(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 1691007 B
(45)授权公告日 2010.06.16

(21)申请号 200510067431.0

(22)申请日 2005.04.22

(30)优先权数据
127122/04 2004.04.22 JP

(73)专利权人 惠普开发有限公司
地址 美国德克萨斯州

(72)发明人 小田弘美

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 程天正 刘杰

(51)Int.Cl.
G06F 17/28(2006.01)
G06F 17/30(2006.01)

(56)对比文件
US 2003/0120639 A1,2003.06.26, 全文.
US 2003/0167267 A1,2003.09.04, 全文.

审查员 吴紫璇

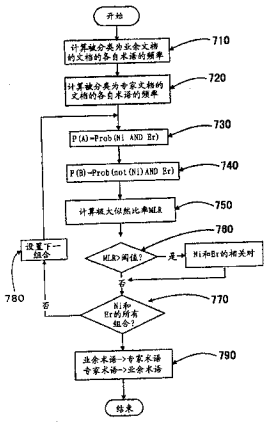
权利要求书 1 页 说明书 12 页 附图 9 页

(54)发明名称

用于文档处理的方法和系统

(57)摘要

本发明涉及用于文档处理的方法和系统,以检测术语。从具有共同主题的第一文档集和第二文档集中检测第二集中对应第一集中专用术语的术语,和/或第一集中对应第二集中专用术语的术语的方法,包括:根据在第一术语列表中列出的每个术语的频率从第一集创建第一术语矩阵;根据在第二术语列表中列出的每个术语的频率从第二集创建第二术语矩阵;根据第一术语矩阵和第二术语矩阵的积计算词汇映射矩阵;依照元素值的降序在词汇映射矩阵的特定行中选择预定数目的术语以把选择的术语采用为在第一集中对应第二集中专用术语的术语;并且依照元素的降序在词汇映射矩阵的特定列中选择预定数目的术语以把选择的术语采用为在第二集中对应第一集中专用术语的术语。



CN 1691007 B

1. 一种从具有共同主题的专家文档集和业余文档集中检测 (a) 在业余文档集中、对应于专家文档集中的专用术语的术语，和 / 或 (b) 在所述专家文档集中、对应于所述业余文档集中的专用术语的术语的方法，已经根据术语列表检索了所述专家和业余文档集，包括 ：

根据在专家术语列表中列出的每个术语的频率，来从所述专家文档集创建专家术语矩阵 ；

根据在业余术语列表中列出的每个术语的频率，来从所述业余文档集创建业余术语矩阵 ；

根据所述专家术语矩阵和所述业余术语矩阵的积来计算词汇映射矩阵 ；

依照元素值的降序来在所述词汇映射矩阵的特定行中选择数目预先确定的术语，以便把所选择的术语采用为在所述专家文档集中、对应于所述业余文档集中的专用术语的术语 ；并且

依照元素的降序来在所述词汇映射矩阵的特定列中选择数目预先确定的术语，以便把所选择的术语采用为在所述业余文档集中、对应于所述专家文档集中的专用术语的术语，

其中 ：(a) 在所述术语列表中术语的数目是 s，(b) 从所述专家文档集中选择的术语数目是 n，(c) 由 s 乘 n 矩阵 P 来表示所述专家术语矩阵，(d) 在所述专家文档集的第 k 个文档中第 i 个术语的频率是 Exp(k，i)，(e) 所述第 i 个术语的总体频率是 Etf(i)，并且 (f) 在第 k 个文档中术语的总数目是 Ewf(k)，矩阵 P 的元素是 ：

[方程式 1]

$$We(k,i)=\frac{Exp(k,i)}{(\mathrm{Etf(i)}*\mathrm{Ewf(k)})}$$

(g) 从所述业余文档集中选择的术语数目是 m，(h) 由 s 乘 m 矩阵 Q 来表示所述业余术语矩阵，并且 (i) 出现在所述业余文档集的第 k 个文档中第 r 个术语的频率是 Naive(k，r)，(j) 所述第 r 个术语的总体频率是 Ntf(r)，并且在所述第 k 个文档中术语的总数目是 Nwf(k)，如下给出矩阵 Q 的元素

[方程式 2]

$$Wn(k,i)=\frac{\mathrm{Naive(k,r)}}{(\mathrm{Ntf(r)}*\mathrm{Nwf(k)})}$$

所述词汇映射矩阵定义为 $T = Q^tP$，其中 t 表示矩阵的转置，如下定义了所述词汇映射矩阵的每个权重值 ：

[方程式 6]

$$W(r,i)=\sum_{k=1}^{s}\left[\frac{Exp(k,i)}{(\mathrm{Etf(i)}*\mathrm{Ewf(k)})}\frac{\mathrm{Naive(k,r)}}{(\mathrm{Ntf(r)}*\mathrm{Nwf(k)})}\right]$$

2. 一种用于执行如权利要求 1 所述的方法的文档处理系统。

用于文档处理的方法和系统

技术领域

[0001] 本发明涉及处理多个具有共同主题的文档集。

背景技术

[0002] 具有采用相同语言的多个描述并且共享相同内容的文档，在那些描述中频繁地使用术语，所述术语的不同取决于作者关于主题所具有的专业知识程度，以及所述作者属于的不同社会层，诸如性别或年龄组。即使所述描述是关于共同主题的，那么由非专家和由专家在他们各自的表达领域中使用的术语也可能是相当不同的。

发明内容

[0003] 本发明的目的是提供一种新的并且改进的方法、设备及其它必要的技术，用于检测由非专家使用且与由专家使用的术语所表达的意思相对应的术语，并且反之用于检测在这种不同的领域之间由专家使用且与由非专家使用的术语所表达的意思相对应的术语。

[0004] 用于转换不同领域的文档的技术的典型例子是翻译机。使计算机执行翻译机的任务的技术已经是已知的。翻译机利用使用术语数据库的计算机程序、用于处理语法规则的程序、用法和例句数据库及其它系统特定组件，来把用自然语言写入的文档自动翻译为另一种自然语言。已经实际应用这种技术，并且存在用于个人计算机的商用语言翻译软件产品。在因特网上也提供某些翻译服务。另外，用于逐词翻译的小型手持装置到处都可以买到。逐词翻译机把用某种语言的一个词转换为用另一种语言、具有同样意思的词。基本上，把预编译词典存储在存储设备中，并且把输入词转换为用另一种语言的对应词。这些常规的技术具有用于把文档从一个领域转换为另一领域的前提；即，在一个领域中的句子必须已知对应于另一个领域中的句子，并且在一个领域中的词必须已知对应于另一个领域中的词。

[0005] 用于把困难的表达转换为用同样语言的容易的表达的意译研究已经问世。例如，在由 Atsushi Fujita 等人 (2003) 和 Masahiro Murayama 等人 (2003) 的研究中报告。在涉及“意译”的研究中，基本技术是寻找将要依照模式匹配规则来由预先确定的表达模式代替的表达模式。在语言翻译中的其它方法利用统计和 / 或概率模型。这些基于模型的方法最初准备一对数据集，其用不同的语言并且具有已知是相同的内容。接下来，根据诸如在每个数据集中句子长度之类的信息，确定用语言 A 和语言 B 的对应句子。最后，根据它们在所述数据集中共同出现的关系来确定在词之间的对应关系。在这种和其它现有技术情况中，存在这样一个前提，即对应于语言 A 的词 Wa，在语言 B 中存在具有合理的语义准确性的词 Wb。

[0006] 专利文档 1 是“Daily Language Computing and its Method”JP 2002-236681 A。

[0007] 专利文档 2 是“Association Method for Words in PaginalTranslation Sentences”JP 2002-328920 A。

[0008] 非专利文档 1 是 http://www2.crl.go.jp/it/a133/kuma/mrs_1i/midisearch.

htm。

[0009] 非专利文档2是Atsushi Fujita,Kentaro Inui,YujiMatsumoto。“Text Correction Processing necessary forParaphrasing into Plain Expressions”。日本第65届信息处理学会全国大会演讲论文集,第五分册,1T6-4,第99-102页,2003年3月。

[0010] 非专利文档3是Masahiro Murayama,Masahiro Asaoka,MasanoriTsuchiya,Satoshi Sato。“Normalization of Terms and Supportfor Paraphrasing Declinable words based on theNormalization”,语言处理学会,第9届年度大会,第85-88页,(2003年3月)。

[0011] 非专利文档4是Dunning,T。(1993).Accurate methods forthe statistics of surprise and coincidence。计算语言学,19(1):61-74

[0012] 如上所述,在常规的机器翻译中,假定在从一种语言翻译到另一种语言时,在两种语言中存在相应的词,而且相应文档集可用。

发明内容

[0013] 本发明的目的是提供一种新的和改进的方法和设备,用于检测用于一个领域的术语,所述术语近似对应于另一个领域中的术语,和/或反之亦然,即使在下列情况下:(1)在目标领域中没有彼此对应的已知词对,(2)没有事先已知彼此对应的文档集对,和/或(3)没有帮助在上述领域中映射的词典或辞典。

[0014] 依照本发明一个方面,为了解决上述问题,

[0015] (1)检索用两种不同的语言表达写入的文档集,其被记述为关于同样的主题(这些文档以下被称为在领域A中的文档和在领域B中的文档),并且

[0016] (2)当给出两种不同的语言表达的这种文档集时,在出现于领域A中文档的术语和出现于领域B中文档的术语之间建立关联。

[0017] 为此,用检索工具使用预先确定的关键词列表来收集候选文档,以便准备用两种不同的语言表达写入的文档集。然而,由于用检索工具检索的候选文档包括大量的所谓的“噪声(noise)”文档,所以在多数情况下,并不能像正常那样使用所述检索结果。从而,本发明的一方面包括从所收集的文档中删除所述“噪声”文档的初始步骤。在此初始步骤之后,根据在所述文档中的术语频率及其它信息来把所述文档分类为专家(expert)文档和业余(naive)文档,其包含不同类型的语言表达。由于出现在目标专家文档和目标业余文档中的术语并不总是相同的,接下来计算在所述两个不同领域中术语之间的相关性。基本概念如下:根据在专家文档集和业余文档集中的术语之间的共同出现关系,来获得出现在专家或业余领域中的一个或一组术语与出现在另一个领域中的一个或一组术语的关联,所述术语记录相同对象。

[0018] 本发明应用的一个例子是适用于打算要购买某些产品或货物的用户的推荐系统。即使文档记述诸如商品之类的相同对象,通常在由具有关于所述对象的高深知识的专家所使用的术语和由具有关于所述对象的很少知识的非专家所使用的术语之间,存在相当多的差异。所述专家常常使用技术术语和特定知识来描述所述对象,而没有这种知识的非专家不得不用基于感知的表达或经由相似的对象或例子来描述所述对象。所述专家试图用他/她的知识来详细地解释所述产品,关于它在哪制造和/或它由什么材料组成,而非专家试

图使用回忆起来的、基于感知的术语来描述相同的产品。普通消费者在所有的专业范围内具有详细的产品知识和涉及产品的专有名称几乎是不可能的。从而，即使专家向非专家解释并推荐特定产品，这事实上要求专业知识来精明地选择，可以设想非专家在购买之前可能不会充分理解所述解释。

[0019] 通过应用本发明，卖方能够用消费者理解的词汇来向所述消费者提供关于产品的充足信息，并且反之，普通消费者可以容易地理解关于产品的信息并且选择适合于他 / 她偏好和品味的信息。

附图说明

[0020] 图 1 是用于执行本发明优选实施例的整个系统图。

[0021] 图 2 是包括在图 1 的系统内的设备图。

[0022] 图 3 是由图 1 的系统执行的算法的流程图。

[0023] 图 4 是由图 2 的设备使用的、用于从图 1 的系统所检索的文档中删除“噪声”文档的方法的流程图。

[0024] 图 5 是由图 2 的设备使用的、用于计算文档的等级相关系数和有效值的方法的流程图。

[0025] 图 6 是由图 2 的设备使用的、用于把文档分类为专家文档和业余文档的方法的流程图。

[0026] 图 7 是由图 2 的设备使用的、用于使用 MLR 方法来执行词汇映射的方法的流程图。

[0027] 图 8a 是专家术语矩阵图。

[0028] 图 8b 是业余术语矩阵图。

[0029] 图 8c 是词汇映射矩阵图。

[0030] 图 9 是由图 2 的设备使用的、用于计算图 8c 的词汇映射矩阵的算法。

具体实施方式

[0031] 图 1 是包括连接到网络 140 的用户 PC 110、站点服务器 (1)120 和站点服务器 (2)130 的系统图。用户访问所述站点服务器 (1)120 和所述站点服务器 (2)130，以便通过使用某种检索工具经由 PC 110 的操作来获得必要的信息。如图 1 的实施例描述了在因特网上的检索。然而，可以使用任何可以检索必要信息的检索系统。所述用户可以通过用在用户 PC 110 上的计算机程序处理所获得的信息，来获得所想要的结果。

[0032] 图 2 是包括外壳 200 的用户 PC 图，所述用户 PC 具有存储设备 210、主存储器 220、输出装置 230、中央处理器 (CPU)240、控制台 250 和网络 I/O 260。所述用户操作所述控制台 250 以便经由所述网络 I/O260 从在因特网上的每个站点获得必要的信息。中央处理器 240 根据存储在所述存储设备 210 中的文档信息来对从因特网检索的信息执行预先确定的数据处理，并且在所述输出装置 230 上显示结果。

[0033] 图 3 是由图 1 的系统和图 2 的 PC 执行的、用于检测在业余和专家文档之间对应术语的操作 (即步骤) 的流程图。所述步骤是 :

[0034] 步骤 310 :使用指定的术语来获得候选文档。

[0035] 步骤 320 :预处理所述候选文档。

[0036] 步骤 330 :删除“噪声”文档。

[0037] 步骤 340 :计算每个文档的特征值。

[0038] 步骤 350 :用判别分析来分类所述文档。

[0039] 步骤 360 :检测在业余和专家文档之间的对应术语。

[0040] 下面详细地描述每个步骤。

[0041] (1) 使用指定的术语来获得候选文档

[0042] 检测对应术语的第一步骤(步骤 310)是准备用于描述相同内容的数据集,所述数据集包括业余文档(由非专家写入的文档,以下称为 N 文档)和专家文档(由专家写入的文档,以下称为 E 文档)对。通过使用术语列表来准备所述数据集。

[0043] 所述术语列表是可以被用作在给定领域中的关键词的术语的列表。例如,当选择“酒”领域时,所述术语列表包括“(产品)酒的名字”。用户使用检索工具依照在所述术语列表中描述的酒名字来在因特网上收集关于酒的信息。指定酒名字,诸如“Auslese-ChateauCure Bon-Chateau Margaux-Vin Santo Toscano”。用那些术语作为关键词来从数据库中检索候选文档。可以使用任何存储这种信息的数据库。现在描述使用搜索引擎来在因特网上检索候选文档的方法。

[0044] 所述用户用酒名字执行检索,所述酒名字被定义为在上述术语列表中的搜索关键词。通过使用搜索引擎来检索酒名字,所述搜索引擎作为商业产品或自由软件可买到。通常,当把酒名字指定为搜索关键词时,检索大量的候选文档。然而,可以依照某些等级排列来选择数目预先确定的候选项。通过使用术语列表对于所有想要的术语可以自动地获得候选文档。

[0045] (2) 预处理候选文档(步骤 320)。

[0046] 用这种方式在因特网上从网页中自动地获得的文档包括各种信息,并且在多数情况下不能像正常那样使用。把对应于垃圾型文档、列表型文档和日记型文档的文档作为“噪声”文档从自动获得的文档中删除。在删除所述“噪声”文档之前,对从所述网页中提取的文档应用预处理。在预处理的第一阶段中,从网页信息中提取可以被认为是文档的部分以便执行文档分析。接下来,把所述文档分段为词来提取实义词、虚词、助词等等,以便能够计算这些文档的特征值 ;即,实义词的数目值,业余词的比例,专有名词的比例,附加专有名词的比例和虚词 / 助词的比例。下面描述为了计算那些特征值而用于该说明书的概念术语。

[0047] (i) 实义词 (content word) 的数目

[0048] 其是包括在网页中文档内的实义词的数目。实义词包括名词、动词、形容词和副词,除了虚词 (particle word)/ 助词 (auxiliaryword)。

[0049] (ii) 业余词的比例=业余词的数目 / 实义词的数目

[0050] 业余词是由在相关领域中的非专家使用的预先确定的词。业余词的比例是出现在一个网页中预先确定的业余词(以下称为“主业余词”)的数目与实义词的数目的比例。

[0051] (iii) 专有名词的比例=专有名词的数目 / 实义词的数目

[0052] 在这里专有名词是通常被称为专有名词的名词。专有名词的比例是出现在一个网页中的专有名词的数目与实义词的数目的比例。

[0053] (iv) 附加专有名词的比例=附加专有名词的数目 / 实义词的数目

[0054] 附加专有名词是通常不被认为是专有名词,但是需要被增加为专有名词以便检测

所对应术语的名词。专有名词的比例是在一个网页中出现的附加专有名词的数目与实义词的数目的比例。

[0055] (v) 虚 / 助词的比例＝虚词的数目 / 助词的数目 / 实义词的数目

[0056] 通过计算在一个网页中出现的虚词的数目与助词的数目的比例，并且通过用实义词的数目除所述比例以便规一化所述比例，来计算虚词 / 助词的比例。

[0057] (vi) 实义词的 n 元语法 (n-gram)

[0058] 通过使用实义词的单语法、实义词的双语法、实义词的三语法和实义词的跳跃双语法来检查在文档之间的相关性。

[0059] 实义词的单语法用于根据一个词（或术语）的频率来确定在文档之间的相关性。在酒领域情况下，可以使用诸如“酒”、“香味”和“饮料”之类的词的频率。

[0060] 实义词的双语法用于根据两个连续词的频率来确定在文档之间的相关性。在酒领域的情况下，使用诸如“酒精 - 百分比”、“这种 - 酒”和“生产国家 - 年代”之类的两个连续词的频率。

[0061] 实义词的三语法用于根据三个连续词的频率来确定在文档之间的相关性。在酒领域的情况下，使用诸如“酒 - 饭 - 喝酒方式”、“白色 - 法国 -1990”和“红色 - 德国 - 优质干白葡萄酒”之类的三个连续词的频率。

[0062] 实义词的跳跃双语法使用在三个连续词中的第一和最后词以便根据这些词的频率来确定在文档之间的相关性。举例来说，可以把“高质量”和“生产”指定为双语法模式的第一词和最后词。由于作为结果的模式要求“高质量 -XXX- 生产”，所以诸如“高质量 - 水果串 - 生产”或“高质量 - 雷司令白葡萄酒 - 生产”满足所述条件。XXX 表明任意的词。

[0063] (vii) 虚词 / 助词的 n 元语法

[0064] 类似地，使用虚词 / 助词的单语法、双语法、三语法、虚词 / 助词双语法、虚词 / 助词的三语法和虚词 / 助词的跳跃双语法。

[0065] 虚词 / 助词的单语法的例子包括“no”、“ga”和“ni”。虚词 / 助词双语法的例子包括“no-ga”、“no-no”、“no-ni”。虚词 / 助词三语法三语法的例子包括“no-ga-ga”、“no-no-ga”和“no-ni-ga”。

[0066] 虚词 / 助词的跳跃双语法的例子包括“no-X-ga”、“no-X-ga”和“no-X-ga”。注意，“X”是任意的虚词或助词。

[0067] (viii) 等级相关系数及其有效值

[0068] 在该实施例中，使用 Spearman 公式来计算等级相关系数和有效值。作为例子将用实义词的单语法来解释。首先，确定用于主业余文档的、诸如“sake（液体）”、“kaori（香味）”、“nomu（饮料）”、“aji（味道）”、“kanjiru（感觉）”和“owom（考虑）”之类的词的频率。类似地，确定用于从某个网络站点获得的文档的、诸如“sake（液体）”、“kaori（香味）”、“nomu（饮料）”、“aji（味道）”、“kanjiru（感觉）”和“owom（考虑）”之类的词的频率。接下来，对于各自的文档计算这些词的频率等级。根据这些个等级信息来计算 Spearman 的等级相关系数，并且计算所述相关系数的有效值。

[0069] (ix) 主业余文档集（或主专家文档集）

[0070] 主业余文档集是包括由非专家在某个领域中使用的术语的文档收集。主专家文档集是包括由专家在某个领域内使用的术语的文档收集。

[0071] (3) 删除“噪声”文档

[0072] 必须删除作为来自从因特网上的网页中检索的文档的“噪声”文档的垃圾型文档、列表型文档和日记型文档。通常认为在“噪声”文档中不包括为检测用于一个领域的术语所必须的信息，所述术语近似对应于用于另一领域中的术语。图 4 是由图 1 的系统执行的、用于删除“噪声”文档的步骤的流程图。

[0073] 410 :删除垃圾型文档。

[0074] 420 :删除列表型文档。

[0075] 430 :删除日记型文档。

[0076] 440 :确认对于所有文档已经执行了删除。

[0077] 450 :设置下一文档。

[0078] 以下描述删除垃圾型、列表型文档和日记型文档。

[0079] (A) 垃圾型文档

[0080] 把满足所有下列条件的文档定义为垃圾型文档。垃圾型文档字面上是垃圾并且不能用于从一个领域到另一领域的术语检测。下面定义用于选择垃圾型文档的准则。

[0081] (a) 其实义词的数目少。

[0082] (b) 其业余词的比例低。

[0083] (c) 其专有名词比例低。

[0084] (d) 其与“主业余文档”的相关系数低。

[0085] 所述“主业余文档集”是事先作为由非专家写入的文档而选择的一组文档。作为选择，可以把由专家事先选择作为文档的一组文档用作为“主专家文档集”。

[0086] (B) 列表型文档

[0087] 把满足所有下列条件的文档定义为列表型文档。这发生在下列情况，其中把关于在某个领域中的对象信息简单地存储为因特网上站点的列表。

[0088] (a) 其专有名词的比例高。

[0089] (b) 其基于实义词和虚词 / 助词与“主业余文档”的相关系数低。

[0090] (c) 日记型文档

[0091] 把满足所有下列条件的文档定义为日记型文档。日记型文档是这样一种文档，其中例如描述了关于液体和酒的信息，但是主要讨论了其它主题或信息。这种文档可能出现在个人日记或在线百货商店的因特网站点上，其涉及液体或酒并且包括许多其它主题。

[0092] (a) 其涉及某个领域的专有名词的比例低。

[0093] (b) 其基于实义词 n 元语法与主文档的相关性低。

[0094] (c) 其基于虚词 / 助词 n 元语法与主文档的相关性高。

[0095] 根据上述定义，因为把垃圾型文档、列表型文档和日记型文档都认为是“噪声”文档，所以在考虑术语领域检测过程中把它们删除。

[0096] (4) 用判别分析来分类所述文档

[0097] 在除去所述“噪声”文档之后，应用判别分析来把其余文档分类为业余文档或专家文档。为了执行所述判别分析，从各自输入文档中提取特征值。使用的特征值具有五种比例 ;即实义词的数目，业余词的比例，专有名词的比例，附加专有名词的比例和虚词 / 助词的比例。此外，使用根据实义词 n 元语法计算的 Spearman 相关系数及其有效值，和根据虚

词 / 助词 n 元语法计算的 Spearman 等级相关系数及其有效值。

[0098] 在下面描述了根据 Spearman 公式来计算等级相关系数及其有效值。图 5 是图 2 的计算机执行用于根据 Spearman 公式来计算等级相关系数及其有效值的操作的流程图。

[0099] 510 :在主业余文档 (Y) 中 n 元语法的频率。

[0100] 520 :在输入文档 (X) 中 N 元语法的频率。

[0101] 530 :依照 X 和 Y 来计算 Spearman 等级相关系数 (ri) 和有效值 (ei)。

[0102] 540 :对于所有 N 元语法确认计算。

[0103] 550 :设置下一 n 元语法。

[0104] 560 :获得所有 n 元语法的等级相关系数和有效值。

[0105] 以下详细描述等级相关系数 / 有效值。

[0106] 把实义词单语法用作为解释的例子。使用它们根据单个词的频率来计算在文档之间的相关性。在酒领域情况下,根据所选择的文档和主业余文档集 (或主专家文档集) 来计算诸如“酒”、“香味”和“饮料”之类的词的频率。把这些频率数字指定为 Y(y1,y2,y3,...,yh)(步骤 510)。

[0107] 接下来,根据输入文档来计算相似的特征值 ;并且把相似的特征值指定为 X(x1,x2,x3,...,xh)(步骤 520)。这里,h 表示数据或词类型的数目,对于所述数据或词类型计算频率。基于 Spearman 公式根据这些数据来计算等级相关系数和有效值。

[0108] r1 = F(X,Y)

[0109] e1 = G(X,Y),

[0110] 其中 r1 是依照 Spearman 相关系数公式而计算的等级相关系数,而 e1 是依照 Spearman 有效值公式计算的等级相关系数的有效值 (步骤 530)。采用相同的方式,对于实义词双语法计算 r2、e2,并且对于其它 n 元语法也进行类似地计算。同样,采用相同的方式对于虚词 / 助词 n 元语法计算等级相关系数和有效值 (步骤 540 和 550)。结果,计算 R = (r1,r2,...,rd) 和 E = (e1,e2,...,ed)(步骤 560)。这里,d 表示实义词 n 元语法和虚词 / 助词 n 元语法的总数目。

[0111] 在该实施例中,对于四种实义词的 n 元语法计算 Spearman 相关系数及其有效值 ;所述四种实义词的 n 元语法即,实义词的单语法、实义词的双语法、实义词的三语法和实义词的跳跃双语法。因此,计算八个特征值作为 Spearman 相关系数及其有效值。类似地,根据虚词 / 助词来计算八个特征值作为 Spearman 相关系数及其有效值。增加上述五个特征值,总共使用 21(= 5+8+8) 个特征值。

[0112] 接下来,使用 Mahalanobis 距离函数来区分输入文档以便把所述输入文档分类为业余文档或专家文档。图 6 是图 2 的计算机执行用于把输入文档分类为业余文档、专家文档及其它文档的操作的流程图。

[0113] 610 :计算主业余文档和主专家文档的特征值。

[0114] 620 :计算每个输入文档的特征值。

[0115] 630 :计算在所述输入文档和所述主业余文档之间的距离 (Db) 和在所述输入文档和所述主专家文档之间的距离 (Da)。

[0116] 640 :如果在所述输入文档和所述主业余文档之间的距离小于阈值,那么把所述输入文档分类为业余文档。

[0117] 650 :如果在所述输入文档和所述主专家文档之间的距离 (Da) 小于阈值,那么把所述输入文档分类为专家文档。

[0118] 660 :把不对应于主业余文档或主专家文档的文档分类为“其它”文档。

[0119] 670 :确认所有文档被分类

[0120] 680 :设置下一文档

[0121] 在下面详细描述各自的步骤。首先,计算所述主业余文档和所述主专家文档的特征值。当使用判别式来判别文档时,这些构成了各自集的基本总数。所述主业余文档是具有这样显著特征的一组文档,即主业余文档选自“主业余文档集”。计算构成主业余文档的各自文档的特征值,并且计算所述特征值的平均值。所述主专家文档也选自所述“主专家文档集”,并且计算各自文档的特征值,并且采用相同的方式来计算所述特征值的平均值 (步骤 610)。

[0122] 接下来,计算所述输入文档的特征值 (步骤 620)。通过使用所述输入文档的特征值和所述主业余文档的特征值,来使用 Mahalanobis 公式 (表达式 1) 计算在所述输入文档和所述主业余文档之间的距离 (Db)。类似地,使用所述输入文档的特征值和所述主专家文档的特征值用 Mahalanobis 公式 (表达式 2) 来计算在所述输入文档和所述主专家文档之间的距离 (Dc)(步骤 630)。

[0123] (表达式 1)$Db = (A-B)^t \Sigma\, b^{-1}(A-B)$

[0124] (表达式 2)$Dc = (A-C)^t \Sigma\, c^{-1}(A-C)$

[0125] 其中 :A 表示从各自文档中获得的特征值并且被表示为 $A^t = (a1, a2, \ldots, ap)$,B 表示所述业余文档的特征值的平均值,并且被表示为 $B^t = (b1, b2, \ldots, bp)$,C 表示所述专家文档的特征值的平均值,并且被表示为 $C^t = (c1, c2, \ldots, cp)$,p 表示特征向量维度的数目,t 表示矩阵的转置。$\Sigma\, b$ 和 $\Sigma\, c$ 表示各自集的协方差矩阵 (covariancematricies),而 $\Sigma\, b^{-1}$ 和 $\Sigma\, c^{-1}$ 表示所述协方差矩阵的逆矩阵。

[0126] 如果 Db 小于所述预先确定的阈值,那么把所述文档分类为业余文档 (步骤 640)。如果 Dc 小于所述预先确定的阈值,那么把所述文档分类为专家文档 (步骤 650)。

[0127] 把既没有被分类为业余文档也没有被分类为专家文档的文档认为是不可分类的,并且认为其是“其它”文档 (步骤 660)。

[0128] 对于所有的文档执行上述步骤 (步骤 670 和 680)。

[0129] (6) 检测在所述业余文档和所述专家文档之间的对应术语。

[0130] 作为上述处理的结果,可以获得由 N 文档和 E 文档组成的文档对,所述 N 文档和 E 文档描述了特定的共同主题。以下描述在用于 N(业余) 文档和 E(专家) 文档的术语之间的关联。

[0131] 在所述业余文档 (N 文档) 和所述专家文档 (E 文档) 中使用不同的术语。然而,由于所述文档描述共同的主题,所以可以推测使用具有相似含义的对应术语。从而,可以开发出一种标识来自 E 文档和 N 文档的、具有相似含义的词对的方法。所述方法如下 :检测对应于 E 文档中的第 r 个词 Er 的一列业余词,并且检测对应于 N 文档中的第 i 个词 Ni 的一列专家词。下面描述所述细节。

[0132] (I) 极大似然比率测试

[0133] 首先,描述了使用所述极大似然比率测试的计算方法。图 7 是图 2 的计算机结合

极大似然比率 (MLR) 测试执行操作的流程图。

[0134] 710 :计算被分类为业余文档的文档的各自术语的频率。

[0135] 720 :计算被分类为专家文档的文档的各自术语的频率。

[0136] 730 :计算 P(A) = Prob(Ni AND Er)。

[0137] 740 :计算 P(B) = Prob(Not(Ni)AND Er)。

[0138] 750 :根据 P(A) 和 P(B) 来计算 MLR。

[0139] 760 :在 MLR 超出阈值的情况下提取 (Ni) 和 (Er) 的组合。

[0140] 770 :确认对于所有组合执行了计算。

[0141] 780 :设置下一组合。

[0142] 790 :从双向检测对应术语。

[0143] 参考图 7 的流程图,特别描述了图 1 的系统用来检测极大似然比率的方法。

[0144] 考虑下列情况 :假定 (1) 从文档 N 中提取了 m 个术语并且 N 的第 i 个术语是 Ni,(2) 从文档 E 中提取了 n 个术语并且 E 的第 r 个术语是 Er,并且 (3)Ni 和 Er 频繁共同出现。换句话说,假定当 Ni 频繁出现时 Er 也频繁出现,并且当 Ni 很少出现时 Er 也很少出现。描述了用于确定这种情况的概率太高以至于不能被认为一致的条件。另外,将要描述用数值来表示所述概率的可信度的方法。

[0145] 下面描述了用于为业余术语 (在 N 文档中的术语) 寻找对应专家术语 (在 E 文档中的术语) 的方法。

[0146] 考虑一对文档,其根据一个题目被提取并且被分类为业余文档或专家文档。事先确定应该被处理的术语,而不是处理在业余文档和专家文档中的所有术语。为此准备的业余术语列表和专家术语列表存储了那些对应于各自领域的术语。所述业余术语列表存储了与人类感觉和主观判断有关的表达。

[0147] 所述专家术语列表存储了满足下列准则的术语 :

[0148] (a) 包括在所述术语列表内的术语和与那些术语相关的术语

[0149] (b) 未包括在所述业余术语列表中的术语

[0150] (c) 以等于或高于预先确定频率的频率出现的术语

[0151] 假定存在来自所述业余术语列表的 n 个术语,其出现在所述业余文档,并且所述业余术语列表的第 i 个术语是 Ni(i = 1 到 m)。计数所述第 i 个术语的频率 (步骤 710)。类似地,假定在业余术语列表中的术语之间存在在所述专家文档中的 m 个术语,并且所述专家列表的第 r 个术语是 Er(r = 1 到 n)。计数所述专家术语列表的第 r 个术语的频率 (步骤 720)。用于计数所述频率的单位是术语单语法、术语双语法或术语三语法之一。根据在各自文档中 Ni 和 Er 的频率,来如下定义 Ni 和 Er 共同出现的概率 P(A)(步骤 730) 和 Ni 出现而 Er 不出现的概率 P(B)(步骤 740)。

[0152] P(A) = Prob(Ni|Er)

[0153] P(B) = Prob(Not(Ni)|Er)

[0154] 接下来,计算极大似然比率 (MLR)(步骤 750)。把 MLR 计算为下列概率的比例 :(1) 概率 P(H 0),其是如果假定在 P(A) 和 P(B) 之间没有差异 (零假设) 的概率,和 (2) 概率 P(H1),其是如果假定存在差异 (择一假设) 的概率。通过把关注的术语对 (Ni 和 Er) 考虑为依照二项式分布的两个随机过程来计算 MLR。如下给出用于计算一个随机变量的二项式

分布概率的表达式。

[0155] [方程式 1]

[0156] (公式 3) $$b(p,k,n)=\binom{n}{k}p^{k}(1-p)^{(n-k)}$$

[0157] 其中:k 表示某个词实际出现的数目,n 表示所述词出现的最大可能数目,而 p 表示基本出现概率。如果在 H0(零假设)情况下假定概率是 p0,在 H1(择一假设)情况下 P(A) 的假定最大概率是 p1,并且 P(B) 的假定最大概率是 p2,那么把 P(H0) 与 P(H1) 的比例表示为:

[0158] [方程式 2]

[0159] (公式 4) $$\lambda=\frac{P(H0)}{P(H1)}=\frac{b(p0,k1,n1)b(p0,k2,n2)}{b(p1,k1,n1)b(p2,k2,n2)}$$

[0160] 根据所述词出现的数目容易地计算 k1、n1、k2 和 n2 的值。似然比的 MLR 是:

[0161] [方程式 3]

[0162] (公式 5)$MLR=-2\log\lambda$

[0163] 通常已知所述 MLR 基本上遵循具有自由度为 1 的 X2 分布。如果利用这个,那么很容易设置所述阈值。换句话说,如果 MLR 值超出某个数值,那么可以说两个术语 Ni 和 Er 共同出现的概率太高以致于不能被认为是一致的(步骤 760)。

[0164] 利用上述原理,图 2 的计算机使用下列方法来选择词汇映射候选:在相对于所有目标术语的组合,即:{(Ni,Er)i = 1 到 m,r = 1 到 n} 计算所述 MLR(步骤 770 和 780)之后,采用所述数值的降序选择超出预先确定阈值的对,所述阈值例如为 5%。检索在所述专家列表中对应于 N 中的第 i 个术语的术语,所述术语具有超出所述阈值的 MLR 值,并且采用所述 MLR 值的降序来在所述术语之间选择数目预先确定的术语,借此获得对应于业余术语的专家术语(步骤 780)。

[0165] 接下来,描述了图 2 的计算机用于从专家术语(在 E 文档中的术语)中寻找对应的业余术语(在 N 文档中的术语)的方法。

[0166] 采用如同上述类似的方式,从所存储的列表中检索在 N 中对应于 E 中的第 r 个术语的术语,所述术语具有超出所述阈值的 MLR 值,并且采用所述 MLR 值的降序来在所述术语之间选择数目预先确定的术语,借此获得对应于专家术语的业余术语(步骤 780)。

[0167] (ii) 基于词汇映射矩阵计算的方法

[0168] 接下来,描述了基于词汇映射矩阵 T 计算的方法,权重依照文档的长度和术语频率来调整。

[0169] 图 9 是图 1 的系统结合词汇映射矩阵执行操作的流程图。

[0170] 810:创建 s 乘 n 专家术语矩阵 P。

[0171] 820:创建 s 乘 m 业余术语矩阵 Q。

[0172] 830:计算 m 乘 n 词汇映射矩阵 T。

[0173] 840:把业余术语转换为专家术语,并且把专家术语转换为业余术语。

[0174] 以下详细描述了各自步骤 810-840。首先,根据被分类为专家文档的文档集来创建专家术语矩阵 P。这里考虑把在术语列表中的第 k 个术语 (k = 1 到 s) 作为关键词来检索

的文档。处理那些被分类为专家文档的文档以便计算用于所述文档的术语的频率。

[0175] 要加以处理的术语是在上述专家术语列表中的术语。把上述操作应用于文档，所述文档是对在所述术语列表中所有术语检索并且被分类为专家文档的文档，借此计算与在专家术语列表中的术语对应的术语的频率。计算表示专家术语频率的 s 乘 n 矩阵 P_0（未示出），假定 n 是在所述专家文档中术语的数目。

[0176] 类似地，当把 m 假定为在业余文档中术语的数目时，计算表示业余术语频率的 s 乘 m 矩阵 Q_0（未示出）。

[0177] 在相互已经共同出现的两个词之间的连接强度应该更高，然而高频率的词常常与许多其它词共同出现。为此，有必要低估这种词作为词汇映射候选的重要性。类似地，当一个文档长并且包含大量的词时，出现在这种文档中单个词的重要性必须被低估。

[0178] 从而，通过如下转换矩阵 P_0 的元素来创建 s 乘 n 专家术语矩阵（图 8a)（步骤 810）：

[0179] [方程式 4]

[0180] $$We(k,i)=\frac{Exp(k,i)}{(\mathrm{Etf(i)}*\mathrm{Ewf(k)})}$$

[0181] 其中出现在专家文档的第 k 个文档的词的频率是 Exp(k,i)，在所有文档中词的频率是 Etf(i)，而出现在所述第 k 个文档中词的总数是 Ewf(k)。

[0182] 类似地，通过如下转换矩阵 Q_0 的元素来创建 s 乘 m 业余术语矩阵 Q（图 8b)（步骤 820）：

[0183] [方程式 5]

[0184] $$Wn(k,i)=\frac{\mathrm{Naive(k,r)}}{(\mathrm{Ntf(r)}*\mathrm{Nwf(k)})}$$

[0185] 其中出现在业余文档的第 k 个文档中的词的频率是 Naive(k,r)，出现在所有文档中词的频率是 Ntf(r)，而出现在所述第 k 个文档中词的总数是 Nwf(k)。

[0186] 创建 s 乘 n 矩阵 P 和 s 乘 m 矩阵 Q 的目的是计算用于表明那些各自词的组合强度的权重值，以便获得 m 乘 n 词汇映射矩阵 T。从而，如下定义所述矩阵 T：

[0187] $T = Q^tP$

[0188] 其中 t 表示矩阵的转置，如下定义了所述词汇映射矩阵 T 的每个权重值：

[0189] [方程式 6]

[0190] $$W(r,i)=\sum_{k=1}^{s}\left[\frac{Exp(k,i)}{(\mathrm{Etf(i)}*\mathrm{Ewf(k)})}\frac{\mathrm{Naive(k,r)}}{(\mathrm{Ntf(r)}*\mathrm{Nwf(k)})}\right]$$

[0191] 从所述词汇映射矩阵中提取用于映射的候选词。例如，为了提取对应于第 i 个业余术语 Ni 的候选专家术语，查阅词汇映射矩阵 T 的第 i 行并且依照权重值的降序来选择所希望术语的数目就足够了（步骤 840)。

[0192] 另一方面，为了提取对应于第 r 个专家术语的候选业余术语，查阅词汇映射矩阵 T 的第 r 行并且依照权重值的降序来选择所希望术语的数目是就足够了（步骤 840)。在这两种情况中，优选地是，把具有最高权重值的十个词，不包括那些具有零值的词挑选为候选词。

[0193] 然而，由于十个挑选的候选词可能包括不必要的信息，所以所述方法可以不必是

实际应用的。从而，可以利用使用包括在所述术语列表内的术语来进一步过滤候选术语的方法。例如，只将在术语列表中描述的“酒名字”的数据保持在输出中。另外，还可以选择满足非专家的偏好信息的业余术语候选项。例如，可以输出用单语法表示偏好信息的非专家术语，所述单语法诸如“karakuchi（不甜的）”，“shitazawari-ga-yoi（好构造）”和“ajiwai-bukai（美味的）”或者对应于表示非专家的术语的双语法组合的“酒名字”。因此，匹配非专家偏好的“酒名字”可以是以非专家偏好信息为基础的。下面讨论在应用该过滤之后的输出例子。

[0194] 下面示出了检索的取样结果。

[0195] 下列例子是作为那些对应于业余术语而检索的专家术语的示例。当在日本因特网站点搜索领域“nihonshu（日本米酒）”时，检测下列业余（非专家）术语：“atsui（强烈的）”、“yutaka（醇厚的）”、“tanrei（明亮且精美的）”、“sararitof（醇合的）”、“bimi（味美的）”、“fukami（浓度）”等。对应于那些业余术语的专家术语分别在检索下列酒名字时产生：“Isojiman”对应“强烈的”和“醇厚的”，“Koshinokanbai”对应“明亮的和精美的”和“醇合的”，而“Kamomidori”对应“味美的”和“浓度”。

[0196] 当在日本因特网站点搜索“酒”领域时，检测到了下列非专家术语：“bimi（味美的）”、“koi（稠的）”、“umami（美味的）”、“suppai（酸的）”、“shitazawari（构造）”、“kire（清晰度）”、“pittari（确切匹配）”、“fukami（浓度）”、“sawayaka（淡的）”、“yawarakaf（不含酒精的）”、“amaroyakaf（醇合且不含酒精的）”等。对应于那些业余术语的专家术语分别在检索下列酒名字时产生：“Au Bon Climat”对应“味美的”、“稠的”、“美味”、“酸的”等，而“Zonnebloem”对应“构造”、“清晰度”、“匹配”、“浓度”、“淡的”、“不含酒精的”、“醇合且不含酒精的”等。

[0197] 下列例子是作为对应于专家术语检测的那些示例业余术语。

[0198] 当在日本因特网站点搜索“nihonshu（日本米酒）”领域时，检测到了是酒名字的下列专家术语：“Kagatobi”、“Hanano-mai”、“Kakubuto”等。作为对应于这些酒名字而检索的业余术语包括以下：“oishii（鲜美的）”、“mizumizushii（凉爽的）”对应“Kagatobi”，“johin（优雅的）、tanrei（明亮且精美的）”对应“Hanano-mai”，而“nameraka（不含酒精的并且醇美的）、sawciyaka（凉且淡的）、subarashii（极好的）”对应“kakubuto”。

[0199] 当在日本站点搜索“酒”领域时，检测到了是酒名字的下列专家术语：“Coltassala”、“Sansoniere”等。作为那些对应于这些酒名字而检索的业余术语包括以下：“awai（半透明的）、kihm（优雅的）、honoka（暗淡的）、karui（明亮的）、kokochiyoi（舒适的）”对应“Coltassala”，而“horonigai（略苦）、karai（不甜的）、johin（优雅的）、yuuga（雅致的）”对应“Sansoniere”。

[0200] 采用上述两种词汇映射方法，可以通过依照术语权重值的降序选择术语，来在双向上选择对应于专用术语的候选术语，所述双向为N->E（非专家到专家）和E->N。

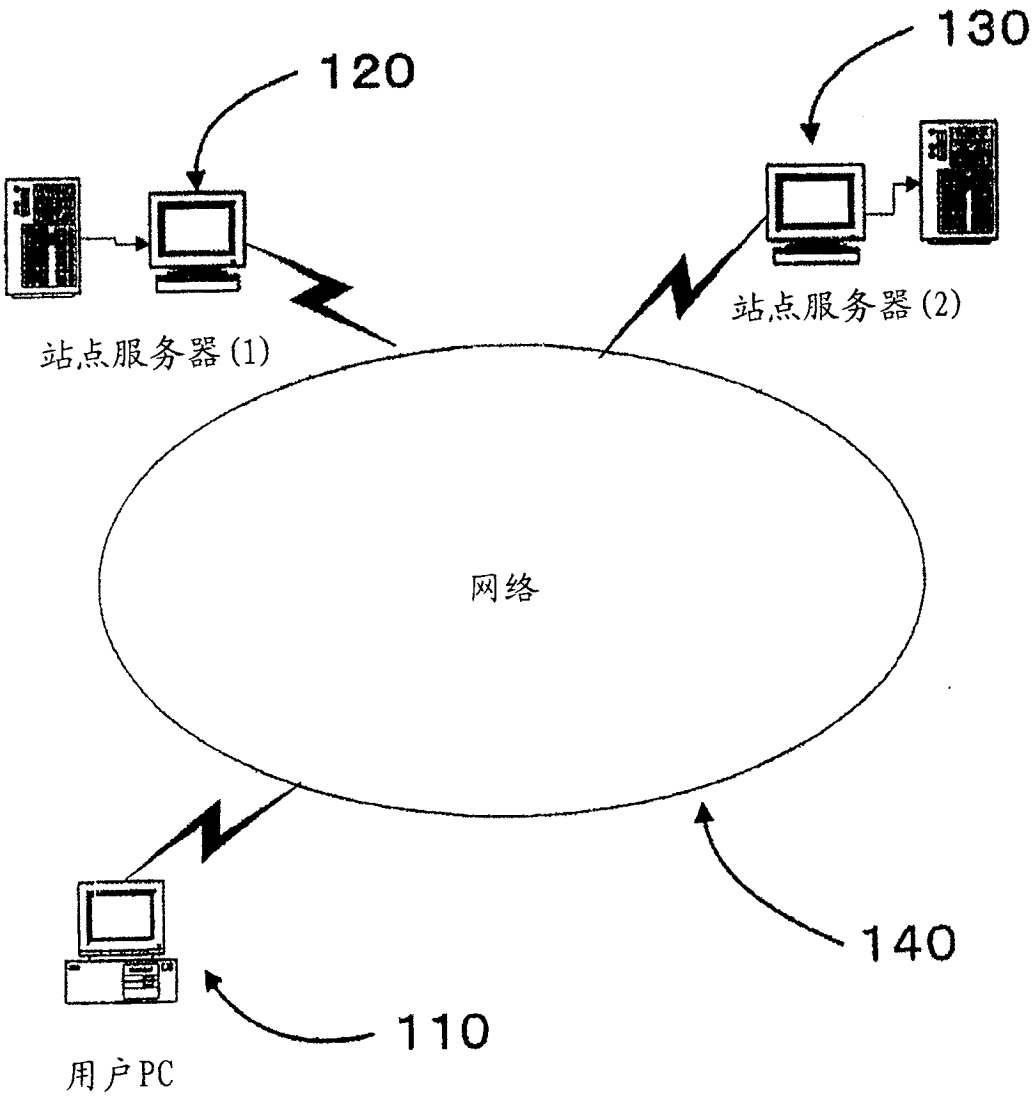


图 1

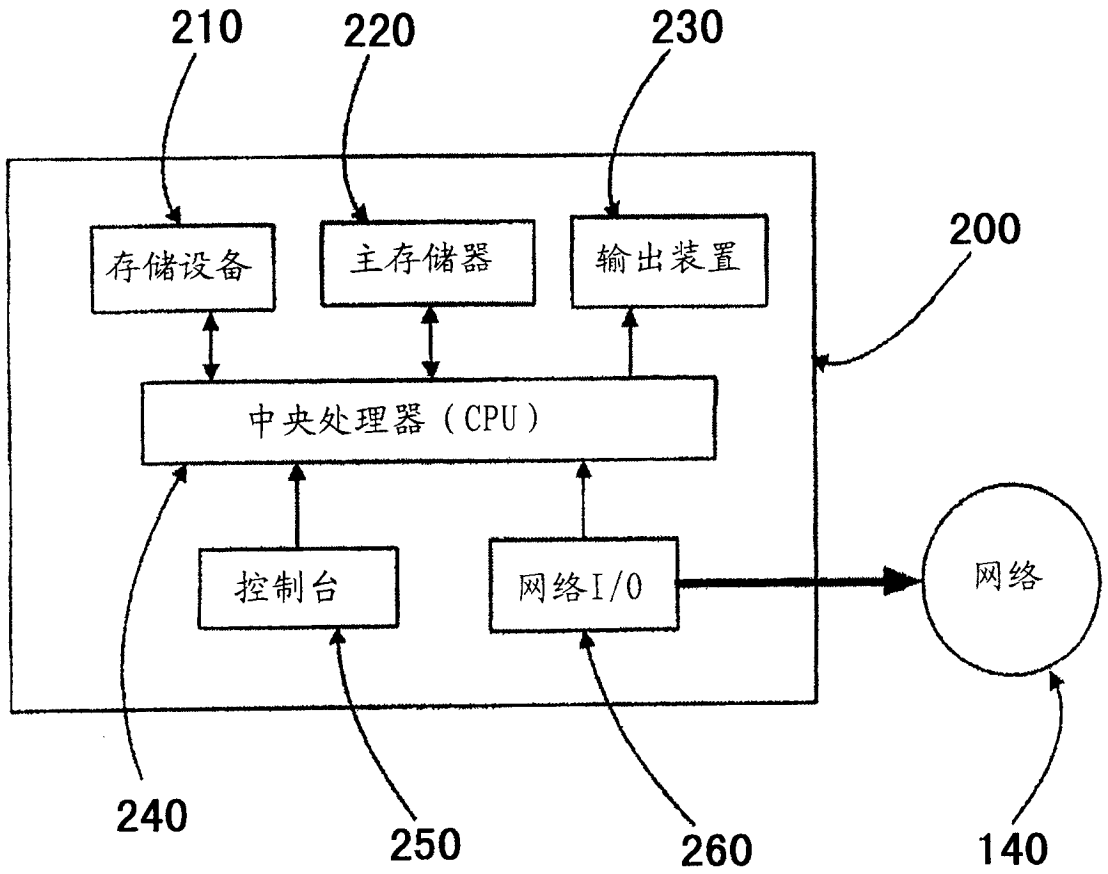

210
220
230
200
存储设备
主存储器
输出装置
中央处理器（CPU）
控制台
网络I/O
网络
240
250
260
140

图 2

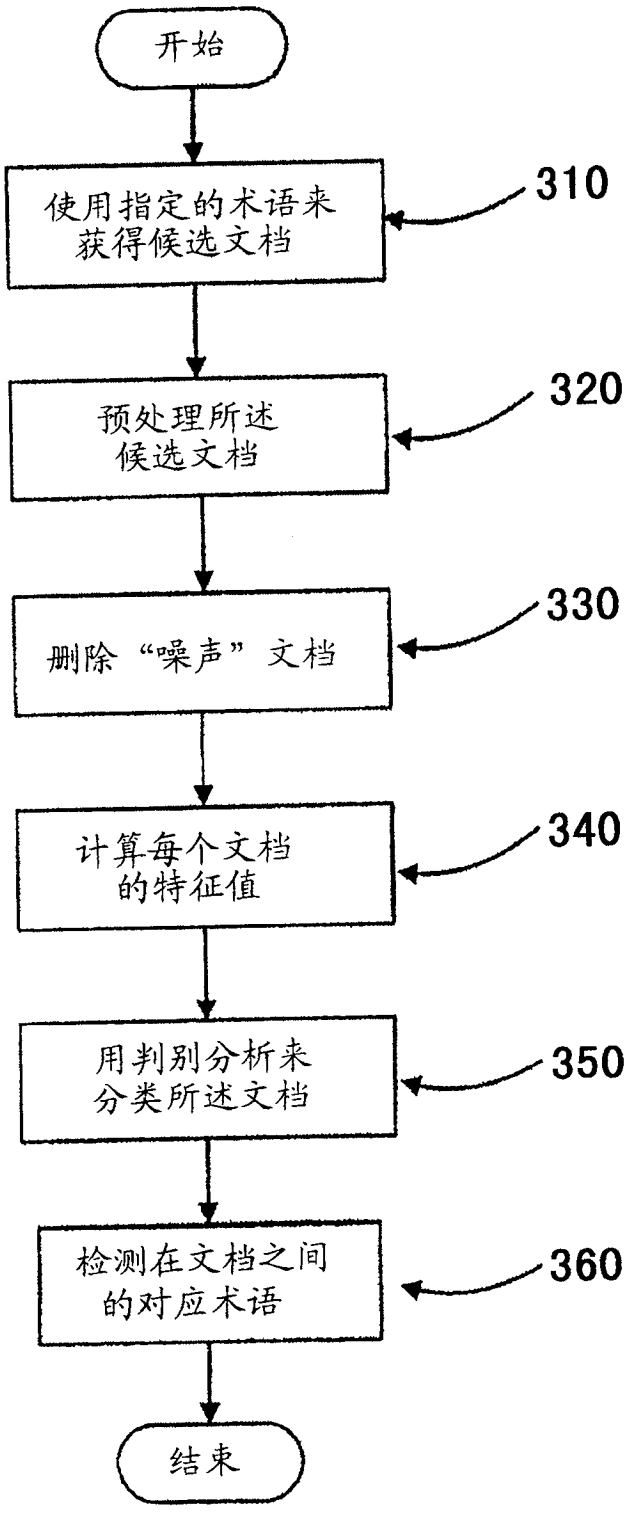

开始
使用指定的术语来获得候选文档
310
预处理所述候选文档
320
删除“噪声”文档
330
计算每个文档的特征值
340
用判别分析来分类所述文档
350
检测在文档之间的对应术语
360
结束

图 3

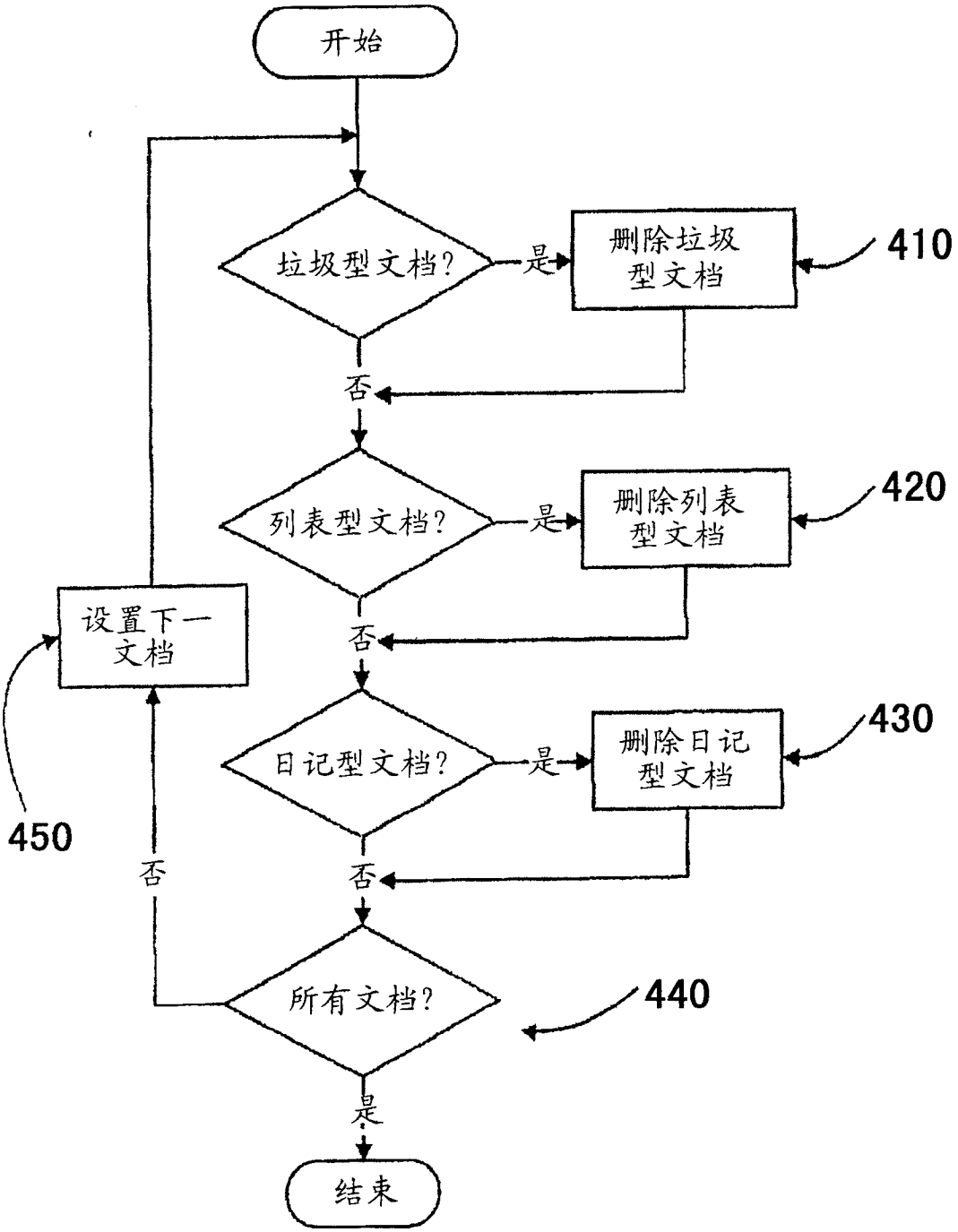

开始
垃圾型文档?
是
删除垃圾型文档
410
否
列表型文档?
是
删除列表型文档
420
否
日记型文档?
是
删除日记型文档
430
否
所有文档?
440
是
结束
设置下一文档
450
否

图 4

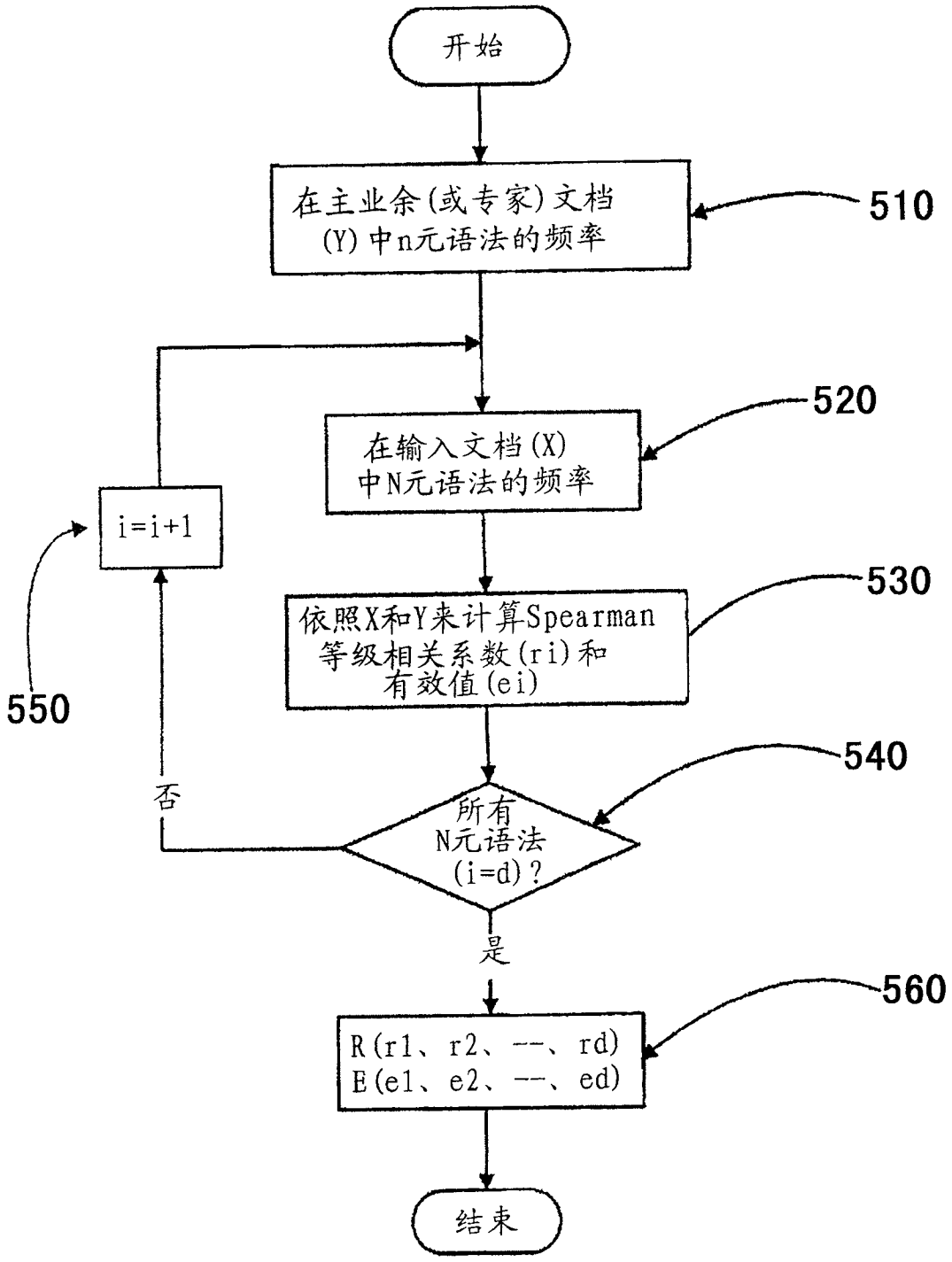

开始
在主业余(或专家)文档(Y)中n元语法的频率
510
在输入文档(X)中N元语法的频率
520
依照X和Y来计算Spearman等级相关系数(ri)和有效值(ei)
530
所有N元语法(i=d)?
540
否
是
i=i+1
550
R(r1、r2、--、rd)
E(e1、e2、--、ed)
560
结束

图 5

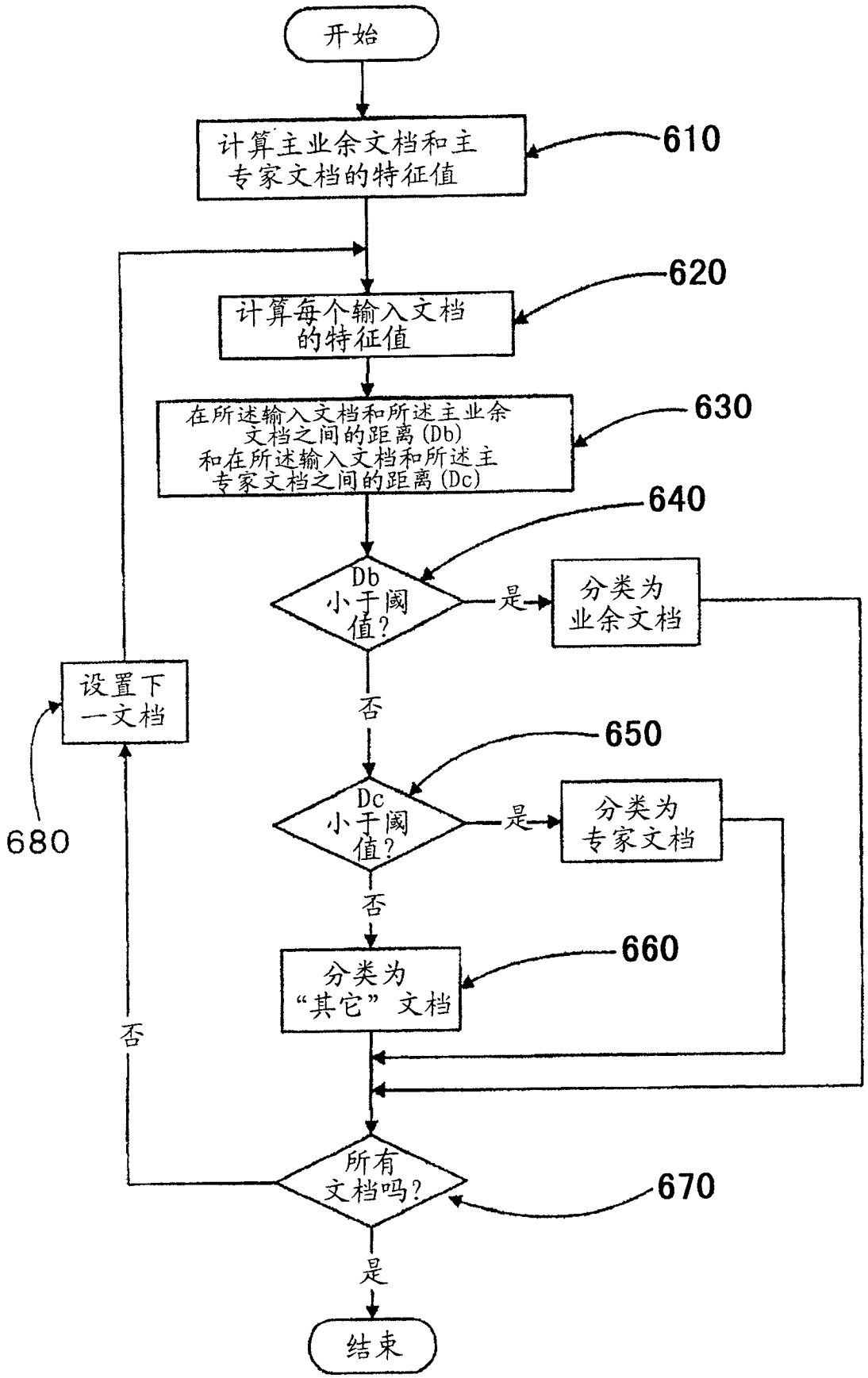

开始
计算主业余文档和主专家文档的特征值
610
计算每个输入文档的特征值
620
在所述输入文档和所述主业余文档之间的距离(Db)和在所述输入文档和所述主专家文档之间的距离(Dc)
630
640
Db小于阈值?
是
分类为业余文档
否
650
Dc小于阈值?
是
分类为专家文档
否
660
分类为“其它”文档
所有文档吗?
670
是
结束
否
设置下一文档
680

图 6

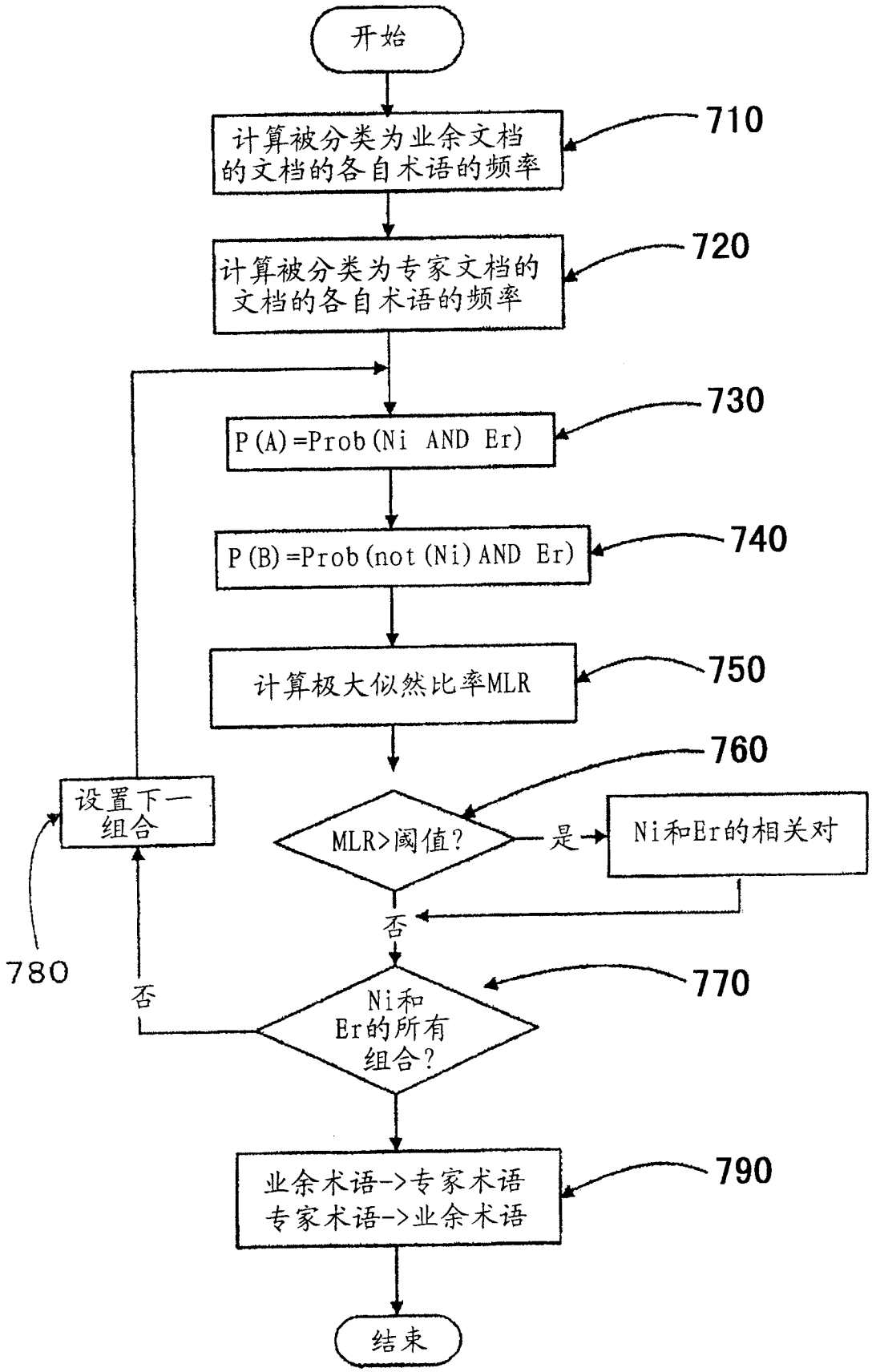

开始
计算被分类为业余文档的文档的各自术语的频率
710
计算被分类为专家文档的文档的各自术语的频率
720
P(A)=Prob(Ni AND Er)
730
P(B)=Prob(not(Ni) AND Er)
740
计算极大似然比率MLR
750
760
MLR>阈值?
是
Ni和Er的相关对
否
780
设置下一组合
否
Ni和Er的所有组合?
770
业余术语->专家术语
专家术语->业余术语
790
结束

图 7

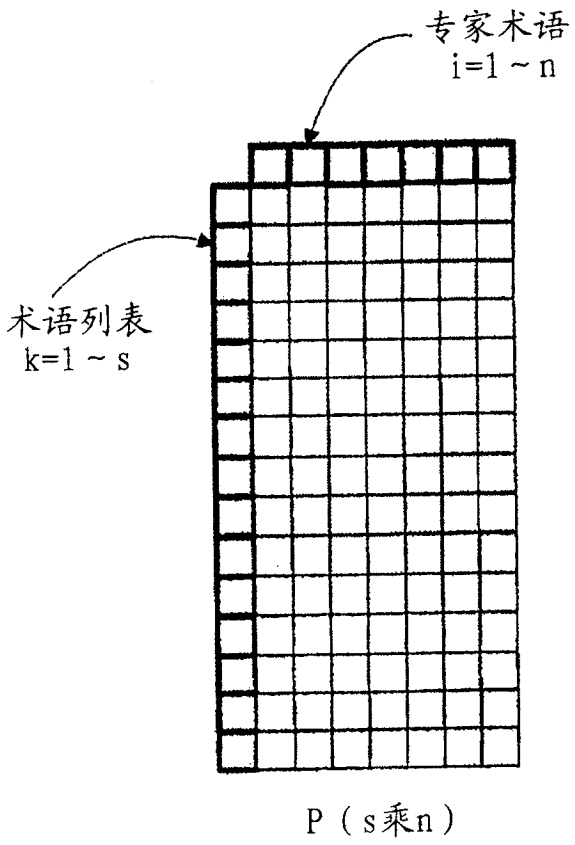

专家术语
i=1～n
术语列表
k=1～s
P（s乘n）

图 8a

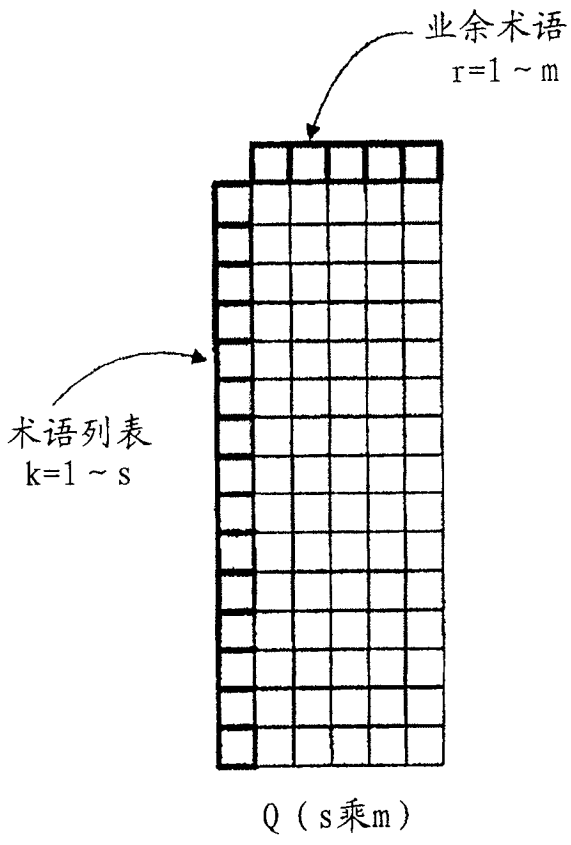

业余术语
r=1～m
术语列表
k=1～s
Q（s乘m）

图 8b

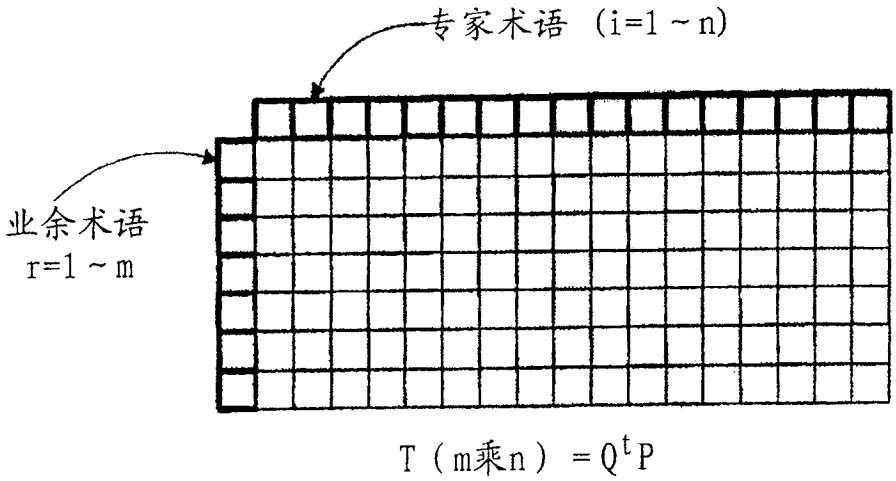

专家术语 (i=1～n)
业余术语
r=1～m
T（m乘n）=Q^tP

图 8c

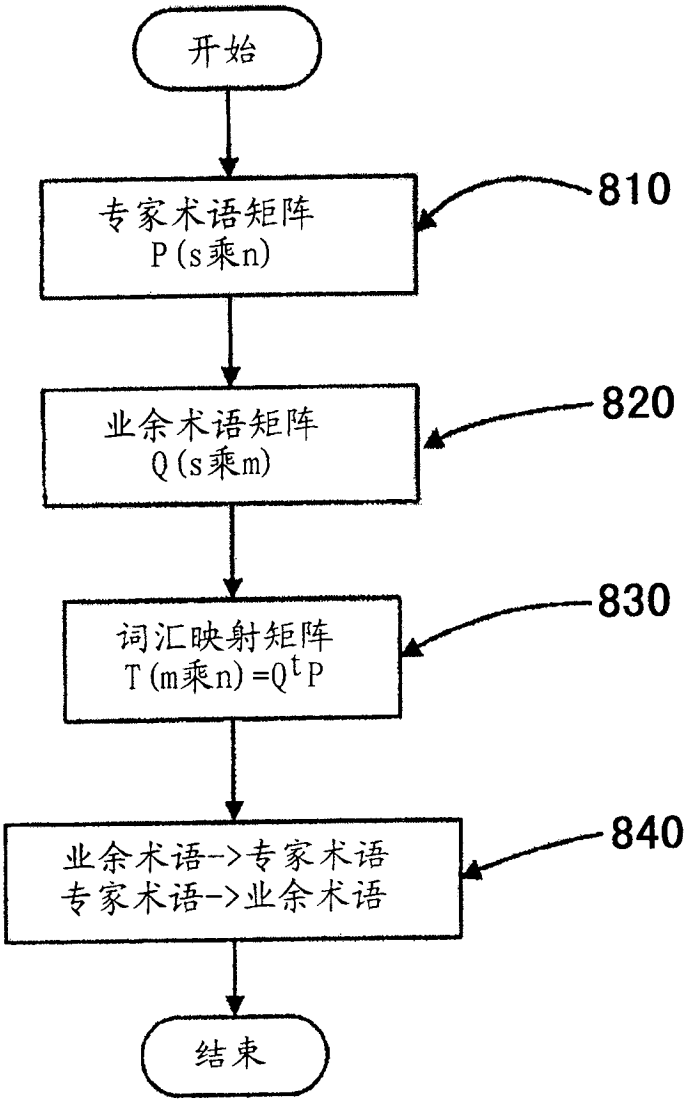

开始
专家术语矩阵
P(s乘n)
810
业余术语矩阵
Q(s乘m)
820
词汇映射矩阵
T(m乘n)=Q^tP
830
业余术语->专家术语
专家术语->业余术语
840
结束

图 9