的语义短语、提及和主题可以存储在一组标准化的关系数据库表中，以允许对提取数据的有效统计处理。例如，在一个实施方式中，关系模式可以每行存储一个提及/主题，以允许使用SQL统计和分析函数的排名处理。

[0030] 另外，在统计处理器150-4中，可以对从文档200提取的所有短语施加几个统计排名函数，以确定最相关和重要的语义短语。可以理解，对于某些类型的文档内容，可能适合不同的统计和排名方法。因此，语义分析器150为许多类型的文档提供语义短语。这种类型的文档包括但是不局限于，诸如具有一个中心主题和少量子话题的新闻故事的长文章、诸如具有大量领域词汇和公知的文档结构(摘要、介绍、结论、参考，等等)的会议和期刊会议记录的科技期刊和科技论文，电子邮件，信件，以及特征在于较少文档结构而且可能具有许多相同重要性的不同想法的通告。

[0031] 最后，在报告组件150-5中，根据统计数据创建语义短语的已排名列表210。创建已排名列表210的视图，并且经由显示器130提供给用户108。语义分析器150允许用户108测试哪些已排名语义短语对于文档200是最佳的。经由报告组件150-5，用户108从列表210中选择一个或多个语义短语。内容预览220被创建并显示给用户。内容预览220向用户显示由选择的语义短语所界定的内容的集合视图。

[0032] 图3为由语义分析器150执行的处理步骤310-340的流程图300。流程图300中的步

骤涉及在图2的框图中示出的特征。语义分析器150(例如,图1中的语义分析器应用150-1和/或语义分析器过程150-2)自动分析文档以根据语义关联性来排名短语。具体地,在步骤310,语义分析器150从文档200提取文本,而且在步骤315,从文档200提取纯文本。可以理解,在文本提取之前,文本和文档200与任意预先选择的一个或多个关键字没有关系。在步骤320,语义分析器150对提取出的文本执行语义分析。在步骤330,作为语义分析的结果,语义分析器150提供多个已排名的语义短语,而且在步骤340,将语义短语与该文档200相关联。语义短语定义出与文档200一起呈现的内容,该内容是广告、到远程信息资源的链接或第二文档。

[0033] 图4为根据这里的实施方式,由语义分析器150执行对提取文本的语义处理的处理步骤410-450的流程图400。流程图400中的步骤涉及在图2的框图中示出的语义处理器150-2。在步骤410,语义分析器150在提取的纯文本内识别至少一个标志,该标志表示文档中的一串文本和字符。具体地,语义分析器150基于在文档200中的出现而保留标志(例如,文档200中的单词)。然而,不需要保留标点、算数符号、数字、介词、限定词和逻辑算子。

[0034] 在步骤420,语义分析器150为表示标志在文档中出现的总次数的标志值列表。在步骤430,语义分析器150在提取的纯文本中识别至少一个短语,该短语包含相邻标志的分组。在步骤440,语义分析器150为短语(x_j)的短语值(n)列表,该值表示该短语在文档中出现的总次数。

[0035] 例如,假设有关于清醒梦境的主题内容的文档200具有在文档200中出现的每个短语。短语“梦状态”提及(例如,单个出现)2次。短语“清醒梦境”提及(例如,单个出现)8次。短语“我的梦提及4次,而短语“梦世界”仅仅提及1次。每个短语都指定有该短语在文档中单个出现的总次数的提及值。因此,由于提及(例如,单个出现)了2次短语“梦状态”,所以“梦状态”的提及值为2。

[0036] 在这些短语中,标志“梦”都出现其中,出现总共15次。诸如“梦”的标志的标志出现值是标志作为短语的一部分在文档的所有提取短语中出现的总次数。因此,“梦”的标志出现值是15。

[0037] 在步骤450,语义分析器150为短语的至少一个提及偏置(offset)列表,该提及偏置表示短语在文档中单个出现(即,提及)的位置。例如,短语“清醒梦境”在文档200中被提及(例如,单个出现)8次,所以“清醒梦境”的提及值是8(n=8)。“清醒梦境”的8个单次出现中的每一个都将具有表示它在文档200中的位置的提及偏置。“清醒梦境”的首次出现将具有提及偏置,其表示比“清醒梦境”的第7次出现的提及偏置在文档200中的位置更早位置。因此,“清醒梦境”的第7次出现将具有这样的提及偏置,其表示了在文档200中比“清醒梦境”的第8次(n^{th})出现的提及偏置的位置更早的位置。

[0038] 图5为根据这里的实施方式,由语义分析器150执行对提取文本的统计处理的处理步骤510-565的流程图500。标志和短语的统计处理的目的是产生分数以对每个语义短语进行排名,从而帮助评估语义短语返回诸如高度关联性广告的内容的合适性。流程图500描述每个文档类型的最佳排名算法。流程图500中的步骤涉及在图2的框图中示出的统计处理器150-4。在步骤510,语义分析器150计算至少一个短语统计量。可以理解,步骤515-565的下列讨论描述一个或多个可以分别和共同(取决于文档200的特定类型)计算的短语统计量。

[0039] 在步骤515,语义分析器150计算短语的标志频率($tf(x_j)$),作为短语中标志的标

志值的函数，该函数包括短语中标志的标志出现值的平均值和中间值中的至少一个。例如，诸如“梦状态”的短语的标志频率为它的所有标志的标志出现值的函数。因此，“梦状态”的标志频率为“梦”(15)和“状态”(2)的标志值的函数。即使可以使用任意的函数，用于计算短语x_j的平均标志频率$tf(x_j)$的函数是各个标志出现值的平均值或中间值。

[0040] 在步骤520，语义分析器150计算短语的平均偏置(moffset(x_j))。短语的平均偏置由下列等式确定：

[0041] $$moffset(x_j)=\frac{1}{n}\sum_{i=1}^{n}offset(m_i)$$

[0042] 可以理解，$offset(m_i)$表示短语(x_j)在文档200内的一组n次出现中的单个出现(即，提及)的提及偏置。因此，总共8次出现(n＝8)的“清醒梦境”的第7次出现的提及偏置由$offset(m_7)$表示。可以理解，n可以等于或者大于1(n＝1)。

[0043] 利用中间偏置，在步骤525，语义分析器150计算短语的偏置标准差(soffset(x_j))。偏置标准差由下面的等式确定：

[0044] $$soffset(x_j)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(offset(m_i)-moffset(x_j))^2}$$

[0045] 在步骤530，当文档200是讨论多个中心话题的长文章时，语义分析器150计算文章分值(ascore(x_j))。文章分值由下面的等式确定：

[0046] $$ascore(x_j)=\frac{n(soffset(x_j))\sqrt{tf(x_j)}}{\sqrt{moffset(x_j)}}$$

[0047] 也就是，对于文章分值，分子是短语的“宏频”(由短语在文档200中的出现次数(n)测定)、这些短语的出现的“分布”(由这些出现的位置的标准差测定)、作为短语的标志频率的平方根的短语“微频”(由文档中200中标志的频率的中间值测定)的乘积。由于假定主题提及均匀分布，所以这被中间偏置的平方根约分(例如，除)。

[0048] 在步骤530的ascore(x_j)测量中，短语频率作出积极贡献，而且其标志还贡献由于其平方根而具有较小影响的平均频率。而且，短语在文档中的分布程度(或在开始或最后频繁提及)还对整体分值做出积极贡献。

[0049] 在步骤535，当文档是科技文章时，语义分析器150计算科技文章分值(tscore(x_j))。相反，如果短语的提及在文档200中平均分布，步骤530的先前测量ascore(x_j)使用短语在文档中位置的标准差。接着，ascore(x_j)的值最大。然而，对于步骤535的科技文章，文档开始和结尾的短语可能只有一半，接着标准差也最大。因此，只有当提及(例如，单个发生)在文档200中平均分布时，科技分值ascore(x_j)才最大。通过获取短语的相邻提及间差异的标准差，而利用了短语的提及的分布平均性。科技分值由下列等式确定：

[0050] $$tscore(x_j)=\frac{n}{rsdiff(x_j)}$$

[0051] 在步骤540，语义分析器150计算短语的两个单个出现的至少一个差(r)。具体地，r定义为下列等式中的一个：$r_1=offset(m_1)-0$、$r_i=offset(m_i)-offset(m_{i-1})$和$r_{n+1}=doclen-offset(m_n)$。可以理解，doclen表示文档200中的文本和字符的总数。

[0052] 在步骤545，语义分析器150计算中间差距(r(x_j))。平均差距由下面等式确定：

[0053] $$r(x_j)=\frac{1}{n}\sum_{i=1}^{n}r_i$$

[0054] 在步骤550，语义分析器150计算差距标准差(rsdiff(x_j)。差距标准差由下面等式确定：

[0055] $$rsdiff(x_j)=\frac{1}{n}\sqrt{\sum_{i=1}^{n}(r_i-r(x_j))^2}$$

[0056] 在步骤555，当文档200是信件(例如，信件、电子邮件、通告)时，语义分析器150计算标准差信件分值和微频信件分值。标准差信件分值假定：短语提及在文档200中的分布(由其提及偏置的标准差测定)与出现次数(n)等同关联性。一般地，在文档200中均匀提及的短语具有最大分值，在开始和结尾处同等(equally)提及的短语也是这样。经常提及但是在分离部分的短语具有较低的权重。对于大部分电子邮件、通告和信件，这非常合适。微频信件分值包括对“微频”的测量，典型地，其对于使用更多先进或科技语言的文档产生更好的结果。

[0057] 在步骤560，语义分析器150定义标准差信件分值(sdlscore(x_j))。标准差信件分值由下列等式确定：sdlscore(x_j)＝n(soffset(x_j))。在步骤565，语义分析器150定义微频信件分值(mflscore(x_j))。微频信件分值由下列等式确定：$mflscore(x_j)=n\sqrt{tf(x_j)}(soffset(x_j))$。

[0058] 可以理解，对于更快的计算机，这些公式可以封装到一组关系数据库视图定义中，这些定义使用文本标志化的高效实施来计算标志频率。接着这些视图定义可以用于呈现和格式化计算结果。

[0059] 图6为根据这里的实施方式，由语义分析器150执行提供语义短语的已排名列表以及内容预览的处理步骤610-635的流程图600。流程图600中的某些步骤涉及在图2的框图中示出的报告组件150-5。另外，用于多个作为语义分析结果而提供已排名语义短语的步骤330在步骤610-620中详细描述。

[0060] 在步骤610，语义分析器150提供多个已排名语义短语的列表，该列表根据一个或多个短语统计量对语义短语进行排名。例如，用户接口(例如，G.U.I，网页浏览器)可以显示和概述出列出的统计排名算法的结果。在用户接口中，可以显示语义短语(基于提取的文本)及其各自语义统计量。可以理解，已排名语义短语210的列表是可排序的，而且能够被用户操纵以调整排名结果。

[0061] 在步骤615，语义分析器150从列表中选择一个或多个语义短语，以创建由选中的一个或者多个语义短语中每一个所定义的内容的预览(例如，弹出窗口、G.U.I、网页浏览器)，该预览显示由选中的一个或者多个语义短语中每一个所确定的内容的聚合。具体地，在用户接口中列出的已排名语义短语被使能，从而能够呈现定义的内容(例如，实际广告)以确定选中的语义短语的有效性。

[0062] 例如，用户可以从已排名语义短语210的列表中选择两个短语。响应于该选择，语义分析器150可以创建内容预览220，并且显示由选中的语义短语确定的内容(例如，广告)。在一个实施方式中，如果从已排名语义短语210的列表中选择的短语是“化学工程”和“学院”，则内容预览220接着显示与“化学工程”和“学院”相关的广告，从而用户可以判定语义短语“化学工程”和“学院”是否提供了保证指定“化学工程”和“学院”作为文档200的语义关键字的期望广告。如果用户确定选择的语义短语的内容预览220是可接受的，则语义分析器

150向用户提供为文档200创建关键字的功能性。具体地，在步骤620，语义分析器150从列表中指定一个或多个语义短语，以作为语义关键字与文档200相关联。

[0063] 用于将至少一个语义短语与文档200相关联的步骤340的细节包括，在步骤625，语义分析器150将该至少一个语义短语插入到文档的元数据部分。在步骤630，语义分析器150将该至少一个语义短语插入到文档的可扩展元数据平台(XMP)部分。可以理解，元数据平台或称XMP，是使用在PDF、摄影和照片编辑应用中的可扩展标记语言。XMP定义可以与任意已定义的元数据项目的集合一起使用的元数据模型。XMP还为基本属性定义了特定方案，用于记录资源(例如，文档)经过多个处理步骤的历史，例如，从被摄影、扫描，或创作为文本；或者经过照片编辑步骤(诸如，剪裁或调色)，整装成最终图像。XMP可以允许每个软件程序或设备沿着该途径，增加自身信息到数字资源(例如文档)，其随后可以在最终数字文件中予以保留。

[0064] 在步骤635，语义分析器150为文档200指定策略。该策略使该文档200能够利用与该文档200相关联的至少一个语义短语作为语义关键字，以当文档200被呈现时显示该内容。在可选方式中，该策略使文档200能够忽略与该文档相关联的该至少一个语义短语，以当文档200被呈现时不显示该内容。

[0065] 再次注意，这里的技术良好适用于配置用以自动分析文档和提供反映文档主题及话题的语义短语的已排名列表(基于提取的文本)的语义分析器。然而，应当注意，这里的实施方式并不局限于应用在这些应用中，这里讨论的技术还良好适用于其它应用。

[0066] 虽然参照优选实施方式特别示出和描述了本发明，本领域技术人员可以理解，可以不脱离由附录权利要求所限定的本发明的精神和范围，作出各种形式上或详细的改变。期望本发明的范围覆盖这些变化。因此，本申请实施方式的前面描述并不意欲为限制性的。相反，在下面权利要求中呈现对本发明的任意限制。

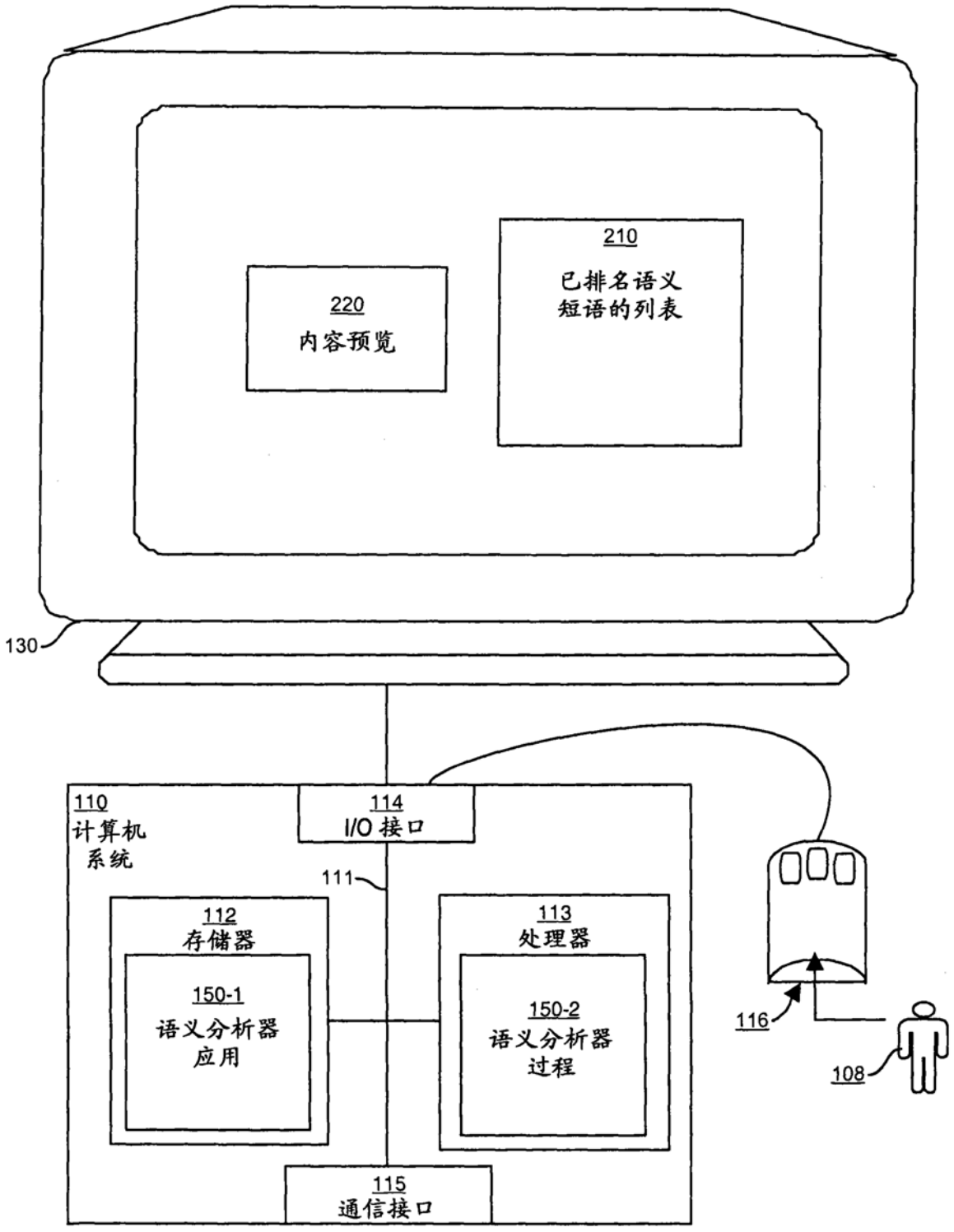

210
已排名语义
短语的列表
220
内容预览
130
110
计算机
系统
114
I/O 接口
111
112
存储器
150-1
语义分析器
应用
113
处理器
150-2
语义分析器
过程
115
通信接口
116
108

图1

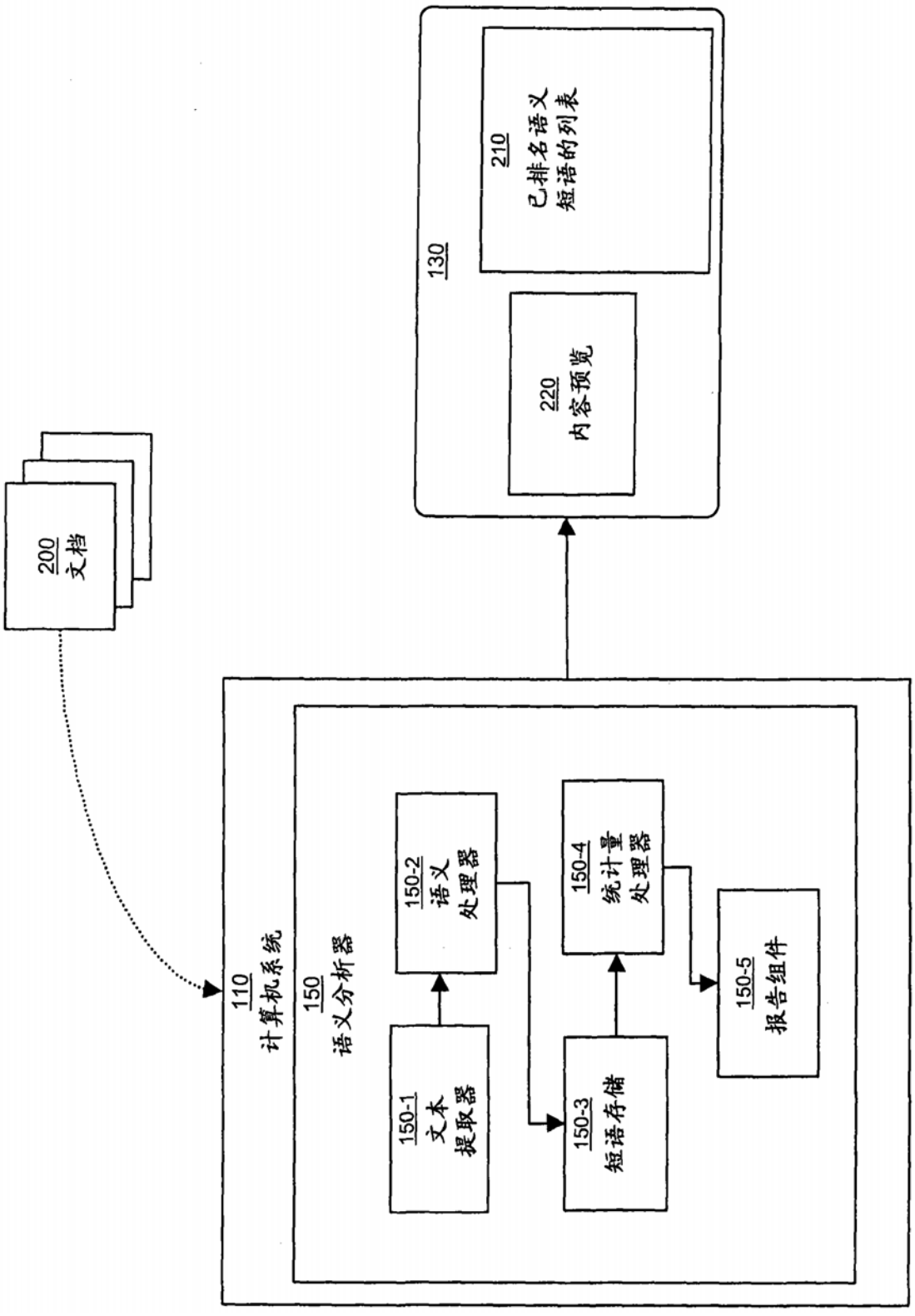

200
文档
130
210
已排名语义
短语的列表
220
内容预览
110
计算机系统
150
语义分析器
150-1
文本
提取器
150-2
语义
处理器
150-3
短语存储
150-4
统计量
处理器
150-5
报告组件

图2

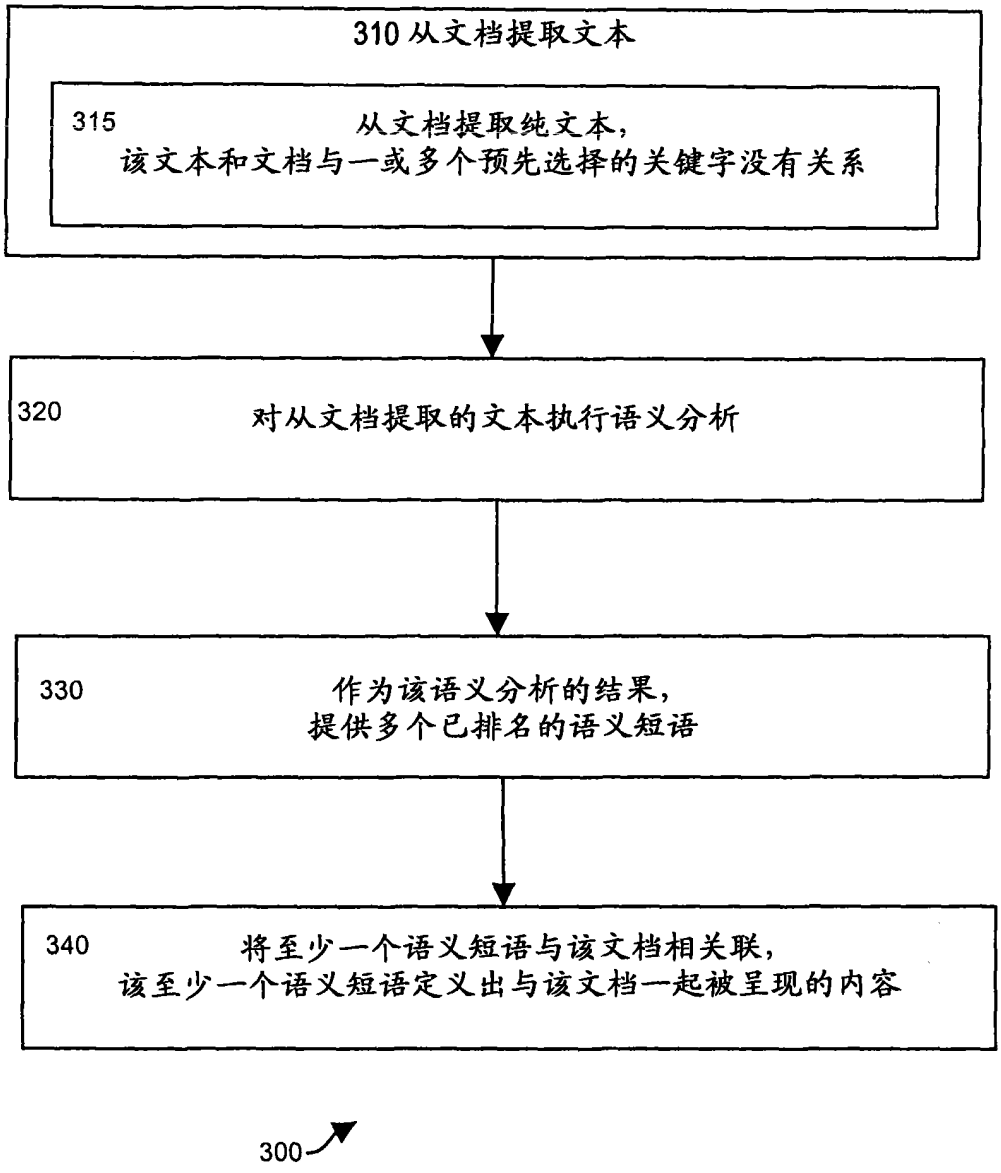

310 从文档提取文本
315
从文档提取纯文本，
该文本和文档与一或多个预先选择的关键字没有关系
320
对从文档提取的文本执行语义分析
330
作为该语义分析的结果，
提供多个已排名的语义短语
340
将至少一个语义短语与该文档相关联，
该至少一个语义短语定义出与该文档一起被呈现的内容
300

图3

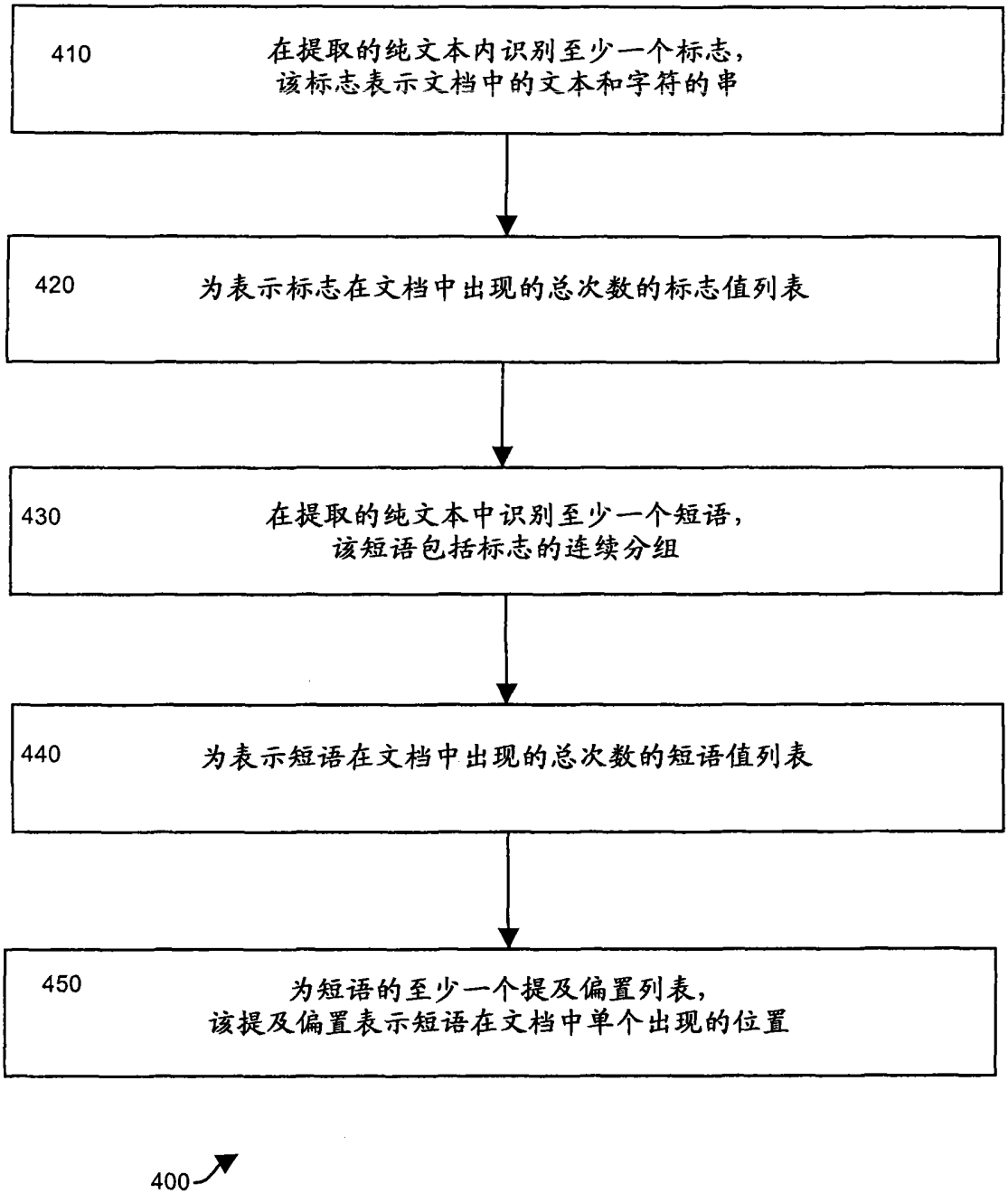

410
在提取的纯文本内识别至少一个标志，
该标志表示文档中的文本和字符的串
420
为表示标志在文档中出现的总次数的标志值列表
430
在提取的纯文本中识别至少一个短语，
该短语包括标志的连续分组
440
为表示短语在文档中出现的总次数的短语值列表
450
为短语的至少一个提及偏置列表，
该提及偏置表示短语在文档中单个出现的位置
400

图4

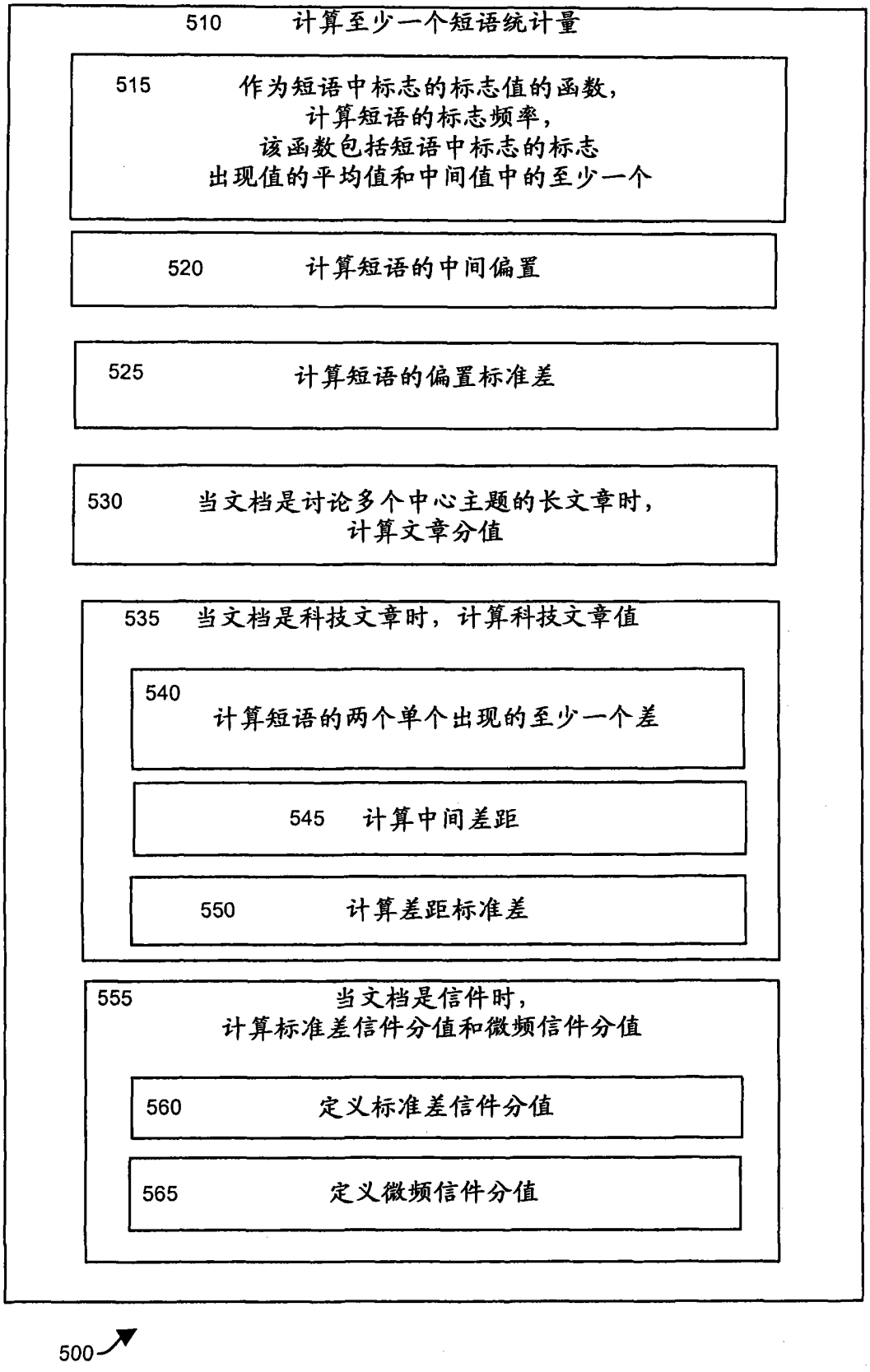

510 计算至少一个短语统计量
515 作为短语中标志的标志值的函数，计算短语的标志频率，该函数包括短语中标志的标志出现值的平均值和中间值中的至少一个
520 计算短语的中间偏置
525 计算短语的偏置标准差
530 当文档是讨论多个中心主题的长文章时，计算文章分值
535 当文档是科技文章时，计算科技文章值
540 计算短语的两个单个出现的至少一个差
545 计算中间差距
550 计算差距标准差
555 当文档是信件时，计算标准差信件分值和微频信件分值
560 定义标准差信件分值
565 定义微频信件分值
500

图5

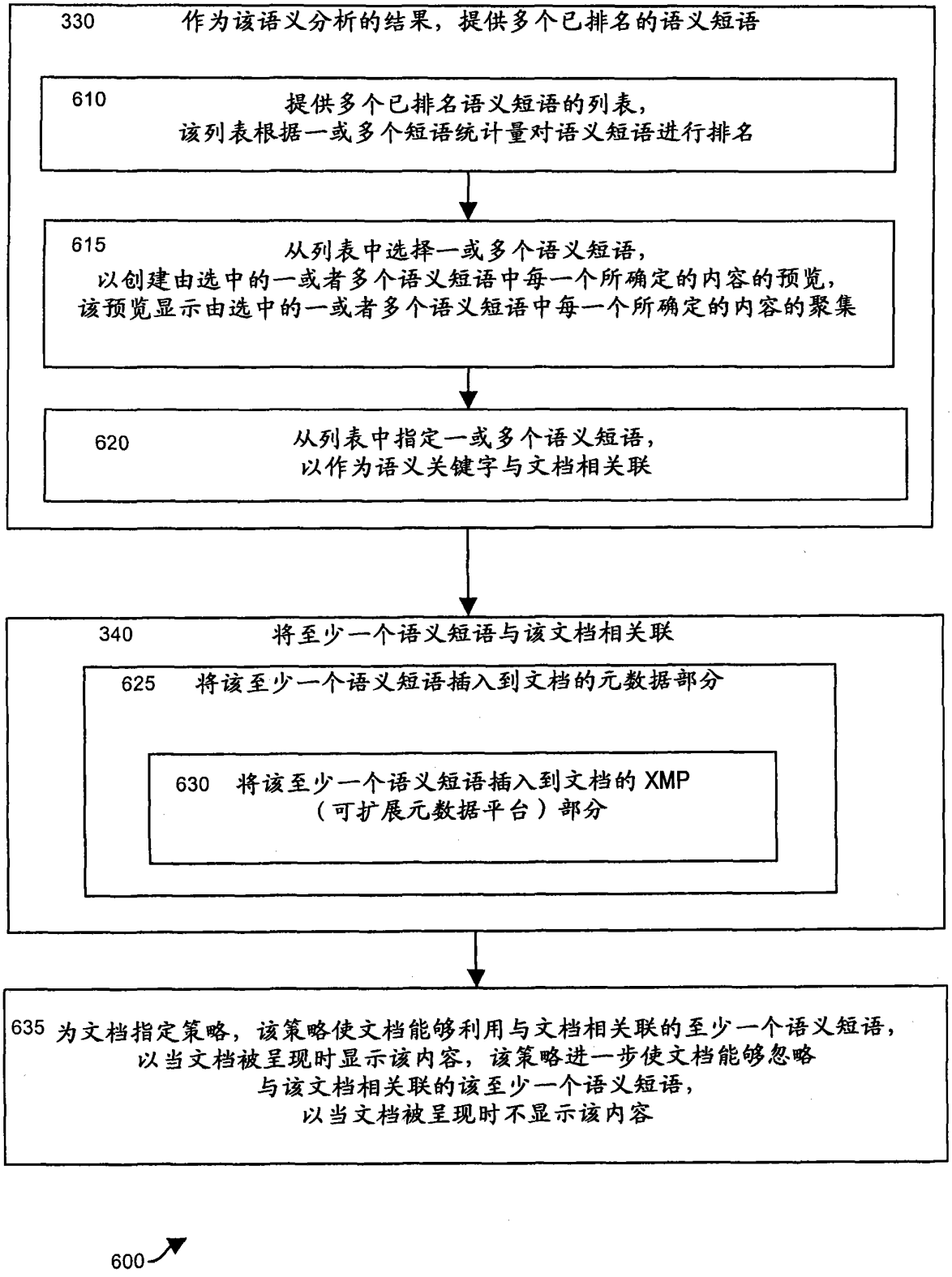

330 作为该语义分析的结果，提供多个已排名的语义短语
610 提供多个已排名语义短语的列表，该列表根据一或多个短语统计量对语义短语进行排名
615 从列表中选择一或多个语义短语，以创建由选中的一或者多个语义短语中每一个所确定的内容的预览，该预览显示由选中的一或者多个语义短语中每一个所确定的内容的聚集
620 从列表中指定一或多个语义短语，以作为语义关键字与文档相关联
340 将至少一个语义短语与该文档相关联
625 将该至少一个语义短语插入到文档的元数据部分
630 将该至少一个语义短语插入到文档的 XMP（可扩展元数据平台）部分
635 为文档指定策略，该策略使文档能够利用与文档相关联的至少一个语义短语，以当文档被呈现时显示该内容，该策略进一步使文档能够忽略与该文档相关联的该至少一个语义短语，以当文档被呈现时不显示该内容
600

图6