(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 1716294 B
(45)授权公告日 2013.09.11

(21)申请号 200510082404.0
(22)申请日 2005.06.30
(30)优先权数据
10/881,867 2004.06.30 US
(73)专利权人 微软公司
地址 美国华盛顿州
(72)发明人 B·章 H-J·曾 马维英 陈正
(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 胡利鸣
(51)Int.Cl.
G06F 17/30(2006.01)
G06F 21/62(2013.01)
H04L 12/58(2006.01)
(56)对比文件
US RE.35861 E,1998.07.28, 摘要第1-4行,第2栏第28-39行.
US 2003/0149687 A1,2003.08.07, 第0005、0021、0022段,图2、3.
US 6658626 B1,2003.12.02, 第3栏第1-32行,第5栏第17-28、40-43行,第6栏第14-16、32-42行,第7栏第6-27行,第8栏第45行-第9栏第5行,第11栏第3-13行,第11栏第65行-第12栏第29行,第14栏第29-58行,第21栏第8-35行,图1A、2、4A.
审查员 俞立文

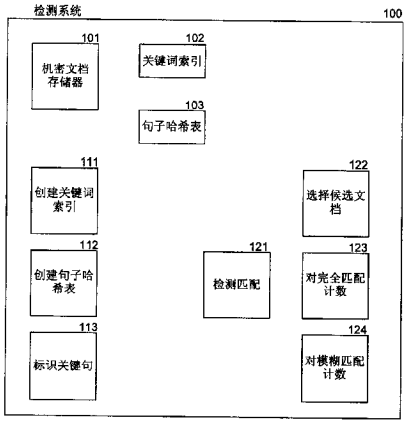
权利要求书4页 说明书8页 附图10页

(54)发明名称

用于检测外发通信何时包含特定内容的方法和系统

(57)摘要

提供了一种用于检测外发通信是否包含机密信息或其它目标信息的方法和系统。检测系统带有包含机密信息的文档集合,称为"机密文档"。当向检测系统提供外发通信时,该系统把外发通信的内容与机密文档的内容相比较。如果外发通信包含机密信息,则检测系统就防止在机构外部发送该外发通信。检测系统基于外发通信内容和已知包含该机密信息的机密文档的内容之间的相似性来检测机密信息。



CN 1716294 B

1. 计算机系统中一种用于标识外发通信是否包含机密信息的方法，所述方法包括：

提供包含机密信息的文档；

生成一个把关键词映射到包含所述关键词的文档的关键词索引，其中关键词是其所计算的重要性大于一重要性阈值的单词；

生成一个把关键句的哈希码映射到包含所述关键句的文档的句子哈希表，其中对于包含所述关键句的文档的每个段落，所述段落中其关键词与该段落关键词最为相似的句子被标识为该段落的关键句；

接收外发通信；

标识所述外发通信的关键词；

基于所述文档的关键词和所述外发通信的所标识的关键词之间的相似性使用生成的关键词索引来定位候选文档；

生成所述外发通信的关键句哈希码；

使用生成的句子哈希码以标识那些包括其哈希码与生成的所述外发通信的关键句哈希码相同的关键句的候选文档；

将所标识的候选文档的关键句与所述外发通信的关键句进行比较；以及

在确定所述外发通信的至少一阈值条关键句与所标识的候选文档的关键句匹配后，将所述外发通信标记为包含机密信息。

2. 如权利要求 1 所述的方法，其特征在于，所提供的文档和外发通信是电子邮件。

3. 如权利要求 2 所述的方法，其特征在于，所述计算机系统是一电子邮件服务器。

4. 如权利要求 1 所述的方法，包括在确定接收到的外发通信包含机密信息时，禁止把接收到的外发通信传送到其目标受信者。

5. 如权利要求 1 所述的方法，其特征在于，所述单词基于检索词频率乘以反转文档频率度量被标识为关键词。

6. 如权利要求 1 所述的方法，包括生成一个把关键词映射到包含所述关键词的文档的句子的关键词索引，其中所述比较包括使用关键词索引来定位包含接收到的外发通信的关键词的句子。

7. 如权利要求 6 所述的方法，其特征在于，当所定位的句子与接收到的外发通信的句子相似时，接收到的外发通信包含机密信息。

8. 如权利要求 1 所述的方法，其特征在于，所述外发通信是一电子邮件。

9. 如权利要求 1 所述的方法，其特征在于，所述外发通信是电子邮件的附件。

10. 如权利要求 1 所述的方法，其特征在于，所述外发通信是一即时消息。

11. 如权利要求 1 所述的方法，其特征在于，所述外发通信是一语音通信。

12. 如权利要求 1 所述的方法，其特征在于，所述外发通信是一互联网记录。

13. 一种用于标识文档是否包含与目标文档内容相似的内容的方法，所述方法包括：

基于目标文档和所述文档的关键词之间的相似性从目标文档中选择候选文档，其中所述选择包括；

创建一关键词索引，所述关键词索引把目标文档的关键词映射到包含该关键词的目标文档，其中关键词是其所计算的重要性大于一重要性阈值的单词；

标识所述文档的关键词；以及

使用所述创建的关键词索引从所述目标文档中标识候选文档，各个候选文档包括与所述文档的标识的关键词相似的关键词；以及

把候选文档与所述文档相比较以确定所述文档是否包含与候选文档相似的内容。

14. 如权利要求 13 所述的方法，其特征在于，所述关键词基于检索词频率乘以反转文档频率度量来标识。

15. 如权利要求 13 所述的方法，包括生成一句子哈希表，所述句子哈希表把从句子导出的哈希码映射到包含所述句子的目标文档，其中所述比较包括使用句子哈希表来定位包含与文档句子相匹配的句子的候选文档。

16. 如权利要求 15 所述的方法，其特征在于，所述句子哈希表映射到目标文档的关键句。

17. 如权利要求 13 所述的方法，包括生成一关键词索引，所述关键词索引把关键词映射到包含所述关键词的目标文档的句子，其中所述比较包括使用关键词索引来定位包含文档关键词的候选文档的句子。

18. 如权利要求 13 所述的方法，其特征在于，所述目标文档包含机密信息。

19. 如权利要求 18 所述的方法，其特征在于，当所述文档是包含机密信息的外发通信时，禁止发送所述外发通信。

20. 如权利要求 13 所述的方法，其特征在于，所述文档是电子邮件，所述比较找到相关的电子邮件。

21. 如权利要求 13 所述的方法，包括生成一句子哈希表，所述句子哈希表把从句子导出的哈希码映射到包含所述句子的目标文档，还包括生成一关键词索引，所述关键词索引把关键词映射到包含所述关键词的目标文档的句子，其中所述比较包括使用句子哈希表来定位包含与文档句子匹配的句子的候选文档，当句子不匹配时，使用所生成的关键词索引来确定所述文档的句子是否与候选文档的句子相似。

22. 计算机系统中一种用于标识外发通信是否包含机密信息的系统，所述系统包括：

用于提供包含机密信息的文档的装置；

用于生成一个把关键词映射到包含所述关键词的文档的关键词索引的装置，其中关键词是其所计算的重要性大于一重要性阈值的单词；

用于生成一个把关键句的哈希码映射到包含所述关键句的文档的句子哈希表的装置，其中对于包含所述关键句的文档的每个段落，所述段落中其关键词与该段落关键词最为相似的句子被标识为该段落的关键句；

用于接收外发通信的装置；

用于标识所述外发通信的关键词的装置；

用于基于所述文档的关键词和所述外发通信的所标识的关键词之间的相似性使用生成的关键词索引来定位候选文档的装置；

用于生成所述外发通信的关键句哈希码的装置；

用于使用生成的关键句哈希码以标识那些包括其哈希码与生成的所述外发通信的关键句哈希码相同的关键句的候选文档的装置；

用于将所标识的候选文档的关键句与所述外发通信的关键句进行比较的装置；以及

用于在确定所述外发通信的至少一阈值条关键句与所标识的候选文档的关键句匹配

后，将所述外发通信标记为包含机密信息的装置。

23. 如权利要求 22 所述的系统，其特征在于，所提供的文档和外发通信是电子邮件。

24. 如权利要求 22 所述的系统，其特征在于，所述关键词基于检索词频率乘以反转文档频率度量来标识。

25. 如权利要求 22 所述的系统，包括用于生成一个把关键词映射到包含所述关键词的文档的句子的关键词索引的装置，其中所述用于比较的装置包括用于使用关键词索引来定位包含接收到的外发通信的关键词的句子的装置。

26. 如权利要求 22 所述的系统，其特征在于，所述外发通信包含机密信息。

27. 如权利要求 26 所述的系统，其特征在于，当所述外发通信是包含机密信 息的外发通信时，禁止发送所述外发通信。

28. 如权利要求 22 所述的系统，其特征在于，所述外发通信是电子邮件，所述比较找到了相关的电子邮件。

29. 一种用于标识文档是否包含与目标文档内容相似的内容的系统，所述系统包括 ：

用于基于目标文档和所述文档的关键词之间的相似性从目标文档中选择候选文档的装置，其中所述用于选择的装置包括 ；

用于创建一关键词索引的装置，所述关键词索引把目标文档的关键词映射到包含该关键词的目标文档，其中关键词是其所计算的重要性大于一重要性阈值的单词 ；

用于标识所述文档的关键词的装置 ；以及

用于使用所述创建的关键词索引从所述目标文档中标识候选文档的装置，各个候选文档包括与所述文档的标识的关键词相似的关键词 ；以及

用于把候选文档与所述文档相比较以确定所述文档是否包含与候选文档相似的内容的装置。

30. 如权利要求 29 所述的系统，其特征在于，所述系统是一电子邮件服务器，并且所述文档是电子邮件。

31. 如权利要求 30 所述的系统，进一步包括包括用于当确定电子邮件包含与所述候选文档相似的内容时，禁止把电子邮件传送到目标受信者。

32. 一种用于标识通信是否包含目标信息的方法，所述方法包括 ：

提供包含目标信息的文档 ；

生成一个把关键词映射到包含所述关键词的文档的关键词索引，其中关键词是其所计算的重要性大于一重要性阈值的单词 ；

生成一句子哈希表，所述句子哈希表把关键句的哈希码映射到包含所述关键句的文档，其中对于包含所述关键句的文档的每个段落，所述段落中其关键词与该段落关键词最为相似的句子被标识为该段落的关键句 ；

接收一通信 ；

标识所述通信的关键词 ；

生成所述通信的关键句的哈希码 ；

使用生成的句子哈希表来从候选文档中标识包括其哈希码与生成的所述通信的关键句哈希码相同的关键句的文档，其中所述候选文档是从包含与所述通信的关键词相似的关键词的文档中标识的 ；

将所标识的文档的关键句与所述通信的关键句进行比较以确定它们是否匹配；以及

基于匹配的程度，将所述通信标记为包含目标信息。

33. 如权利要求 32 所述的方法，其特征在于，所提供的文档和接收到的通信是电子邮件，所述目标信息是机密的。

34. 如权利要求 32 所述的方法，其特征在于，当确定接收到的通信包含目标信息时，禁止把接收到的通信传送到其目标受信者。

35. 如权利要求 32 所述的方法，其特征在于，其中所述比较包括基于接收到的通信的关键词、使用关键词索引来定位候选文档。

36. 如权利要求 32 所述的方法，包括生成一关键词索引，所述关键词索引把关键词映射到包含所述关键词的文档的句子，其中所述比较包括使用关键词索引来定位包含接收到的通信的关键词的句子。

37. 如权利要求 32 所述的方法，其特征在于，接收到的通信是一电子邮件。

38. 如权利要求 32 所述的方法，其特征在于，接收到的通信是一网页。

39. 如权利要求 32 所述的方法，其特征在于，所提供的文档是网页。

40. 如权利要求 32 所述的方法，其特征在于，所述通信是电子邮件的附件。

41. 如权利要求 32 所述的方法，其特征在于，所述通信是一即时消息。

42. 如权利要求 32 所述的方法，其特征在于，所述通信是一语音通信。

43. 如权利要求 32 所述的方法，其特征在于，所述通信是一互联网记录。

用于检测外发通信何时包含特定内容的方法和系统

技术领域

[0001] 所述技术一般涉及检测文档何时有相似的内容，尤其涉及检测外发通信何时包含特定的内容。

背景技术

[0002] 许多机构都开发了机密的、商业秘密的、所有的信息以及对于每一个这样的机构的成功运作重要的其它信息。在许多情况下，机构确保该信息不在机构外部被公开是至关重要的。如果这种信息在机构外部被公开，信息就可能变得毫无用处，或者会对机构造成实质性的损害。例如，制造公司可以开发一系列特征以结合在产品的下一版本中。如果竞争者能够在发布下一版本前确认这一系列特征，则竞争者就能使用该信息以便有益于他们的竞争。举另一个例子，机构可能需要对违反了机构某一规则的雇员采取内部惩罚措施。如果违规变得公开，它就可以表示机构的公众关系问题。为了确保他们的机密信息未被不适当地公开，许多机构实现了昂贵的手段来确保不出现这种公开。例如，一些公司对他们的雇员实施训话以确保他们理解保持商业秘密的机密性的重要性、确保雇员知道要把包含商业秘密的所有文档都标记为机密、等等。

[0003] 尽管电子通信允许机构的雇员有效地且高效地通信，然而电子通信也使机密信息容易且快速地散步在机构外。例如，如果设计队伍的领导者向队伍的成员们发送了一电子邮件、详细说明了产品下一版本的新特征，那么队伍的任一成员都可能把邮件转发给公司的其它雇员、或甚至转发给竞争者公司的雇员。这种机密信息到竞争者公司雇员的散步会是疏忽的或故意的。例如，雇员可能希望把详细说明新特征的电子邮件转发给公司市场队伍的几个成员。在转发电子邮件时，雇员可能输入目标受信者的部分名字。然而，如果目标受信者具有与竞争者公司雇员相似的名字，那么电子邮件程序可能把该部分名字解析到竞争者公司雇员的电子邮件地址。即使公开是疏忽所至，公司仍然会被严重损害。当雇员故意把有机密信息的电子邮件转发到未被授权接收该信息的某人时，将产生更大的问题。在这一情况下，雇员可能通过例如从电子邮件中删除机密性的告示（例如“该文档包含 Acme 公司的机密、所有权的以及商业秘密信息。”），从而尝试掩盖该信息的机密性。此外，机密信息未经授权的公开不限于电子邮件；未经授权的公开可以采取其它形式的电子通信。例如，雇员可以经由互联网新闻和讨论组、即时消息传递系统、电子邮件的附件、通讯稿、电子介绍、出版物等等来公开机密信息。

[0004] 一些电子邮件系统具有过滤电子邮件以确保它们不包含不适当内容的特征。例如，这一系统可以扫描外发消息中是否有机密信息的指示，比如单词“所有权”、“机密的”或“商业秘密”。如果在邮件中找到这样的单词，则系统会禁止发送该邮件。然而，不是所有包含机密信息的电子邮件都包含这样的单词。例如，设计队伍中的雇员会向其它人频繁地发送电子邮件以便得到关于新理念的非正式反馈。在这种情况下，电子邮件一般不会包含机密性的告示。此外，故意要把机密信息发送给竞争者的雇员可以通过在转发前从邮件中删除这些单词，从而容易地避免这种系统的检测。

[0005] 期望有一种系统，该系统能容易地检测到电子邮件中、更为一般的是在任何外发通信（例如出版、新闻组记录和电子邮件附件）中机密信息的存在。在电子邮件的情况下，这一系统会能检测到：雇员不加任何修改而仅仅转发原始的电子邮件、雇员把原始电子邮件的部分剪切和粘贴到新的电子邮件中、雇员加上附加评论来转发原始电子邮件的各部份、雇员修改原始电子邮件的内容、等等。此外，由于机构可能生成的电子邮件的容量，因此希望这一系统能快速地检测电子邮件中的这种机密信息，而不会显著地延迟传递、并且无需对附加的硬件和软件作出重大投资以便支持这种检测。

发明内容

[0006] 提供了一种用于标识通信是否包含与目标文档内容相似的内容的基于计算机的方法和系统。该系统把候选文档标识为包含与通信关键词相似的关键词的那些目标文档。然后，系统把候选文档与通信相比较以确定通信是否包含与候选文档相似的内容。当通信是一外发通信时，比如包含与候选文档相似内容的电子邮件，系统可以禁止外发通信的传递。

附图说明

[0007] 图 1 是说明一实施例中一检测系统的组件的框图。

[0008] 图 2 是说明一实施例中图 1 的检测系统的数据结构的框图。

[0009] 图 3 是说明一实施例中创建关键词索引的处理的流程图。

[0010] 图 4 是说明一实施例中创建句子哈希表的处理的流程图。

[0011] 图 5 是说明一实施例中、基于句子关键词和段落关键词的相似性进行的标识关键句成分的处理的流程图。

[0012] 图 6 是说明一实施例中、基于反转句子频率进行的标识关键句成分的处理的流程图。

[0013] 图 7 是说明一实施例中、检测匹配成分的处理的流程图。

[0014] 图 8 是说明一实施例中、选择候选文档成分的处理的流程图。

[0015] 图 9 是说明一实施例中、对完全匹配分量计数的处理的流程图。

[0016] 图 10 是说明一实施例中、对模糊匹配分量计数的处理的流程图。

具体实施方式

[0017] 提供了一种用于检测外发通信是否包含机密信息或其它目标信息的方法和系统。在一实施例中，检测系统带有包含机密信息的文档集合，称为“机密文档”。例如，当外发通信是一电子邮件时，机密文档可能是前面发送的包含机密信息的电子邮件。当向检测系统提供外发通信时，该系统把外发通信的内容与机密文档的内容相比较。如果比较表明外发通信包含机密信息，则检测系统就防止在机构外部发送该外发通信。例如，检测系统可以作为机构的内部电子邮件用户和外部电子邮件用户之间的电子邮件网关的一部分来实现。这样，检测系统基于外发通信内容和已知包含该机密信息的机密文档的内容之间的相似性来检测机密信息，并且不需要依赖于可被容易删除的机密性告示。

[0018] 由于机构的雇员每天可能在机构外发送成千上百个电子邮件，且机构可能有几千

个机密文档，因此仅仅把每个外发通信的每个句子与每个机密文档的每个句子相比较是不切实际的。实际上，比较的计算复杂度可能为O(N×M)，其中N是机密文档的数目，M是外发通信的数目。在一实施例中，检测系统用各种辅助的数据结构来组织机密文档，以确保能够快速标识外发通信中的机密信息。检测系统可以生成一索引，该索引把机密文档的关键词映射到包含关键词的那些机密文档。例如，几个机密文档可能包含短语“新产品发布”。在这一情况下，关键词“新”、“产品”和“发布”可能被映射到那些机密文档的每一个。当检测系统接收到外发通信时，它标识外发通信的关键词。然后，检测系统可以使用关键词索引来标识哪些机密文档包含相似的关键词。例如，检测系统可以选择和外发通信有大量共同关键词的那些机密文档。然后，检测系统可以把外发通信的内容与所标识机密文档（也称为候选文档）的内容相比较，以确定外发通信是否真的包含机密信息。例如，如果外发通信包含关键词“新”、“产品”和“发布”，但每个关键词都在不同的句子中使用，则检测系统可能标识出包含短语“新产品发布”的几个机密文档。然而，当检测系统把外发通信的内容与候选文档的实际内容相比较时，它不会检测到相似性，因此会允许发送外发通信。检测系统可以使用各种技术来标识文档内的关键词。例如，检测系统可以使用检索词频率乘以反转文档频率度量（即“TF*IDF”）来标识关键词。本领域的技术人员会理解，可以使用其它度量。例如，给定文档中的单词、文档的元数据（例如关键词属性、摘要属性和标题属性）等等，检测系统可能以侧重点（例如字体大小、字体磅值和下划线）为因素。通过使用关键词索引，检测系统能有效地把机密文档限定为一组候选文档，以便进一步比较。

[0019] 在一实施例中，检测系统使用辅助的数据结构，比如哈希表，来帮助标识哪些候选文档类似于外发通信。检测系统可以生成一哈希表，该哈希表把为每个句子导出的哈希码映射到包含那些句子的机密文档。检测系统通过向机密文档的每个句子应用哈希函数以便为每个句子生成一哈希码，从而生成句子哈希表。然后，检测系统保存哈希码到机密文档内相应句子的映射。在检测系统为外发通信标识了候选文档后，检测系统为外发通信的句子生成哈希码。检测系统使用所生成的哈希码来标识哪些候选文档包含具有相同哈希码的句子。检测系统接着可以把所标识的句子与外发通信的相应句子相比较，以确定它们是否匹配（即完全匹配或者类似）。根据匹配程度（例如外发通信的句子与机密文档的句子匹配的次数），检测系统可以把外发通信标记为包含机密信息。为了加速外发通信的过程，检测系统可以仅分析机密文档和外发通信的“关键句”。例如，关键句可以对应于一个段落的主题句。

[0020] 在一实施例中，检测系统可以使用另一种辅助数据结构，比如关键句索引，来帮助标识哪些候选文档类似于外发通信。检测系统可以生成一关键句索引，该索引把机密文档的关键词映射到包含那些关键词的那些机密文档内的句子。在检测系统标识了候选文档后，检测系统可以使用关键句索引来计算外发通信的每个句子和候选文档的每个句子之间的相似性。检测系统可以使用各种相似性度量的任一个，比如余弦相似性和编辑距离。基于相似性程度（例如机密文档中与外发通信的句子相似的句子数），检测系统把外发通信标记为包含机密信息。

[0021] 在一实施例中，检测系统将其分析基于机密文档和外发通信的“关键句”，而不是对每个句子执行其分析。“关键句”是表示机密文档或机密文档内一个段落的关键思想的句子。检测系统可以以各种方式来标识文档的关键句。检测系统可以计算一个段落的每个句

子与该段落的相似性。具有与该段落的最高相似性的句子可以被视为该段落的关键句，它代表了该段落的机密信息并因此是该段落最重要的句子。为了计算相似性，检测系统可以用其关键词来表示该段落和每个句子。然后，检测系统计算每个句子的关键词与段落关键词之间的相似性。检测系统把具有最高相似性的句子选择作为关键句。或者，检测系统可以用检索词频率乘以反转句子频率度量（即 TF*ISF）来标识关键句，以计算句子对于段落的重要性。反转句子频率像反转文档频率一样，反映了文档中的句子数除以包含该单词的句子数。检测系统通过把一单词在一句子中的出现次数和该句子的反转句子频率相乘，从而计算该单词对于该句子的重要性。然后，检测系统可以把每个句子的重要性设为单词在句子内的平均重要性。检测系统把具有最高重要性的句子选择作为关键句。本领域的技术人员会理解，关键句可以从句子对文档的总体重要性或相似性而导出，而不是逐段地导出。

[0022] 在一实施例中，检测系统可以以各种方式把文档加入机密文档的集合。检测系统可以提供一用户接口，管理员通过该用户接口能向所述集合提交机密文档。此外，检测系统可能有一子系统，该子系统可以分析一文档全集，并且检测哪些文档具有机密性告示。例如，指示可以是文档的页脚或页眉上的单词“机密”。

[0023] 本领域的技术人员会理解，除了检测经由电子邮件系统发送的机密信息以外，检测系统可用来检测多种环境下的类似内容。检测系统可用来检测任一类到来或外发通信中的相似内容，比如新闻和讨论组记录、即时消息、电子邮件附件、通讯稿、电子介绍、出版物、由语音通信系统分发的消息、网页等等。在对基于 web 的讨论组记录的情况下，检测系统可以集成有 web 浏览器。检测系统也可以被实现为对通信内容进行适当的解密和加密。检测系统也可以用来标识任一类目标通信，并且不限于电子邮件的机密信息。目标信息可用来监视雇员正在发送哪一类电子邮件。例如，目标信息可以是表示雇员所发送的典型邮件的模板电子邮件的集合，比如日程安排邮件、个人邮件、问题汇报邮件、帮助邮件等等。检测系统可用来检测所访问的网页是否包含不期望的内容。

[0024] 图 1 是说明一实施例中的检测系统的组件的框图。检测系统 100 包括文档存储数据结构 101-103、初始化数据结构组件 111-113 以及检测组件 121-124。文档存储数据结构包括一机密文档存储器 101、关键词索引 102 和句子哈希表 103。检测系统可以在把机密文档置于机密文档存储器内以前处理它们。例如，在电子邮件的情况下，检测系统可以删除“来自：”、“发送至：”和“主题：”信息，并且删除内容的问候语和结束语部分。检测系统也可以以各种方式使其余内容标准化，比如删除大写、调节单词内非字母数字的字符（例如“n*w d*sign”），并且作出其它调节以便抵消发送者想要模糊机密信息的尝试。关键词索引把机密文档的关键词映射到包含那些关键词的机密文档。在一实施例中，关键词索引也可以标识包含该关键词的每个机密文档内的句子。或者，文档存储数据结构也可以包括一关键词 / 关键句索引，该索引把关键词映射到包含那些关键词的机密文档的关键句。句子哈希表把句子（例如关键句）的哈希码映射到包含那些句子的机密文档。初始化数据结构组件包括创建关键词索引组件 111、创建句子哈希表组件 112 和标识关键句组件 113。创建关键词索引组件为机密文档存储器的文档创建了关键词索引。创建关键词索引组件可以基于检索词频率乘以反转文档频率度量来标识关键词。创建句子哈希表组件初始化句子哈希表，以便把关键句的哈希码映射到机密文档内的句子。创建句子哈希表组件调用了标识关键句组件来标识关键句。检测组件包括一检测匹配组件 121，检测匹配组件 121 调用了选

择候选文档组件 122、对完全匹配计数组件 123 和对模糊匹配计数组件 124。检测匹配组件首先调用选择候选文档组件来标识类似于外发通信的候选文档。检测匹配组件接着调用对完全匹配计数组件来确定候选文档的句子是否语外发通信的句子匹配。如果是，则根据匹配程度，查找匹配组件表明外发通信包含机密信息。如果完全匹配的程度不足以表示机密信息，则检测匹配组件可以调用对模糊匹配计数组件来标识外发通信的句子是否与候选文档的句子相似（例如模糊匹配，而不是完全匹配）。如果是，则根据相似性程度，检测匹配组件表明外发通信包含机密信息。本领域的技术人员会理解，检测匹配组件的各种组合可用来实现检测系统。例如，检测系统可以使用选择候选文档组件和对模糊匹配计数组件，而不使用对完全匹配计数组件。检测系统也可以提供匹配程度的等级（例如极可能、高度可能、可能以及不可能），使得可以采取适当的行动（例如通知安全人员并禁止外发通信的发送）。本领域的技术人员会理解，外发通信可能对应于在一组预定的受信者外部发送的任何通信。检测系统也可以把目标通信分隔成几个关注级别（例如极度机密、高度机密以及机密）。检测系统可以定义不同组的受信者，所述受信者被授权接收具有不同关注级别的通信。

[0025] 其上实现检测系统的计算设备可以包括：中央处理单元、内存、输入设备（例如键盘和指示设备）、输出设备（例如显示设备）以及存储设备（例如磁盘驱动器）。内存和存储设备是包含能实现检测系统的指令的计算机可读介质。此外，数据结构和消息结构可以被保存或经由数据传输介质被发送，比如通信链路上的信号。可以使用各种通信链路，比如互联网、局域网、广域网或点对点拨号连接。

[0026] 检测系统可以在各种操作环境中实现，包括个人计算机、服务器计算机、手持或膝上型设备、多处理器系统、基于微处理器的系统、可编程消费者电子设备、网络 PC、小型计算机、大型计算机、包括上述系统和设备的任一个在内的分布式计算环境等等。

[0027] 检测系统可以在计算机可执行指令的一般环境中描述，比如由一台或多台计算机或其它设备执行的程序模块。一般而言，程序模块包括执行特定任务或实现特定的抽象数据类型的例程、程序、对象、组件、数据结构等等。一般而言，程序模块的功能在各个实施例中可以根据需要而组合或分布。

[0028] 图 2 是说明一实施例中、图 1 的检测系统的数据结构的框图。关键词索引 201 和句子哈希表 211 把关键词和句子映射到机密文档存储器（即目标信息存储器）的机密文档 250。关键词索引为机密文档的每个关键词包含一条目 202。每个条目为包含该关键词的每个文档包含一子条目 203。在一实施例中，关键词索引也可以包括一辅助数据结构，该辅助数据结构把关键词映射到包含那些关键词的机密文档的关键句。句子哈希表为每个句子哈希码包含一条目 212。每个条目可以包含子条目 213，该子条目 213 映射到与该句子哈希码相对应的文档内的特定句子。例如，如果两个机密文档包含同一关键句，则这两个句子的句子哈希码会相同。此外，哈希函数可以把两个不同的句子映射到同一哈希码。因此，子条目表示了一系列抵触哈希码。本领域的技术人员会理解，关键词索引和句子哈希码可以用各种数据结构技术来实现，比如数组、二进制树、链表以及哈希表、以及已知表示了检测系统的数据的一个可能逻辑组织的数据结构。

[0029] 图 3 是一实施例中、创建关键词索引的处理的流程图。组件为机密文档的每个词生成一反转文档频率度量，然后使用检索词频率乘以反转文档频率度量来计算每个词对其

文档的重要性。然后，组件选择每个文档中最重要的词作为该文档的关键词，并向关键词索引添加每个关键词的相应条目。在方框 301 中，组件创建了一文档乘单词矩阵，该矩阵表示了每个文档中每个词的数目。组件从该矩阵中导出反转文档频率和检索词频率。在方框 302-304 中，组件循环，为机密文档内每个词计算反转文档频率。组件可以忽视文档中的无用词（例如“和”、“定冠词 the”以及“不定冠词 a”）。在方框 302 中，组件选择机密文档的下一个词。在判决框 303，如果机密文档的全部词都已被选择，则组件继续到方框 305，否则组件继续到方框 304。在方框 304，组件为所选词计算反转文档频率，它是对机密文档数目除以包含所选词的机密文档数目然后取常用对数。然后，组件循环到方框 302 以便选择机密文档的下一个词。在方框 305-311 中，组件循环，选择每个文档并且计算该文档内每个词对该文档的重要性。在方框 305 中，组件选择下一个机密文档。在判决框 306，如果全部机密文档都已被选择，则组件完成，否则组件继续到方框 307。在方框 307，组件选择所选机密文档的下一个词。在判决框 308，如果所选机密文档的全部词都已被选择，则组件循环到方框 305 以选择下一个机密文档，否则组件继续到方框 309。在方框 309，组件计算所选词对所选机密文档的重要性，它是检索词频率（即所选词在所选机密文档内的出现次数）乘以所选词的反转文档频率的乘积。本领域的技术人员会理解，单词对文档的重要性可以以许多不同方式来计算。例如，可以对检索词频率乘以反转文档频率度量进行标准化以便弥补文档内的单词总数。在判决框 310，如果重要性大于一重要性阈值，则组件继续到方框 311，否则组件继续到方框 307 以便选择所选文档的下一个词。在方框 311 中，组件向关键词索引添加一条目，该条目把所选词映射到所选文档。该条目还包含所计算的重要性，该重要性用于确定机密文档的句子是否与外发通信的句子相似。组件接着循环到方框 307 以便选择所选机密文档的下一个关键词。

[0030] 图 4 是说明一实施例中、创建句子哈希表组件的处理的流程图。组件为机密文档的每个关键句向句子哈希表添加一条目。在方框 401 中，组件选择下一个机密文档。在判决框 402 中，如果全部机密文档都已被选择，则组件返回，否则组件继续到方框 403。在方框 403，组件选择所选文档的下一个段落。在判决框 404，如果所选文档的全部段落都已被选择，则组件循环到方框 401 以选择下一个机密文档，否则组件继续到方框 405。在方框 405，组件调用通过所选段落的标识关键句组件。所调用的组件返回所通过段落的关键句的指示。在方框 406，组件调用一哈希函数来为关键句生成一哈希码，然后向句子哈希表为所标识的关键句添加一条目。本领域的技术人员会理解，可以使用各种哈希函数。例如，哈希函数可以从句子的每个关键词的首字母生成一哈希码。组件接着循环到方框 403 以便选择所选文档的下一段落。本领域的技术人员会理解，关键句可以基于它们和所属文档的相似性来导出，而不是逐段地导出。

[0031] 图 5 是说明一实施例中、基于句子关键词和段落关键词之间的相似性进行的标识关键句组件的处理的流程图。组件计算每个句子的关键词和段落关键词之间的相似性。然后，组件选择关键词与段落关键词最为相似的句子作为该段落的关键句。在方框 501 中，组件创建一关键词数组，列出每个关键词在段落内的出现次数。在方框 502 中，组件创建一句子乘关键词矩阵，该矩阵表明每个关键词在段落的每个句子内的出现次数。在方框 503-505，组件循环，计算每个句子和段落的相似性。在方框 503，组件选择段落的下一个句子。在判决框 504，如果全部句子都已被选择，则组件继续到方框 506，否则组件继续到方框

505。在方框 505，组件计算所选句子和段落的相似性。在一实施例中，组件可以把相似性计算为：由矩阵和数组所表示的所选句子和段落间共有的关键词重要性的乘积之和。然后组件循环到方框 503 以选择段落的下一个句子。在方框 506 中，组件选择与段落有最高相似性的句子。然后组件返回。在一实施例中，组件可以标识一段落的多个关键句。在标识了第一关键句后，组件可以从段落的关键词中删除该关键句的关键词、对其余句子重复相似性计算、然后选择在那些相似性中有最高相似性的句子作为另一关键句。组件可以重复该过程，直到标识了期望数量的关键句为止。

[0032] 图 6 是说明一实施例中、基于反转句子频率进行的标识关键句组件的处理的流程图。图 5 和 6 因此表示了标识关键句的可选方式。本领域的技术人员会理解，可以使用任一种方式或所述方式的组合来标识关键句。组件为每个关键词计算反转句子频率。然后，组件为每个句子的每个词计算一重要性，比如检索词频率乘以反转句子频率度量。然后，组件通过把句子关键词的重要性相加来计算句子的重要性。具有最高重要性的句子被视为段落的关键句。在方框 601 中，组件创建一句子乘关键词矩阵。在方框 602-602 中，组件循环，选择段落的关键词并且计算它们的反转句子频率。在方框 602 中，组件选择段落的下一关键词。在判决框 603，如果段落的全部关键词都已被选择，则组件继续到方框 605，否则组件继续到方框 604。在方框 604，组件把所选关键词的反转句子频率计算为：段落内句子数除以包含所选关键词的段落内句子数然后取常用对数。在方框 605-610 中，组件循环，计算每个句子对段落的重要性。在方框 605，组件选择段落的下一个句子。在判决框 606，如果全部句子都已被选择，则组件继续到方框 611，否则组件继续到方框 607。在方框 607 中，组件选择所选句子的下一个关键词。在判决框 608 中，如果所选句子的全部关键词都已被选择，则组件继续到方框 610，否则组件继续到方框 609。在方框 609 中，组件把所选关键词对所选句子的重要性计算为：关键词在句子内的出现次数乘以句子的反转句子频率。然后，组件循环到方框 607 以选择所选句子的下一个关键词。在方框 610 中，组件通过把所选句子的关键词重要性除以所选句子中关键词数目（即平均关键词重要性）相加，从而计算所选句子对段落的重要性。然后组件循环到方框 605 以便选择下一个句子。在方框 611 中，组件把具有最高重要性的句子选择作为关键句，然后返回。

[0033] 图 7 是说明一实施例中、检测匹配组件的处理的流程图。在方框 701 中，组件调用选择候选文档组件来标识匹配的候选文档。在方框 702 中，组件调用对完全匹配计数组件来标识外发通信的句子和机密文档的句子之间完全匹配的程度。在判决框 703 中，如果完全匹配程度超过一阈值，则组件返回已经检测到完全匹配的指示，否则组件继续到方框 704。在方框 704 中，组件调用对模糊匹配计数组件来标识机密文档的句子和外发通信的句子之间的模糊匹配程度。在判决框 705，如果模糊匹配程度超过一阈值，则组件返回已经发现模糊匹配的指示，否则组件返回没有发现匹配的指示。

[0034] 图 8 是说明一实施例中、选择候选文档组件的处理的流程图。组件标识外发通信的关键词，然后标识与候选文档有相似关键词的文档。在方框 801 中，组件创建由外发通信的单词组成的单词数组。在方框 802-804，组件循环，计算外发通信单词的重要性。在方框 802，组件选择外发通信的下一个词。在判决框 803，如果外发通信的全部词都已被选择，则组件继续到方框 805，否则组件继续到方框 804。在方框 804，组件使用一检索词频率乘以反转文档频率度量来计算所选词的重要性，并且循环到方框 802 以选择下一个词。反转文

档频率可以表示机密文档内的反转文档频率。在方框 805-809，组件循环，选择每一个机密文档，并且计算它和外发通信的相似性。在方框 805，组件选择下一个机密文档。在判决框 806，如果全部机密文档都已被选择，则组件完成，否则组件继续到方框 807。在方框 807，组件通过把在机密文档和外发通信间共同的关键词的重要性乘积相加，从而计算所选机密文档和外发通信的相似性。在判决框 808，如果相似性超过一相似性阈值，则组件继续到方框 809，否则组件循环到方框 805 以选择下一个机密文档。在方框 809，组件把所选的文档选择作为候选文档，然后循环到方框 805 以选择下一个机密文档。

[0035] 图 9 是说明一实施例中、对完全匹配计数组件的处理的流程图。组件对外发通信的句子和候选文档内句子相匹配的次数进行计数。在方框 901，组件选择外发通信的下一个段落。在判决框 902，如果全部段落都已被选择，则组件返回，否则组件继续到方框 903。在方框 903，组件调用一标识关键句组件来标识所选段落的关键句。在方框 904 中，组件调用一哈希函数来为关键句生成一哈希码。然后，组件检验句子哈希表的每个被哈希的条目的每个子条目，以确定关键句是否与候选文档的句子相匹配。在判决框 905，如果发现匹配，则组件继续到方框 906，否则组件循环到方框 901 以选择外发通信的下一个段落。在方框 906 中，组件把外发通信的匹配计数递增所发现匹配的数目。然后，组件循环回方框 901 以选择外发通信的下一个段落。

[0036] 图 10 是说明一实施例中、对模糊匹配计数组件的处理的流程图。在方框 1001 中，组件选择外发通信的下一个段落。在判决框 1002，如果全部段落都已被选择，则组件返回，否则组件继续到方框 1003。在方框 1003，组件选择下一个候选文档。在判决框 1004，如果全部候选文档都已被选择，则组件循环到方框 1001 以选择外发通信的下一个段落，否则组件继续到方框 1005。在方框 1005，组件选择所选候选文档的下一个关键句。在方框 1006 中，组件计算所选句子间的余弦相似性或编辑距离。在判决框 1007，如果相似性或距离超过一阈值，则组件继续到方框 1008，否则组件循环到方框 1003 以选择下一个候选文档。在方框 1008，组件把对外发通信的相似性计数递增，然后循环到方框 1003 以选择下一个候选文档。

[0037] 本领域的技术人员会理解，尽管这里为了说明而描述了检测系统的特定实施例，然而可以作出各种修改而不背离本发明的精神和范围。因而，本发明仅受所附权利要求的限制。

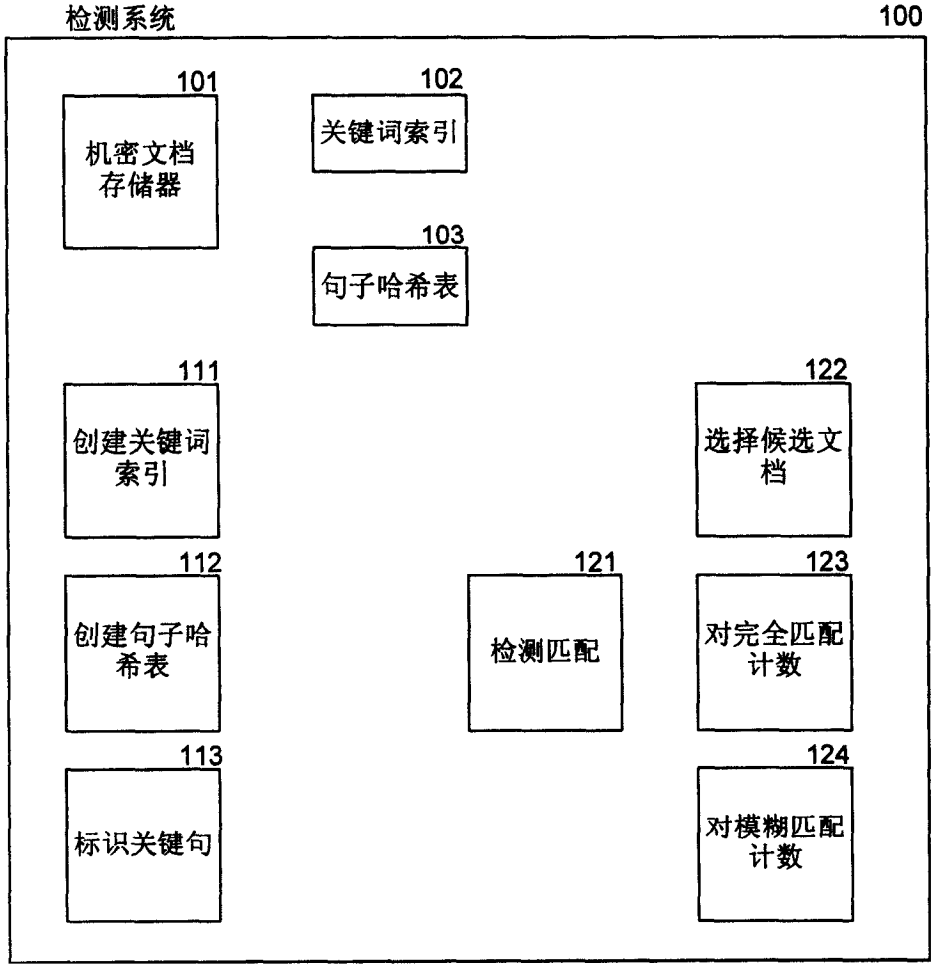

检测系统
100
101
机密文档存储器
102
关键词索引
103
句子哈希表
111
创建关键词索引
112
创建句子哈希表
113
标识关键句
121
检测匹配
122
选择候选文档
123
对完全匹配计数
124
对模糊匹配计数

图 1

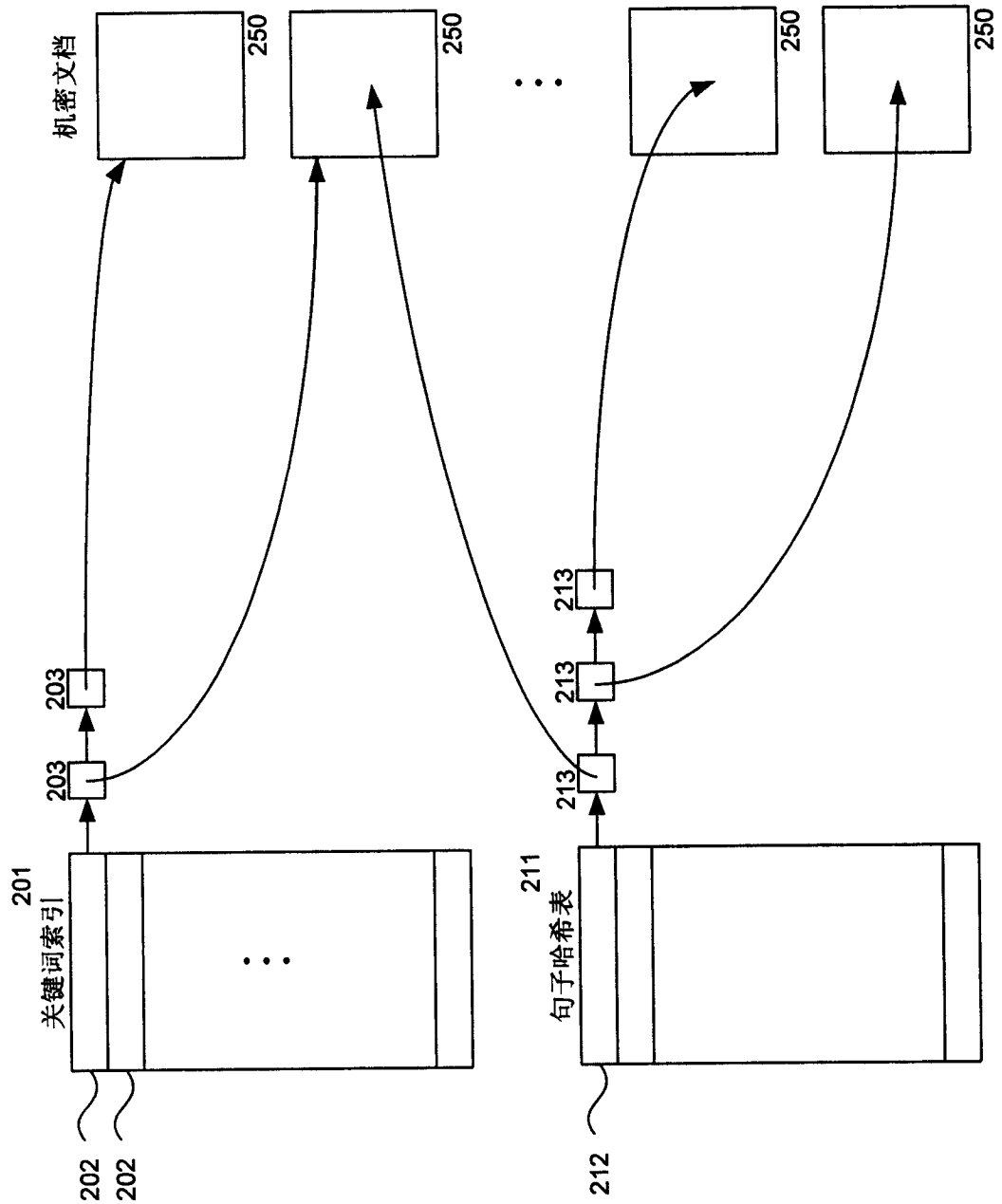

机密文档
250
250
250
250
201
关键词索引
202
202
203
203
211
句子哈希表
212
213
213
213

图 2

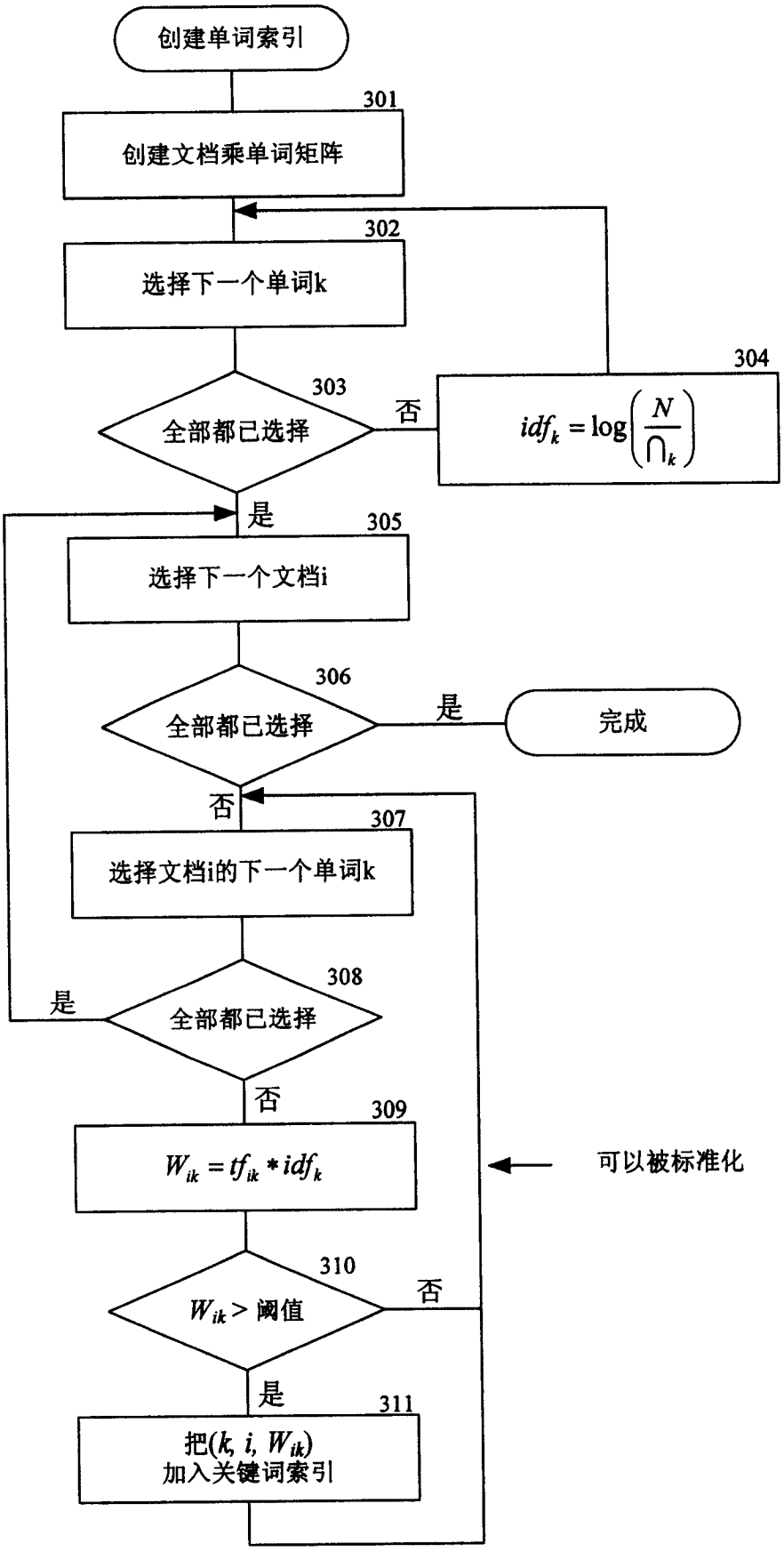

创建单词索引
301
创建文档乘单词矩阵
302
选择下一个单词k
303
全部都已选择
否
304
$idf_k = \log\left(\frac{N}{\cap_k}\right)$
是
305
选择下一个文档i
306
全部都已选择
是
完成
否
307
选择文档i的下一个单词k
308
全部都已选择
是
否
309
$W_{ik} = tf_{ik} * idf_k$
可以被标准化
310
W_{ik} > 阈值
否
是
311
把(k, i, W_{ik})
加入关键词索引

图 3

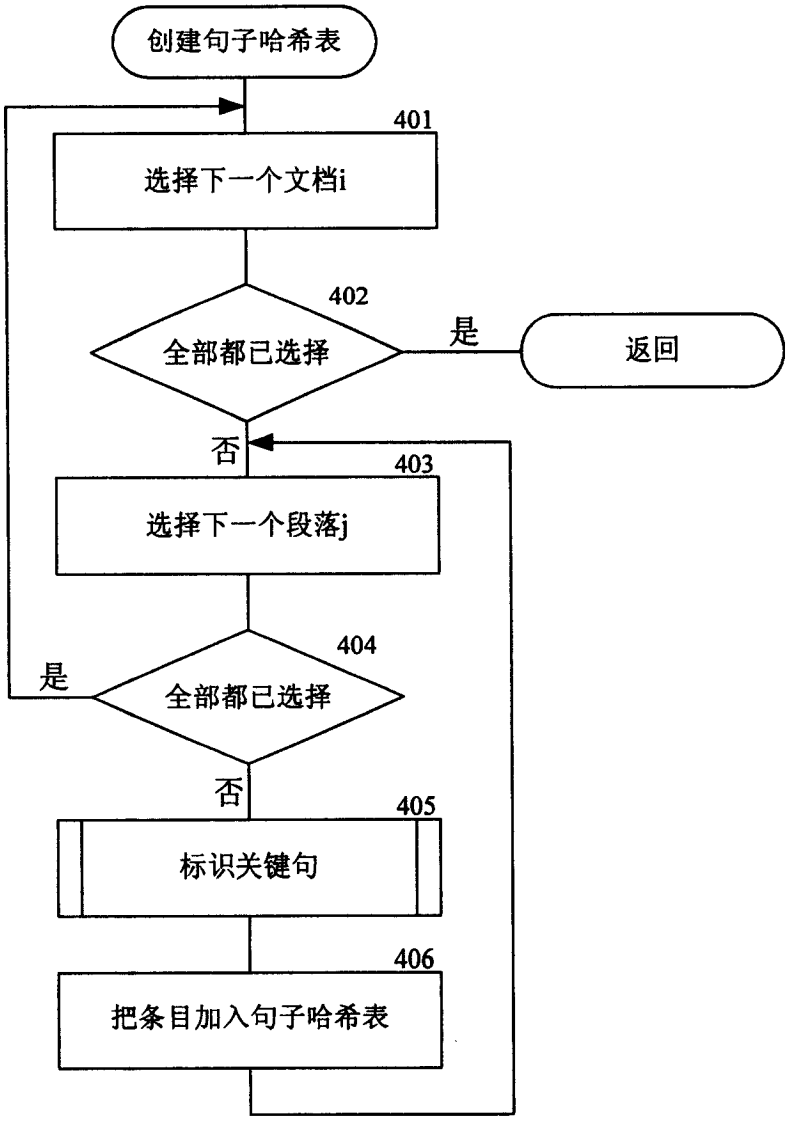

创建句子哈希表
401
选择下一个文档i
402
全部都已选择
是
返回
否
403
选择下一个段落j
404
全部都已选择
是
否
405
标识关键句
406
把条目加入句子哈希表

图 4

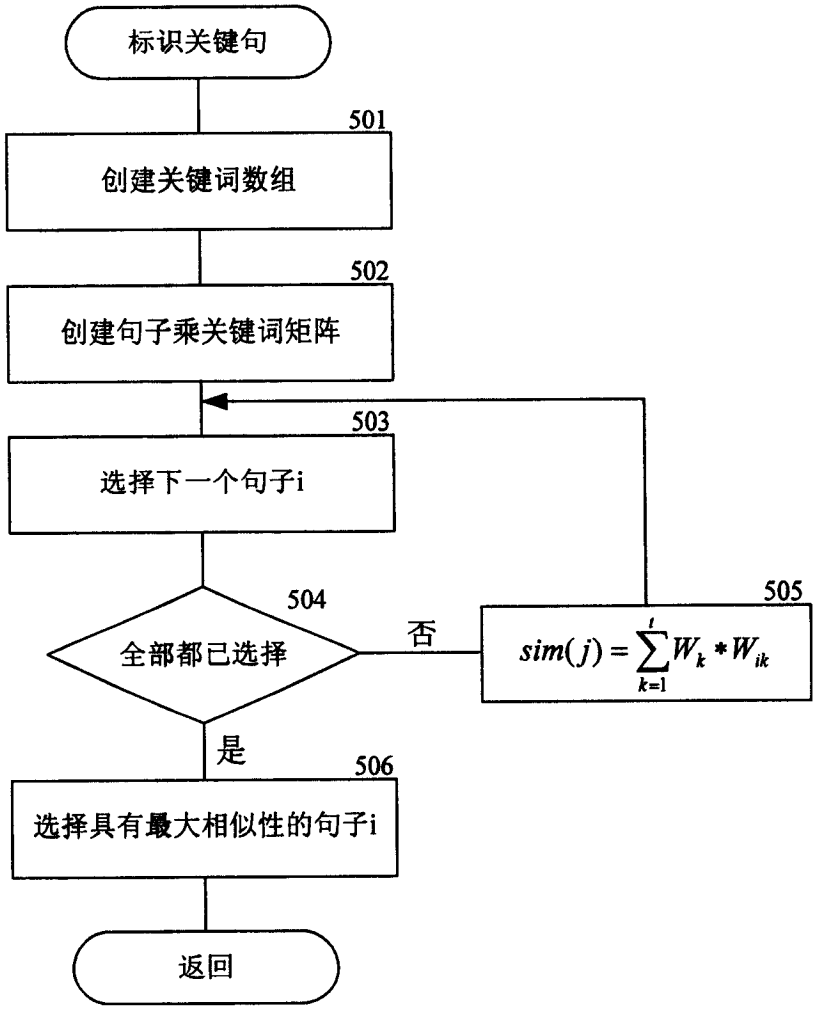

标识关键句
501
创建关键词数组
502
创建句子乘关键词矩阵
503
选择下一个句子i
504
全部都已选择
否
505
$sim(j)=\sum_{k=1}^{t}W_k * W_{ik}$
是
506
选择具有最大相似性的句子i
返回

图 5

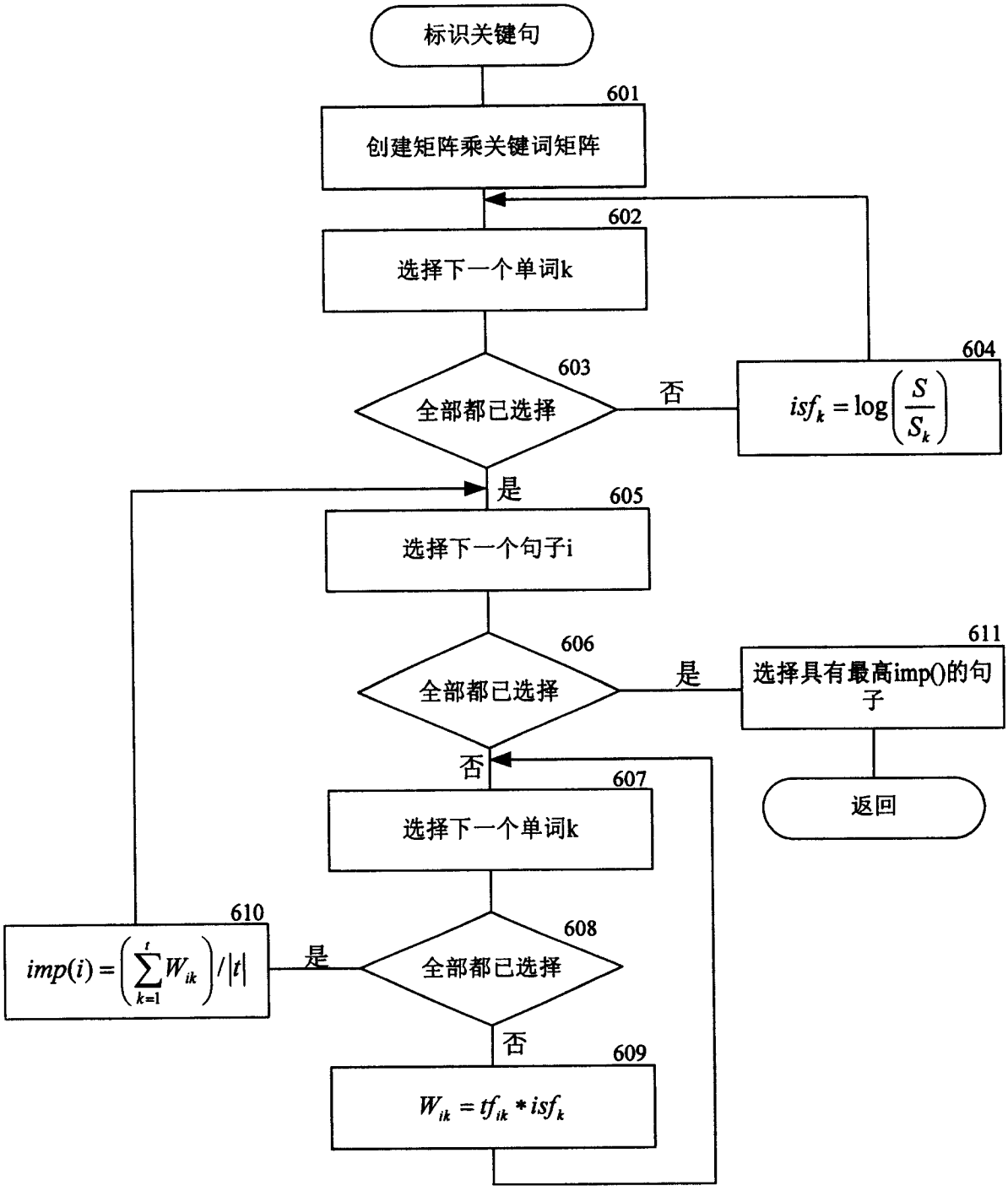

标识关键句
601
创建矩阵乘关键词矩阵
602
选择下一个单词k
603
全部都已选择
否
604
$isf_k = \log\left(\frac{S}{S_k}\right)$
是
605
选择下一个句子i
606
全部都已选择
是
611
选择具有最高imp()的句子
返回
否
607
选择下一个单词k
610
$imp(i) = \left(\sum_{k=1}^{t} W_{ik}\right) / |t|$
是
608
全部都已选择
否
609
$W_{ik} = tf_{ik} * isf_k$

图 6

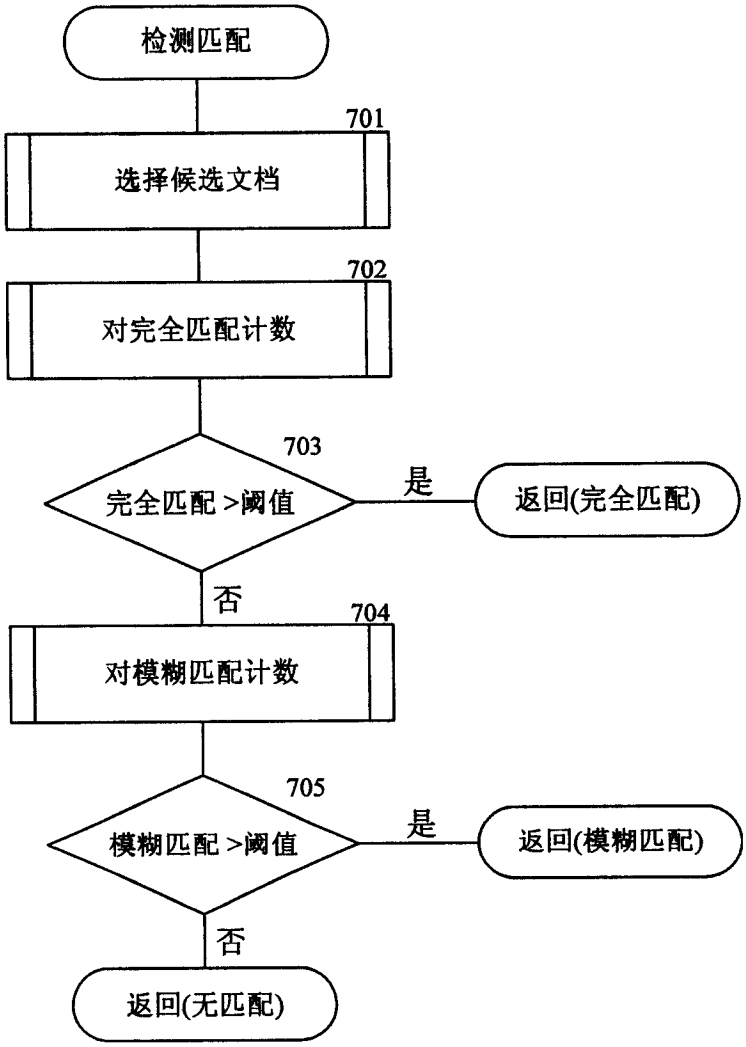

检测匹配
701
选择候选文档
702
对完全匹配计数
703
完全匹配 >阈值
是
返回(完全匹配)
否
704
对模糊匹配计数
705
模糊匹配 >阈值
是
返回(模糊匹配)
否
返回(无匹配)

图 7

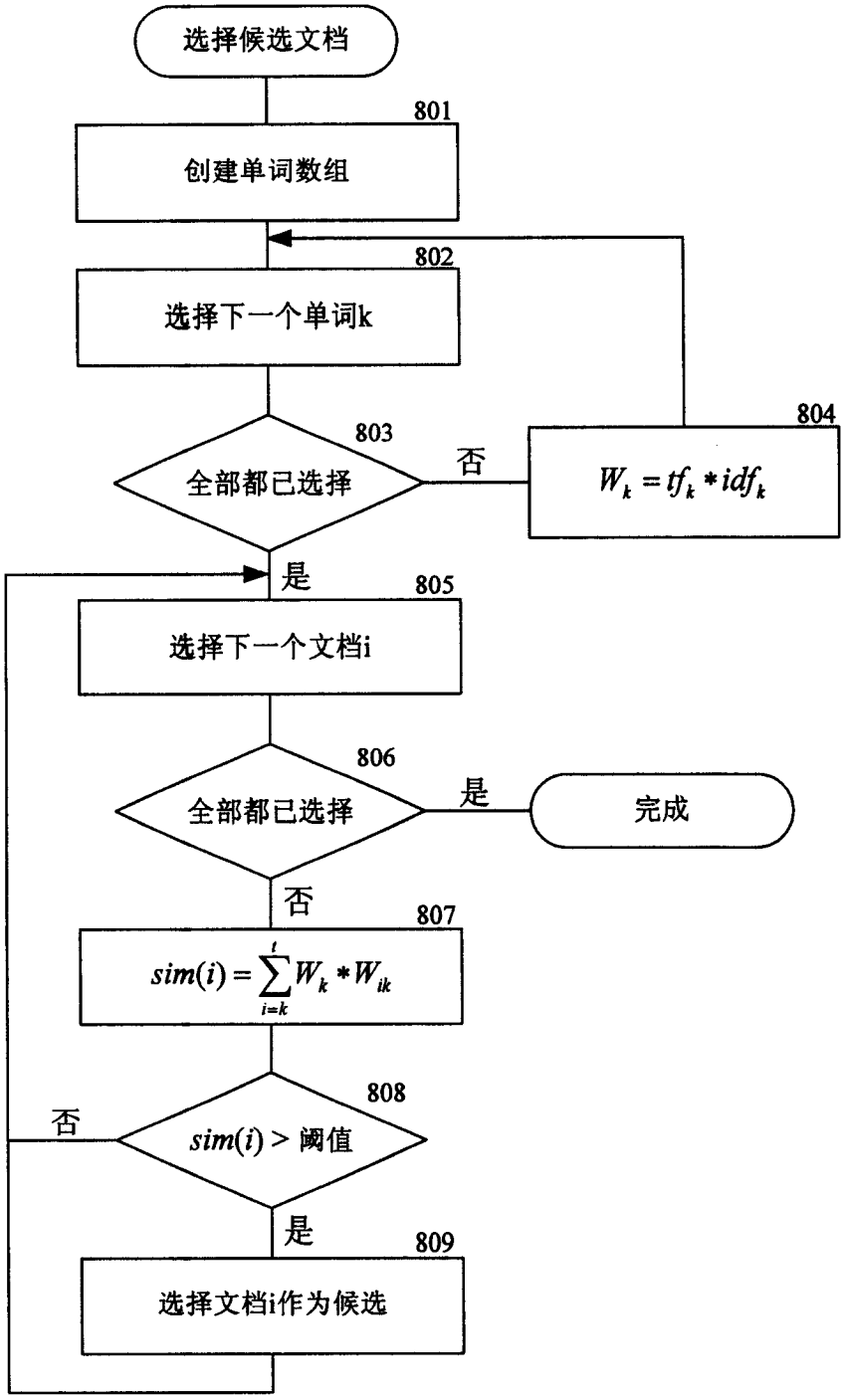

选择候选文档
801
创建单词数组
802
选择下一个单词k
803
全部都已选择
否
804
$W_k = tf_k * idf_k$
是
805
选择下一个文档i
806
全部都已选择
是
完成
否
807
$sim(i) = \sum_{i=k}^{t} W_k * W_{ik}$
808
$sim(i) >$ 阈值
否
是
809
选择文档i作为候选

图 8

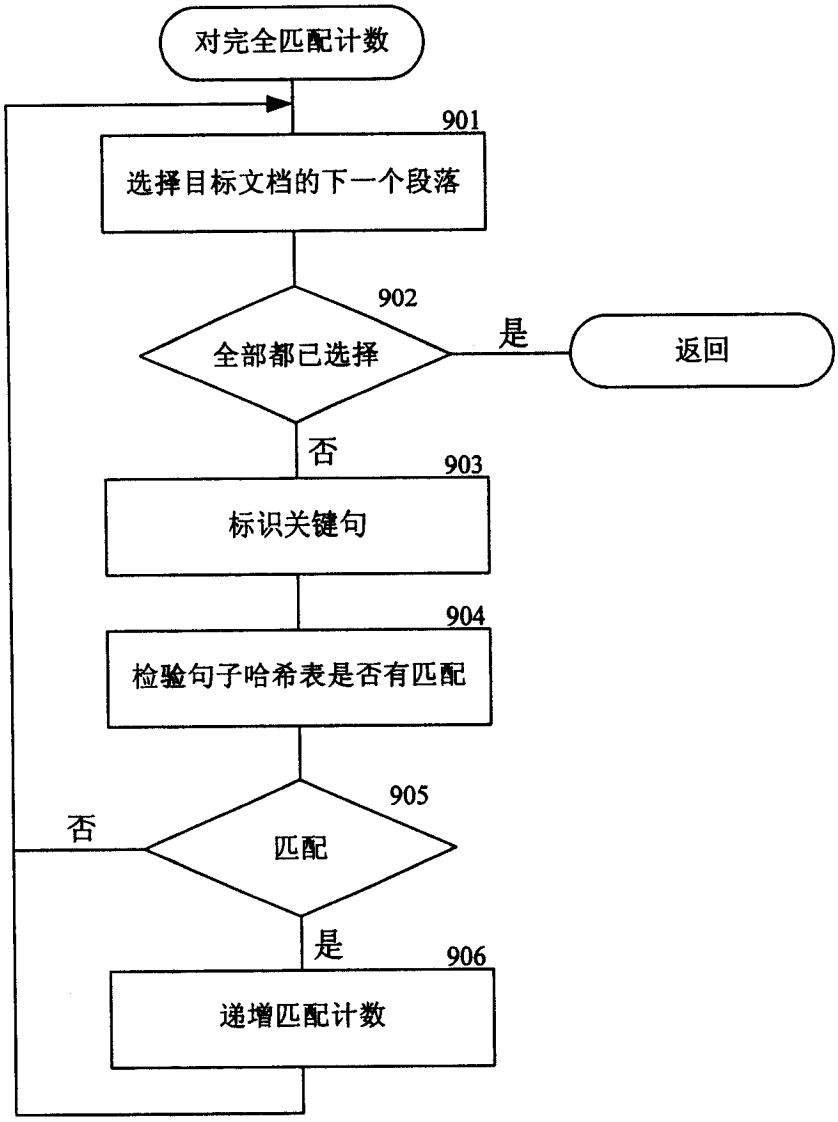

对完全匹配计数
901
选择目标文档的下一个段落
902
全部都已选择
是
返回
否
903
标识关键句
904
检验句子哈希表是否有匹配
905
匹配
否
是
906
递增匹配计数

图 9

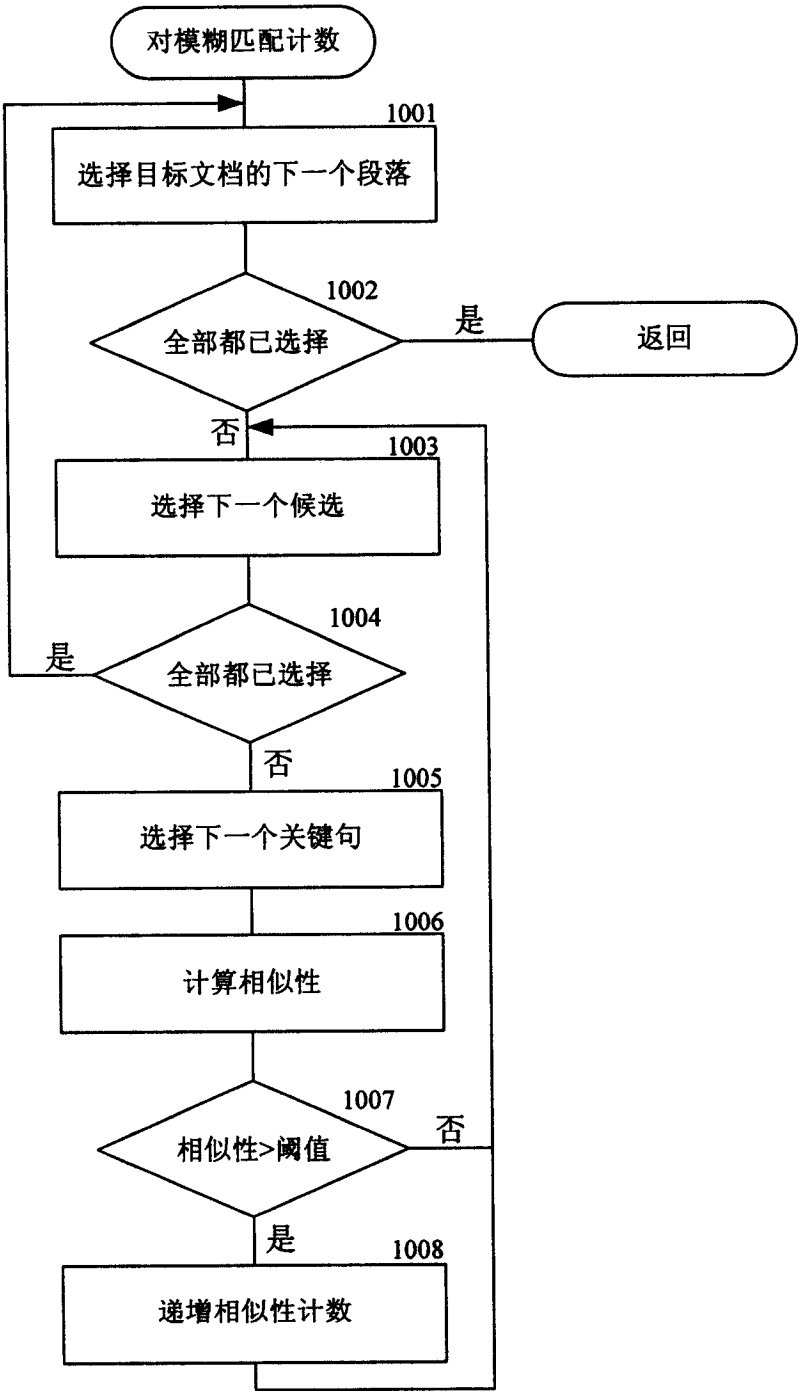

对模糊匹配计数
1001
选择目标文档的下一个段落
1002
全部都已选择
是
返回
否
1003
选择下一个候选
1004
全部都已选择
是
否
1005
选择下一个关键句
1006
计算相似性
1007
相似性>阈值
否
是
1008
递增相似性计数

图 10