



(12) 发明专利

(10) 授权公告号 CN 101084500 B

(45) 授权公告日 2010.05.12

(21) 申请号 200580043755.9

(22) 申请日 2005.05.10

(30) 优先权数据

10/969,488 2004.10.20 US

(85) PCT申请进入国家阶段日

2007.06.20

(86) PCT申请的申请数据

PCT/US2005/017492 2005.05.10

(87) PCT申请的公布数据

W02006/043988 EN 2006.04.27

(73) 专利权人 甲骨文国际公司

地址 美国加利福尼亚州

(72) 发明人 理查德·C·约翰逊

(74) 专利代理机构 北京东方亿思知识产权代理
有限责任公司 11258

代理人 王怡

(51) Int. Cl.

G06F 17/30(2006.01)

G06F 17/28(2006.01)

(56) 对比文件

W0 00/79436 A2, 2000.12.28, 全文.

W0 82/00442 A, 1982.02.18, 说明书第1页
第1行至第25行, 第2页第33行至第6页第32
行, 第7页第19行至第30行, 第12页第10行至
第14页第6行, 图1-4.

CN 1192003 A, 1998.09.02, 全文.

审查员 孙芳

权利要求书 5 页 说明书 14 页 附图 9 页

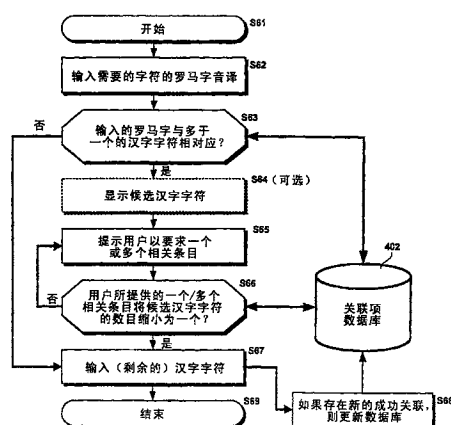
(54) 发明名称

用于输入和搜索非罗马字母表字符的计算机
实现的方法和系统以及相关搜索系统

(57) 摘要

一种计算机实现的方法, 用于从一组非罗马字符或对象中选择需要的罗马或非罗马字母表字符或对象。所述方法可以包括下列步骤: 提供关联项数据库, 该关联项数据库对于所述组中的每个非罗马字母表字符都包括与所述每个非罗马字母表字符相关联的罗马字母表或其他语音音译和与所述每个非罗马字母表字符相关联的多个条目; 接收需要的非罗马字母表字符的语音音译或数据对象和与需要的非罗马字母表字符或其它类似符号输入相关联的至少一个相关条目; 访问关联项数据库并且将所述组中那些与所接收的语音音译和所接收的至少一个相关条目相关联的字符识别为候选字符; 如果候选字符的数目大于1, 则接收另外的相关条目并且重复访问和识别步骤, 直到多个候选字符被缩小为单个候选字符; 以及提供所述单个候选字符作为需要的非罗马字母表字符。而且, 源于上面所描述的原理, 本发明包括各种用于通过使用关联项和为正在搜索的(一个或多个)项目提供环境的其它装置而改善搜索引

擎效率的方法。



1. 一种计算机实现的方法,用于从一组非罗马字符中选择需要的非罗马字母表字符,该方法包括下列步骤:

为每个用户提供个体关联项数据库,该关联项数据库对于所述组中的每个非罗马字母表字符都包括与每个所述非罗马字母表字符相关联的罗马字母表语音音译和与每个所述非罗马字母表字符相关联的多个词或短语;

将每个用户的个体关联项数据库个性化,以使得所述关联项数据库对于每个用户都包含从每个所述用户所接收的相关词或短语,从每个所述用户所接收的相关词或短语中的至少一个是与需要的字符的字典含义无关的依用户而定的关联项;

接收所述需要的非罗马字母表字符的语音音译和至少一个与所述需要的非罗马字母表字符相关联的相关词或短语;

访问所述关联项数据库并且将所述组中那些与所接收的语音音译和至少一个所接收的相关词或短语相关联的字符识别为候选字符;

如果候选字符的数目大于1,则接收另外的相关词或短语并且重复所述访问和识别步骤,直到多个候选字符被缩小为单个候选字符;以及

提供所述单个候选字符作为所述需要的非罗马字母表字符。

2. 如权利要求1所述的计算机实现的方法,其中所述相关词或短语中的至少一个与所述需要的字符的含义无关。

3. 如权利要求1所述的计算机实现的方法,其中所述关联项数据库中的相关词或短语中的至少一个是由用户提供的。

4. 如权利要求1所述的计算机实现的方法,其中如果接收的相关词或短语当前不与所述需要的字符相关联,则执行将所接收的相关词或短语与所述需要的字符相关联并且将所接收的相关词或短语存储在所述关联项数据库中的步骤。

5. 如权利要求1所述的计算机实现的方法,其中如果接收的相关词或短语当前不在所述关联项数据库中,则执行将所接收的相关词或短语增加到所述关联项数据库中的步骤。

6. 如权利要求1所述的计算机实现的方法,其中所述提供步骤是在所述一组非罗马字符包括日语字符的情况下执行的。

7. 如权利要求1所述的计算机实现的方法,其中所述提供步骤是在所述一组非罗马字符包括汉语字符的情况下执行的。

8. 如权利要求1所述的计算机实现的方法,其中所述提供步骤是在所述一组非罗马字符包括朝鲜语字符的情况下执行的。

9. 如权利要求1所述的计算机实现的方法,其中所述提供步骤是在所述一组非罗马字符包括希伯来语字符的情况下执行的。

10. 如权利要求1所述的计算机实现的方法,其中所述提供步骤是在所述一组非罗马字符包括阿拉伯语字符的情况下执行的。

11. 如权利要求1所述的计算机实现的方法,其中所述提供步骤是在所述关联项数据库包括具有多行的关联项表的情况下执行的,每一行包括所述组中的单个非罗马字母表字符、所述单个非罗马字母表字符的罗马字母表语音音译和与所述单个非罗马字母表字符相关联的多个词或短语。

12. 如权利要求1所述的计算机实现的方法,其中所述接收步骤是通过到计算机的语

音输入而执行的,该计算机耦合到所述关联项数据库。

13. 如权利要求 1 所述的计算机实现的方法,其中所述接收步骤是通过到计算机的键盘输入而执行的,该计算机耦合到所述关联项数据库。

14. 如权利要求 1 所述的计算机实现的方法,其中所述接收步骤是通过到计算机的机器视觉而执行的,该计算机耦合到所述关联项数据库。

15. 如权利要求 1 所述的计算机实现的方法,还包括排列所述关联项数据库中的相关词或短语的步骤,以使得那些所接收的在所述访问和识别步骤的较少迭代中引导至所述单个候选字符的相关词或短语的级别高于在所述访问和识别步骤的比较多的迭代中引导至所述单个候选字符的相关词或短语。

16. 如权利要求 1 所述的计算机实现的方法,还包括搜索计算机网络以查找包括所提供的非罗马字母表字符的文档。

17. 如权利要求 1 所述的计算机实现的方法,其中所述接收步骤是在所述相关词或短语与所述需要的非罗马字母表字符的语言相同的情况下执行的。

18. 如权利要求 1 所述的计算机实现的方法,其中所述接收步骤是在所述相关词或短语与所述需要的非罗马字母表字符的语言不同的情况下执行的。

19. 如权利要求 1 所述的计算机实现的方法,还包括接连返回到所述语音音译接收步骤以选择多个非罗马字母表字符。

20. 一种计算机系统,用于从一组非罗马字符中选择需要的非罗马字母表字符,所述计算机系统包括:

第一提供装置,用于为每个用户提供个体关联项数据库,该关联项数据库对于所述组中的每个非罗马字母表字符都包括与每个所述非罗马字母表字符相关联的罗马字母表语音音译和与每个所述非罗马字母表字符相关联的多个词或短语;

用于将每个用户的个体关联项数据库个性化的装置,以使得所述关联项数据库对于每个用户都包含从每个所述用户所接收的相关词或短语,从每个所述用户所接收的相关词或短语中的至少一个是与需要的字符的字典含义无关的依用户而定的关联项;

第一接收装置,用于接收所述需要的非罗马字母表字符的语音音译和至少一个与所述需要的非罗马字母表字符相关联的相关词或短语;

访问和识别装置,用于访问所述关联项数据库并且将所述组中那些与所接收的语音音译和至少一个所接收的相关词或短语相关联的字符识别为候选字符;

第二接收装置,用于如果候选字符的数目大于 1,则接收另外的相关词或短语并且重复所述访问和识别装置所执行的步骤,直到多个候选字符被缩小为单个候选字符;以及

第二提供装置,用于提供所述单个候选字符作为所述需要的非罗马字母表字符。

21. 如权利要求 20 所述的计算机系统,其中所述相关词或短语中的至少一个与所述需要的字符的含义无关。

22. 如权利要求 20 所述的计算机系统,其中所述关联项数据库中的相关词或短语中的至少一个是由用户提供的。

23. 如权利要求 20 所述的计算机系统,还包括用于如果接收的相关词或短语当前不与所述需要的字符相关联,则将所接收的相关词或短语与所述需要的字符相关联并且将所接收的相关词或短语存储在所述关联项数据库中的装置。

24. 如权利要求 20 所述的计算机系统,还包括用于如果接收的相关词或短语当前不在所述关联项数据库中,则将所接收的相关词或短语增加到所述关联项数据库中的装置。

25. 如权利要求 20 所述的计算机系统,其中所述第一提供装置在所述关联项数据库包括具有多行的关联项表的情况下执行提供,每一行包括所述组中的单个非罗马字母表字符、所述单个非罗马字母表字符的罗马字母表语音音译和与所述单个非罗马字母表字符相关联的多个词或短语。

26. 如权利要求 20 所述的计算机系统,其中所述第一和第二接收装置通过到计算机的语音输入而执行提供,该计算机耦合到所述关联项数据库。

27. 如权利要求 20 所述的计算机系统,其中所述第一和第二接收装置通过到计算机的键盘输入而执行提供,该计算机耦合到所述关联项数据库。

28. 如权利要求 20 所述的计算机系统,其中所述第一和第二接收装置通过到计算机的机器视觉而执行提供,该计算机耦合到所述关联项数据库。

29. 如权利要求 20 所述的计算机系统,还包括用于排列所述关联项数据库中的相关词或短语的装置,以使得那些所接收的在所述访问和识别装置所执行的步骤的较少迭代中引导至所述单个候选字符的相关词或短语的级别高于在所述访问和识别装置所执行的步骤的较多的迭代中引导至所述单个候选字符的相关词或短语。

30. 如权利要求 20 所述的计算机系统,还包括用于搜索计算机网络以查找包括所提供的非罗马字母表字符的文档的装置。

31. 如权利要求 20 所述的计算机系统,其中所述第一接收装置在所述相关词或短语与所述需要的非罗马字母表字符的语言相同的情况下执行接收。

32. 如权利要求 20 所述的计算机系统,其中所述第一接收装置在所述相关词或短语与所述需要的非罗马字母表字符的语言不同的情况下执行接收。

33. 如权利要求 20 所述的计算机系统,还包括用于接连返回到由所述第一接收装置执行的语音音译接收步骤以选择多个非罗马字母表字符的装置。

34. 一种计算机实现的方法,用于搜索包括用户未知的一个或多个预定非罗马字母表字符的文档,该方法包括下列步骤:

提供需要的非罗马字母表字符的发音;

提供用户将其与所述需要的非罗马字母表字符相关联的至少一个依用户而定的相关词或短语;

使用户登入;

从关联项数据库中取得一个或多个候选字符,所述候选字符的发音与所提供的发音匹配并且与所述至少一个相关词或短语中的一个或多个相关联,所述登入步骤使得所述取得步骤从为登入用户个性化的个体关联项数据库中取得所述至少一个候选字符,该个体个性化的关联项数据库对于每个用户都包含从每个所述用户接收的相关词或短语,从每个所述用户接收的相关词或短语中的至少一个是与需要的字符的字典含义无关的依用户而定的关联项;

如果候选字符的数目大于 1,则提供另外的相关词或短语并且重复所述取得步骤,直到单个候选字符被取得;以及

执行搜索以查找包括所述单个候选字符的文档。

35. 如权利要求 34 所述的计算机实现的方法,其中在所述发音提供步骤中,所述发音是语音发音。

36. 如权利要求 34 所述的计算机实现的方法,其中在所述相关词或短语提供步骤中,所述相关词或短语中的至少一个与所述预定字符的含义无关。

37. 如权利要求 34 所述的计算机实现的方法,其中如果提供的相关词或短语当前不与所述预定字符相关联,则还执行将所提供的相关词或短语与所述预定字符相关联并且将所接收的相关词或短语存储在所述关联项数据库中的步骤。

38. 如权利要求 34 所述的计算机实现的方法,其中如果提供的相关词或短语当前不在所述关联项数据库中,则执行将所提供的相关词或短语增加到所述关联项数据库中的步骤。

39. 如权利要求 34 所述的计算机实现的方法,其中所述发音提供步骤是在所述非罗马字符是日语字符的情况下执行的。

40. 如权利要求 34 所述的计算机实现的方法,其中所述发音提供步骤是在所述非罗马字符是汉语字符的情况下执行的。

41. 如权利要求 34 所述的计算机实现的方法,其中所述发音提供步骤是在所述非罗马字符是朝鲜语字符的情况下执行的。

42. 如权利要求 34 所述的计算机实现的方法,其中所述发音提供步骤是在所述非罗马字符是希伯来语字符的情况下执行的。

43. 如权利要求 34 所述的计算机实现的方法,其中所述发音提供步骤是在所述非罗马字符是阿拉伯语字符的情况下执行的。

44. 如权利要求 34 所述的计算机实现的方法,其中所述取得步骤是在所述关联项数据库包括具有多行的关联项表的情况下执行的,每一行包括单个不同的非罗马字母表字符、所述单个非罗马字母表字符的罗马字母表发音和与所述单个非罗马字母表字符相关联的多个词或短语。

45. 如权利要求 34 所述的计算机实现的方法,其中所述提供步骤是通过到计算机的语音输入而执行的,该计算机耦合到所述关联项数据库。

46. 如权利要求 34 所述的计算机实现的方法,其中所述提供步骤是通过到计算机的键盘输入而执行的,该计算机耦合到所述关联项数据库。

47. 如权利要求 34 所述的计算机实现的方法,其中所述提供步骤是通过到计算机的机器视觉而执行的,该计算机耦合到所述关联项数据库。

48. 如权利要求 34 所述的计算机实现的方法,其中所述搜索步骤促使搜索计算机网络以查找所述文档。

49. 如权利要求 34 所述的计算机实现的方法,其中所述相关词或短语提供步骤是在所述相关词或短语与所述预定非罗马字母表字符的语言相同的情况下执行的。

50. 如权利要求 34 所述的计算机实现的方法,其中所述相关词或短语提供步骤是在所述相关词或短语与所述预定非罗马字母表字符的语言不同的情况下执行的。

51. 如权利要求 34 所述的计算机实现的方法,还包括接连返回到所述发音提供步骤以取得用于另外的预定非罗马字母表字符的另外的候选字符。

52. 如权利要求 34 所述的计算机实现的方法,其中所述发音提供步骤中的发音被配置

为使用语音字母表。

53. 如权利要求 34 所述的计算机实现的方法,其中所述取得步骤被配置为促使取得罗马字符、非罗马字符、图片、音乐和任何计算机可读文件中的至少一个。

用于输入和搜索非罗马字母表字符的计算机实现的方法和系统以及相关搜索系统

技术领域

[0001] 本发明涉及针对文档和网络的用于输入汉语和日语字符的计算机实现的方法和系统以及其它数据对象、图像和符号对象或概念的搜索。

背景技术

[0002] 与输入日语和汉语字符到计算机中相关的基本问题仅仅是不能使键盘足够大以便容纳一个性将会在日语报纸中发现的数千个字符,更不用说正式的汉语技术文档或者政府文档所需要的大约 40,000 个字符了。相同的问题也出现在其它语言中。另外,基于字符的(汉语的、日语的和一些朝鲜语的)网站的数目上的增长使得在没有足够的方法来输入这样的字符的情况下很难搜索它们。此外,甚至连获得字符并对其进行搜索也不意味着进行搜索的字符的具体实例被分离出来了。诸如阿拉伯语或者希伯来语之类的其它语言具有不适于通过当前方法手动输入的字符或句法和书写风格。虽然存在次优的(suboptimal)方案,但是没有一个能够很好地起作用。虽然传统的搜索系统可以返回需要的结果,但是它们所返回的无用的结果总是它们所返回的需要的结果的很多倍。

[0003] 日语支持语音输入,但是存在关于语音缺陷或者结巴、口音、发音、错误(一个词或字符替代另一个)、方言、第二种或非母语语言的问题。大多数当前的语音识别系统需要很长的训练期以使得机器能够精确地转录用户的语音。而且,经常需要训练用户来练习合适的措辞以使得机器能够工作在可接受的识别水平。

[0004] 首先来看日语的书面形式,最古老的普通方法涉及用户输入需要的字符(或者短语,意思是一串字符)的通过键盘输入的(罗马字或者使用罗马字母表或假名(日语语音字符)的日语字符的语音音译(phonetic transliteration))语音表示。所有相似发音的字符将显示在屏幕上,然后要求用户从出现在显示器上的相似发音的字符中进行选择。该过程很慢、很乏味,而且不是总能产生需要的字符。

[0005] 新近的方案要求用户口述日语或汉语口语,用软件/机器将口语词翻译成字符。然后用户必须编辑结果或者重复语音或者确认结果。然而,使用语音的当前方法被认为缺少足够的速度、精度和对于所有可能的用户的适合性。同样,单独依赖键盘方法来输入罗马字发音的传统方法被认为很慢、麻烦并且反直觉。最后,这些传统方法经常通过要求在很多候选字符之中选择一个或者通过要求用户重复他或她的语音而打断用户的思路。经常,用户可能未识别出候选字符,因此不能从候选字符中选择合适的字符。而且,这些方法经常不能使用户以及及时的方式可靠地选择需要的字符。

[0006] 在想要输入非罗马字母表字符到网络搜索引擎中以搜索包含这样的字符的网站时也出现这些缺点。因此,也需要使得用户能够容易地输入非罗马字母表字符到搜索引擎中并且对所输入的字符进行搜索的系统和方法。

发明内容

[0007] 根据一实施例,本发明是一种计算机实现的方法,用于从一组非罗马字符中选择需要的非罗马字母表字符。词“关联项 (association)”的以下使用都是指本发明的特定用户头脑中的关联项。所述方法可以包括下列步骤:提供关联项数据库,该关联项数据库对于所述组中的每个非罗马字母表字符都包括与所述每个非罗马字母表字符相关联的罗马字母表语音音译和与所述每个非罗马字母表字符相关联的多个条目;接收需要的非罗马字母表字符的语音音译和至少一个与需要的非罗马字母表字符相关联的相关条目;访问关联项数据库并且将所述组中那些与所接收的语音音译和至少一个所接收的相关条目相关联的字符识别为候选字符;如果候选字符的数目大于1,则接收另外的相关条目并且重复访问和识别步骤,直到多个候选字符被缩小为单个候选字符;以及提供所述单个候选字符作为需要的非罗马字母表字符。

[0008] 根据进一步的实施例,相关条目中的一个或多个可以与需要字符的含义无关。关联项数据库中的相关条目中的一个或多个是用户提供的。如果接收的相关条目当前不与需要字符相关联,则所述方法还可以包括将所接收的相关条目与需要的字符相关联并且将所接收的相关条目存储在关联项数据库中的步骤。如果接收的相关条目当前不在关联项数据库中,则所述方法还可以包括执行将所接收的相关条目增加到关联项数据库中的步骤。所述提供步骤可以在例如所述一组非罗马字符包括日语字符、汉语字符、朝鲜语字符、希伯来语字符和/或阿拉伯语字符的情况下执行。所述提供步骤可以在关联项数据库包括具有多行的关联项表的情况下执行,每一行包括所述组中的单个非罗马字母表字符、所述单个非罗马字母表字符的罗马字母表语音音译和与所述单个非罗马字母表字符相关联的多个条目。所述接收步骤例如可以通过到计算机的语音输入、键盘输入和/或机器视觉而执行,该计算机耦合到关联项数据库。将关联项数据库个性化的步骤可以被执行,以使得关联项数据库对于每个用户都包含从所述每个用户所接收的相关条目。所述方法还可以包括排列关联项数据库中的相关条目的步骤,以使得那些所接收的在访问和识别步骤的较少迭代中引导至单个候选字符的相关条目的级别高于在访问和识别步骤的较多的迭代中引导至单个候选字符的相关条目。所述方法还可以包括搜索计算机网络(例如因特网或者企业网)以查找可能包括所提供的非罗马字母表字符的文档的步骤。所述接收步骤可以在相关条目与需要的非罗马字母表字符的语言相同的情况下执行。所述接收步骤可以在相关条目与需要的非罗马字母表字符的语言不同的情况下执行。所述方法还可以包括接连返回到语音音译接收步骤以选择多个非罗马字母表字符。

[0009] 根据另一实施例,本发明是一种机器可读介质,在其上存储表示指令序列的数据,所述指令序列在被计算设备执行时,通过执行步骤而促使计算设备从一组非罗马字符中选择需要的非罗马字母表字符,所述步骤包括:提供关联项数据库,该关联项数据库对于所述组中的每个非罗马字母表字符都包括与所述每个非罗马字母表字符相关联的罗马字母表语音(或者日语假名或其它语音字母的)音译和可以与所述每个非罗马字母表字符相关联的多个条目;接收需要的非罗马字母表字符的语音音译和至少一个与需要的非罗马字母表字符相关联的相关条目;访问关联项数据库并且将所述组中那些与所接收的语音音译和至少一个所接收的相关条目相关联的字符识别为候选字符;如果候选字符的数目大于1,则接收另外的相关条目并且重复访问和识别步骤,直到多个候选字符被缩小为单个候选字

符 ; 以及提供所述单个候选字符作为需要的非罗马字母表字符。

[0010] 根据另一实施例, 本发明是一种计算机系统, 用于从一组非罗马字符中选择需要的非罗马字母表字符, 所述计算机系统包括至少一个处理器 ; 至少一个耦合到所述至少一个处理器的数据存储设备 ; 由所述至少一个处理器产生的多个过程, 所述过程包括用于以下的处理逻辑 : 提供关联项数据库, 该关联项数据库对于所述组中的每个非罗马字母表字符都包括与所述每个非罗马字母表字符相关联的罗马字母表语音音译和与所述每个非罗马字母表字符相关联的多个条目 ; 接收需要的非罗马字母表字符的语音音译和至少一个与需要的非罗马字母表字符相关联的相关条目 ; 访问关联项数据库并且将所述组中那些与所接收的语音音译和至少一个所接收的相关条目相关联的字符识别为候选字符 ; 如果候选字符的数目大于 1, 则接收另外的相关条目并且重复访问和识别步骤, 直到多个候选字符被缩小为单个候选字符 ; 以及提供所述单个候选字符作为需要的非罗马字母表字符。

[0011] 本发明的另一实施例是计算机实现的方法, 用于搜索可以包括用户可能未知的一个或多个预定非罗马字母表字符的文档。这样的方法可以包括下列步骤 : 以任何语音表示形式 (例如日语假名或朝鲜语的非汉语风格字母表字符) 或符号输入 (例如手语) 提供需要的非罗马字母表字符的发音 ; 提供用户将其与需要的非罗马字母表字符相关联的至少一个相关条目 ; 从关联项数据库中取得一个或多个候选字符, 所述候选字符的发音与所提供的发音匹配并且与所述至少一个相关条目中的一个或多个相关联 ; 如果候选字符的数目大于 1, 则提供另外的相关条目并且重复所述取得步骤, 直到单个候选字符被取得 ; 以及执行搜索以查找可能包括所述单个候选字符的文档。希望关联项的使用将会帮助将搜索缩小为搜索针对的字符的期望示例, 而不是挖掘该特定字符的所有事件。

[0012] 根据进一步的实施例, 在发音提供步骤中, 发音例如可以是语音发音。在相关条目提供步骤中, 相关条目中的至少一个可能与预定字符的含义 (例如字典含义) 无关并且可能是给定用户的任意关联项。如果提供的相关条目当前不与预定字符相关联, 则所述方法还可以包括将所提供的相关条目与预定字符相关联并且将所接收的相关条目存储在关联项数据库中的步骤。如果提供的相关条目当前不在关联项数据库中, 则所述方法可以包括将所提供的相关条目增加到关联项数据库中的步骤。所述发音提供步骤可以在所述非罗马字符例如是日语字符、汉语字符、朝鲜语字符、希伯来语字符或阿拉伯语字符等多种可能的语言的情况下执行。所述取得步骤可以在关联项数据库包括具有多行的关联项表的情况下执行, 每一行包括单个不同的非罗马字母表字符、所述单个非罗马字母表字符的罗马字母表发音和与所述单个非罗马字母表字符相关联的多个条目。所述提供步骤例如可以通过到计算机的语音输入、键盘输入和 / 或机器视觉而执行, 该计算机耦合到关联项数据库。所述方法还可以包括登入步骤, 以使得所述取得步骤从为登入用户个性化的关联项数据库中取得所述至少一个候选字符。所述搜索步骤搜索计算机网络以查找文档。所述相关条目提供步骤可以在相关条目与预定非罗马字母表字符的语言相同或不同的情况下执行。所述方法还可以包括接连返回到所述发音提供步骤以得回用于另外的预定非罗马字母表字符的另外的候选字符。所述发音提供步骤中的发音可以被配置为使用任意语音字母表。所述取得步骤可以被配置为取得任何计算机可读文件或数据, 例如罗马字符、非罗马字符、图片或音乐。

[0013] 根据另一实施例, 本发明可以被视为一种计算机实现的方法, 用户用其在计算机网络上或者在用于搜索的计算机内的本地存储装置上搜索文档、图片、符号、概念或者数据

对象（在这里总称“文档”）。所述方法可以包括下列步骤：接收搜索项目；接收至少一个相关条目，所述至少一个相关条目与搜索项目有关；查阅关联项数据库，所述关联项数据库存储多个字典含义和用于每个字典含义的至少一个相关条目，每个相关条目与其相应字典含义有关；将所接收的相关条目与所存储的相关条目和相应字典含义中的一个或多个相匹配；以及搜索计算机网络以查找关于或者包含搜索项目以及所匹配的存储的字典含义和相关条目的文档。所述方法还可以包括在所述搜索步骤成功时将所接收的至少一个相关条目增加到关联项数据库中的步骤。所述方法还可以包括提示用户提供另外的信息的步骤，所述另外的信息例如是类别、特性和 / 或成员资格。所述方法还可以包括从用户接收排除的相关条目的步骤，以使得所述搜索步骤不搜索一些文档，这些文档包括或涉及与所接收的排除的相关条目相对应的搜索项目的一个或多个字典含义。

附图说明

- [0014] 图 1 是示出用于多个个具有罗马字音译“ken”的汉字字符的关联项的表。
- [0015] 图 2 是示出用于多个个具有罗马字音译“kitai”的汉字字符的关联项的表。
- [0016] 图 3 是示出用于多个个具有罗马字音译“kiri”的汉字字符的关联项的表。
- [0017] 图 4 示出了根据本发明一实施例的示例性方式，其中关联项数据库可以被预播种（pre-seed）。
- [0018] 图 5 示出了根据本发明一实施例的关联项数据库的示例性表。
- [0019] 图 6 是根据本发明一实施例的输入字符的方法的流程图。
- [0020] 图 7 示出了根据本发明一实施例的示例性方式，其中关联项数据库可以被个性化。
- [0021] 图 8 示出了采用本发明一实施例的示例性的搜索引擎。
- [0022] 图 9 示出了根据本发明一实施例的由图 7 的搜索引擎所返回的结果。
- [0023] 图 10 示出了根据本发明一实施例的计算机，用其可以实践本发明的实施例。

具体实施方式

[0024] 甚至其书面形式不依赖于罗马字母表的语言（例如日语和很多汉语方言）的母语使用者用眼所识别的字符数通常远远多于他们能够用手再现的字符数。本发明的实施例即使在用户仅具有需要的字符的样子的模糊的、不清晰的记忆时，也能够唯一地选择需要的非罗马字符。这样，本发明的实施例使得用户能够通过提供语音发音和一个或多个相关条目而表现比他们实际所拥有的字符书写的能力和技能明显更高的能力和技能。这样的条目是通过一系列的击键、口述词和 / 或任何其它计算机可感知的输入形式（例如机器视觉）做出的。这样，本发明的实施例帮助用户成功地输入比他们真正知道（或者真正能够通过重建所有的敲击而形成）的字符数更多的字符数。本发明的实施例可以应用于文档撰写以及搜索引擎（例如包括Google®、Yahoo®等）中，用于扫描和搜索用非罗马字母表驱动的书面形式所公布的网站和文档。

[0025] 在附图和以下的描述的书面部分中，重点放在汉字书写的日语形式。然而，本发明的实施例可同等地应用于所有其它书面表达形式，这些书面形式不依赖于文本以其进行书写的罗马字母表。例如，本发明的实施例容易扩展到很多汉语方言、阿拉伯语形式和希伯来

语。因此,本发明所主张权利的实施例明示地希望包括所有这样的不依赖于罗马字母表的书面表达形式,并且为了搜索,希望包括任何字符系统中的词的特定实例以及它们以罗马字母或者任何语音表示形式呈现的相关词,所述任何字符系统中的词包括但不局限于罗马字。

[0026] 现在参照附图,图 1 是示出用于多个个具有罗马字音译“ken”的汉字字符的关联项的表。在图 1 中,在 102 处示出了罗马字音译,在由数字 104 所指出的列中示出了汉字字符,在 106 处可以找到关联项。这些字符称为“同义语”或相似发音的日语字符。在图 1 中,“ken”是汉字字符 104 发音的罗马字符版本;即它们的音译。与每个汉字字符 104 相邻的是多个(在本实施例中是英语)条目(在本示例中是词),这些条目与汉字字符 104 相关联并且其含义与汉字字符相关。关联项 106 可以是任何语言,包括与字符 104 的语言相对应的口语。例如,关联项 106 可以是日语。根据本发明的实施例,相关的条目和含义 106 可以包括语言(该情况中为日语)的母语使用者所熟悉的词和概念,这些母语使用者共享文化和所产生的词/字符关联项。因此,根据本发明的实施例,相关条目 106 不仅可以包括包含在给定字符的相应字典条目中的词,还可以包括例如典型用户可能将其与相应的汉字字符相关联的其它词、概念和/或短语以及声音。例如,虽然图 1 中的第五行示出了字典条目是“right”和“morally good”的汉字字符 108,但是罗马字音译是“ken”的图 1 中第五行的字符 108 的相关条目 106 的列表也可以包括用户可能将其与该字符相关联的其它词、概念和/或短语。例如,可以与汉字字符 108 相关联的条目也可以包括词“Power”、“Authority”、“Influence”、“Wrong”和“Privilege”,即使这些词中的每一个可能都在日语中具有它们自己的汉字符号。根据本发明的实施例,与给定的非罗马字母表字符相关联的条目可以是该字符的精确的英语对应物,可以具有与基本字符紧密相关的含义,可以具有仅与基本字符略微相关的含义或者与非罗马字母表字符的公认含义完全无关。根据本发明的另一技术方案,相关条目根本不必是词,而是可以是用户将其与需要的字符相关联的语音表示和/或任何机器可感知的可以在机器可读介质中存储和访问的用户表现。例如,只要该用户表现可以电子地表示,相关条目就可以包括由用户所作出的声音、移动或者手势的表示。实际上,语音、手势、手语、键盘输入、具有提高的或降低的语调、口音、强调和/或其它说话方式的语音或者动作可以被捕捉、记录并且用作相关条目以便找到需要的一个或多个字符。基于经验的关联项表从不同的键盘、语音和其它输入到相关的关联项数据库 402 的建立实现了与特定字符的关联。通过使用这样的输入,用户可以快速地选择需要的字符,并且该字符可以被快速地用于构建文档或者形成搜索一个或多个包含所需要的(一个或多个)字符的文档的基础。根据本发明,文档可以包括任何计算机可读文件,例如由服务器(例如网络服务器)响应于来自客户端的请求而产生的文本文档或者 XML 或 HTML 文档(例如网页),所述请求例如是 HTTP 请求。这样的 HTML 文档可以包括对图像和/或其它可以在客户端侧分辨的丰富数据源,客户端可能必须发出另外的请求以获得嵌入的图像、声音、JavaScript 代码片段等。不同于传统的通过翻译而进行的语音定向的字符选择,并且不同于传统的基于发音的键盘选择,本发明的实施例均衡每种类型的用户输入作为选择需要的字符中的另一因素。可以重复本方法以建立用于文档写作或者更复杂搜索的字符串。

[0027] 根据本发明的实施例,只要至少一个用户在其头脑中形成了字符(例如图 1 中的汉字字符 108)和相关条目之间的关联即可。例如,如果用户将词“Moon”与字符 108 相关

联,则即使词“Moon”明显与字符 108 无关,该词也可以与字符 108 相关联。

[0028] 类似地,图 2 的第一行中的汉字字符 208 可以被在字典中定义为“Gas”或者“Vapor”的含义。因此,这些含义可以被列为相关条目,连同字符 208 的罗马字音译“kitai”。然而,用户可以形成与该汉字字符 208 的其它关联。这些关联中的一些可以被经历相同文化、培养、流行文化、和教育的其他人共享,而这些关联中的其它关联可以仅由一小部分处于相似境遇的人共享或者甚至由单个人共享。例如,在很多用户的头脑中,词“Air”、“Explode”、“CO2”或者“Bhopal”可以与字符 208 相关联。当搜索字符 208 时,他们可以回想其罗马字音译“kitai”,并且能够提供一个或多个与其相对应的相关条目。图 3 示出了另外的汉字字符、它们所共享的各自罗马字音译(该情况中为“kiri”)和可以与其相关联的多个条目。

[0029] 本发明的实施例包括关联项数据库 402,关联项数据库 402 例如包括来自 Oracle 公司的相关数据库管理系统 RDMS。数据库 402 可以存储一组非罗马字母表字符、它们各自的语音发音或其它音译和一个或多个相关条目。在这里所提供的实施例中,存储在数据库 402 中的非罗马字母表字符是汉字字符并且各自的音译是罗马字音译。其它实施例包括其它非罗马字母表字符和其它发音或音译。

[0030] 图 4 示出了根据本发明一实施例的示例性方式,其中关联项数据库 402 可以被预播种。数据库 402 可以包括一个或多个表,这些表包括多个记录。在 404 处示出一个有代表性的记录。每个记录可以存储为数据 402 中的一行。根据本发明一实施例,可以用一组或多组非罗马字符(例如大约 40,000 个的汉字字符)、它们的罗马字音译(或者例如语音发音)和一个或多个用于每个汉字字符-罗马音译对的相关条目对数据库 402 进行预播种。为了预播种关联项数据库 402,如 408 处所示,现有的数据库、编译和/或字典可能有用。本发明的实施例还要求用一个或多个用于每个汉字字符的相关条目来预播种数据库 402。如图 4 所示,该预播种过程可以包括面试多个参与者,这些参与者被认为或多或少地代表了本发明的期望用户(例如就教育和文化背景而言)。这样的参与者在图 4 的标号 410、414 和 418 处示出。例如,面试过程可以包括向参与者示出汉字字符及其罗马字音译并要求他们对每个汉字字符进行自由关联。在图 4 的示例中,汉字字符 406 可以在日英字典中被列为与词“Ticket”、“Coupon”、“Bond”和“Certificate”相对应。汉字字符 406 的这些普通含义中的每一个可以被作为预播种的相关条目连同音译“ken”一起包括在记录 404 中。根据本发明一实施例,数据库 402 的一个或多个记录可以在每个记录中包括另外的相关条目,这些条目可以如下选择。可以要求每个参与者 410-418 对汉字字符 406 进行自由关联。结果,参与者可以想到多个不同的关联项,如泡泡 412、416 和 420 所示。例如,在被示出字符 406 并且看到/听到音译“ken”之后,参与者 410 可以将词“Ticket”、“Permit”、“Certificate”、“Bond”、“Pink”和“Coupon”与该字符相关联。这些关联形成候选关联项。也示出了其它参与者 414 和 418 的候选关联项。那些在参与者之间匹配的候选关联项用粗体示出。根据本发明一实施例,那些由大多数参与者作出的候选关联项可以被选为另外的预播种相关条目并且被包括在记录 404 中。例如,参与者将词“Certificate”、“Bond”、“Pink”和“Coupon”与字符 406 相关联,这些候选关联项可能已经被作为相关词包括在记录 404 中。不必在记录 404 中复制这些。然而,多于一个的参与者将词“Speeding”、“Concert”和“Traffic”与字符 406 相关联。如果足够(或者可选)数目的参与者用这些关联项进行响应,则这些关

联项也可以被包括在用于字符 406 的记录中,如图 1 所示。选择候选关联项作为相关条目包括在记录 404 中的方式是完全可选的。例如,一种多数算法 (majority algorithm) 可以用于确定候选关联项中的哪些被选择包括在每个记录中。

[0031] 注意到参与者 410 响应于自由关联的请求,除了词“Ticket”、“Bond”、“Certificate”和“Coupon”之外,还将词“Permit”和“Pink”与字符 406 相关联。然而,由于这些候选关联项未被足够数目的参与者所共享,因此这些候选关联项不被选择作为相关条目包括在记录 404 中。同样,候选关联项“Receipt”、“Official Document”和“Money”未被足够数目的参与者所共享,因此不被选择作为相关条目包括在记录 404 中。类似地,由参与者 418 所想到的词“Record”、“Hashimoto”和“Diploma”也不被选择。注意到虽然那些条目可能未包括在用于播种目的的通用数据库中,但是它们将保留在数据库的依用户而定的部分中以供输入了它们的人使用。这里的目的是将任何被尝试的候选关联项或者累积到通用数据库中或者累积到仅由采用了这些候选关联项的特定用户可见的数据库中,如果存在具有多个用户的中央数据库的话。然后,数据库的每个登入用户将具有可用关联项的不同组;关联项的通用组将是最初所播种的关联项加上增加到初始播种存储中的通过算法选择的另外的关联项,最后,用户自己的特别的关联项未被算法标记以包括在通用数据库中。

[0032] 由上可知,相关条目不必具有与字典的直接语义关系或者字符 406 通常所理解的含义。为了被包括在记录 404 中,只要足够数目的参与者认为候选关联项在他们的头脑中由于某种原因而与字符 406 相关就行了。这就是例如候选关联项“Concert”如何被作为相关条目而包括在记录 404 中。

[0033] 图 5 示出了根据本发明一实施例的关联项数据库 402 的示例性关联项表。图 5 中示出了四个表。第一个这样的表 502 是日语表,第二个表 504 是普通话汉语表,第三个表 506 是广东话汉语表,以及第四个表是上海话汉语表。实际上,关联项数据库 402 中可以包括仅一个这样的关联项表或者更多这样的关联项表。例如,关联项表可以被组织得就像是关联项表 502,但是这些表不一定要这样组织。如图 5 所示并且根据本发明一实施例,表 502 的每一行都是一个记录并且每个记录包括多列。第一列可以存储非罗马字母表字符(本示例中为汉字字符)或者与字符相对应的代码(例如与大约 40,000 个汉字字符中的一个相对应的 2 字节的词)。另一列可以存储例如字符的罗马字音译或者其另外某种语音发音。然后每个记录可以包括一个或多个相关条目,例如可以用相对图 4 详述的并且如下所述的方式选择这些相关条目。要了解到图 5 仅是概念上的说明,关联项数据库 402 内的关联项表的物理组织不一定如图 5 所示,而是可以根据需要而改变以优化效率或者其它要求。

[0034] 虽然有很多种汉语方言,但是书面汉语字符只有一个通用列表,尽管给定的字符可能具有可选的传统格式或现代的短格式。在任一种情况中,不论对何种方言而言,它们都是同样的。一个人甚至可以用书面汉语字符来表示美国手语;符号不是必须与特定语音对应物相关。因此,为了使本发明的实施例能够对不同方言的使用者有用,需要提供关联项表以适应针对其而实现本发明的这些特定方言中的每一种的使用者。例如,虽然汉语书面字符可以在所有的汉语关联项表之间共享,但是其罗马字母表音译可能是不同的,与这些字符中的每一个相关联的条目也可能是不同的。实际上,相关条目在汉语方言关联项表的每一个中可能是不同的,这不仅因为方言之间的差别,还因为这些方言的母语使用者和体现本发明技术方案的应用程序的预期用户的不同文化、教育和社会背景。

[0035] 要了解到可以为除了日语和很多汉语方言之外的语言提供和填充另外的表。如上所述,可以为阿拉伯语、希伯来语、朝鲜语或者任何其它在其书面形式中使用非罗马字母表字符的语言开发关联项表。

[0036] 图 6 是根据本发明一实施例的用于从一组非罗马字符中选择需要的非罗马字母表字符的方法的流程图。所述方法开始于 S61,在此可以提示用户输入需要的非罗马字母表字符的罗马字母表音译(或者例如语音发音)。为了说明的目的,虽然了解到本发明不局限于日语汉字字符,但是这里所讨论的方法是关于日语汉字字符的。因此,步骤 S62 提示用户向运行体现本发明技术方案的应用程序的计算机输入(或者提供)需要的汉字字符的罗马字音译。本发明的实施例不局限于通过计算机键盘的用户输入,而是可以被配置为任何形式的用户输入(例如包括音译、语音发音和/或相关条目),例如口语词(使用语音识别技术)、书面形式(使用光学字符识别技术)、视觉输入(例如使用机器视觉和模式识别技术)或者实际上计算机可感知的任何数据输入形式。例如,体现本发明实施例的计算机可以设有用于识别手语、手势或者身体姿势的装置以指出数种可能性。本发明的实施例可以适用于操作例如为那些具有身体或精神损伤的人所开发的多个机器交互装置。

[0037] 如步骤 S62 的要求输入了需要的汉字字符的罗马字音译,然后确定所输入的罗马字音译是否与任何汉字字符相对应。如果不对应,则可以提示用户重试并且输入另一罗马字音译或者先前输入的罗马字音译的变体。然而,如果如步骤 S63 所示,输入的罗马字音译仅与一个汉字字符相对应,则方法进行到步骤 S67,其中唯一的汉字字符被输入。可以在步骤 S67 之前增加用户确认步骤以向用户确认响应于用户提供的罗马字音译而提供的单个汉字字符是需要的(例如正确的)字符。如果提供的汉字字符不是需要的字符,则用户可以返回步骤 S62 进行重试。而且,如果在搜索中返回正确的字符但是错误的使用,则用户可以用不同的关联项重试。

[0038] 然而,如果输入的罗马字音译与多于一个的汉字字符相对应(如同是具有图 1-3 的音译“ken”、“kitai”和“kiri”的情况),则方法可以从步骤 S63 进行到步骤 S64。在此,计算机系统可以向用户显示或者提供候选汉字字符。然而,因为显示多个候选字符可能证明对于用户来说是非常让人分心的,因为他或她正在创作和提供相关条目,所以步骤 S64 可以省略。要注意到这里所描述的和附图中所示的相关条目不一定是所示出的英语词。例如对于日语的母语使用者,相关条目可以是用户用需要的字符作出的关联项的语音(假名)形式。因此,如下详细所述,可以用任何语言和/或任何形式来定义相关条目。候选字符是那些与用户所提供的罗马字音译相对应的汉字字符。例如,可以响应于用户输入罗马字音译“ken”而为用户显示图 1 中所示的字符作为候选汉字字符。然后可以给用户直接选择候选汉字字符中的一个作为需要的汉字字符的机会。然而,本发明的实施例使得用户能够向计算机提供一个或多个相关条目,以试图将候选汉字字符的范围缩小为单个汉字字符。因此,如步骤 S65 所示,计算机可以提示用户以要求一个或多个相关条目,例如图 4 和图 5 所示的相关条目。在步骤 S66 中,确定所提供的相关条目实际上是否真的将候选汉字字符的范围缩小为单个汉字字符。如果不是,则可以提示用户提供一个或多个另外的相关条目并且方法可以返回到步骤 S65。如果仅剩下单个候选汉字字符,则唯一的剩余候选汉字字符应该是用户所寻找的汉字字符,如步骤 S67 所示。如果不是,则可以重复所述过程。方法结束于步骤 S69。

[0039] 本发明的实施例不局限于其中关联项表（见图 5）被预播种（即被用由其他人例如如图 4 中所示的方式所选择的相关条目填充）的情况。实际上，本发明的实施例可以通过允许这里所描述的非罗马字母表字符选择系统从其用户处学习而获得增强的功能。实际上，这里所描述的非罗马字母表字符选择方法和系统不仅可以从其用户总体学习，还可以被配置为学习并适应于个体用户。因为认为没有两个用户具有同样的背景，所以这是需要的。而不同的背景（教育、社会、文化等等）必然使得用户形成不同的关联项和心智图像（mental image）并且因此将不同的条目与不同的字符相关联。因此，给定的条目或词与给定汉字字符的对一个用户来说可能有意义或直观的关联可能不（或者可能将不）对另一用户来说具有意义或者直观，即使是处于相似情形的关联。

[0040] 简短地返回图 4，在 410、414 和 418 处示出的用于预播种关联项数据库的参与者共享多个关联项，例如用于罗马字音译“ken”的“Certificate”和“Speeding”。然而，参与者 418 将“Hashimoto”与汉字字符 406 和音译“ken”408 相关联，这样做的原因可能仅由他或她知道。对于由参与者 410 为同一汉字字符 406 所提供的相关词“Pink”也是如此。虽然这些关联项不符合作为预播种的相关条目包括在关联项数据库 402 中的选择标准，但是这些关联项对于参与者 410、418 来说仍然明显地非常有意义。因此，本发明的实施例有利的允许能够对当前的字符选择系统的每个用户定制关联项表（见图 5）。然后这些用户可以回想那些相同的关联项以选择需要的字符。

[0041] 图 7 示出了根据本发明一实施例的示例性方式，其中对当前字符选择方法和系统的每个用户，关联项数据库 402 可以被个性化。示出了三个用户；即用户 702Alice、用户 710Bob 和用户 718Charlie。根据本发明一实施例，可以用参考图 4 所示出和描述的方式对关联项数据库 402 进行预播种。而且，关联项数据库 402 可以通过将其用户所提供的相关条目包括在其中而被扩展。如图所示，用于 Alice 的字符 406 的记录除了预播种的相关条目之外，还可以包括 Alice 将其与字符 406 和 / 或音译 404 相关联的条目。如 704 处所示，Alice 将其与字符 406 相关联的相关条目是“Form”、“Application”和“Off-Track Betting”。因此，在关联项数据库 402 中，这些依用户而定的关联项将被包括在用于字符 406 的记录中。因此，可以用 Alice 在被呈现以字符 406 时所作出的关联项，如 708 处所示地对关联项数据库 402 进行个性化。因此，当前的字符选择方法和系统可以包括训练部件，其中每个用户训练系统将所选条目与所选字符相关联。在这种方式下，用户 710 可以将条目“Record”和“Satai”与同一汉字字符 406 相关联，如 712 处所示。例如，相关条目“Satai”可能仅对用户 710Bob 有用，而对任何其它用户都没用。然而，该相关条目“Satai”对于 Bob 来说可能是最有效的关联项，因为不管这样的关联背后的原因，这可能是 Bob 将其与字符 406 最接近地相关联的词。用户 718Charlie 可能将金融有关的术语与字符 406 相关联，这反映在他与字符 406 相关联的条目的选择上，如 720 处所示。然后这些个体关联项被合并到用户个体的个性化关联项数据库中，如标号 708、716 和 724 处所示。这些可以是分开的数据库，或者被集成到关联项数据库 402 中。可以实现例如登入过程以使得系统能够区分用户。

[0042] 根据本发明另一实施例，可以为多个用户仅提供单个关联项数据库 402。例如在该情况中，可以为每个用户安排相关条目或发音，最成功的相关条目（即，使得选择单个字符或者使得用最少数目的相关条目成功选择需要的字符的相关条目）出现在不那么成功的相关条目之前。而且，可以排列相关条目以允许关联项数据库内的高效二分搜索（binary

search)。根据另一实施例,仅那些经常使用和/或使得成功选择需要的字符的相关条目可以保留在数据库中。可以从用于需要的字符的关联项数据库中的记录中剔除那些不使用或者未使得高效选择需要的字符的相关条目。在这种方式下,关联项数据库 402 可以最终变得仅仅或者主要由用户中意的关联项(即最成功的相关条目)填充。注意到关联项数据库 402 可以被进一步排列为可以将音译自身放入排序次序中,以使得可以对它们进行二分搜索。关联项表 402 中的所有条目都可以被索引以用于快速访问。

[0043] 通过监视用户使用的模式,本发明的实施例可以预测用户的选择并且增强字符选择过程的性能。通过修改如上所详细描述的数据数据库 402,关联项数据库 402 可以仅在很短的时间段(例如数周的繁重使用)之后变得高度个性化。所存储的用户概况的实现使得个体用户能够访问他们自己的“个性化”关联项数据库 402,例如图 7 中的 708、716 和 724 处所示。

[0044] 本发明的实施例可以包括离散训练部件,如参考图 4 所述。然而,本发明的实施例还包括不依赖于离散训练会话的更集成的训练功能。实际上,当前的字符选择系统和方法在效果上可以被配置为从用户的失败关联项中学习。例如,当用户响应于步骤 S65 中的提示而向系统提供相关条目时,他或她也可以提供未出现在用于汉字字符中任何一个的关联项数据库中的相关条目。这些所提供的相关条目不是立即在缩小候选汉字字符的范围时有效,因为系统不能将所提供的相关条目与关联项数据库 402 中的任何记录相匹配。因为然后对步骤 S66 的回答是“否”,所以系统返回到步骤 S65 以提示用户提供一个或多个另外的相关条目。最终,用户应该向系统提供实际上真将候选字符的范围缩小为单个字符的相关条目,所述单个字符在步骤 S67 中变成所选择和需要的汉字字符。由于用用户所提供的相关条目选择单个汉字字符的过程是成功的,因此关联项数据库 402 可以被更新以包括那些用户所提供的先前未存储在关联项数据库中的相关条目,如步骤 S68 所示。然后,这些新的相关条目可以被存储在与步骤 S67 中所选择的汉字字符相对应的记录中。用户下一次想要选择同一汉字字符时,新增加的相关条目可以在选择需要的汉字字符或者至少在缩小候选汉字字符的范围(缩小候选汉字字符的数目)时有效。在这样的搜索情形中,用户在被给定所使用的关联项的情况下将更有可能找到由用户的语音表示形式和所提供的关联项确定的需要的引用。

[0045] 因为关联项数据库 402 中的表可以具有用于每个汉字字符的多个相关条目,所以可以预见一些相关条目将是多于一个的记录所公共的;也就是说,与多于一个的汉字字符相关联。在该情况中,候选字符(可能或者可能不提供给用户)的数目大于 1,并且表示其记录包含公共相关条目的字符的数目。为了解决这个问题并且将候选字符的数目筛选为 1,用户可以提供一个或多个另外的相关条目,这些条目将会减少包含这样的公共相关条目的记录的数目。因为用户提供另外的相关条目并且那些条目被应用于候选字符的一直缩减的整体,所以该数目将会很快地减少为单个候选字符。

[0046] 本发明的实施例可以被配置为搜索引擎或者可以被配置为通过提升搜索的功效率和通过返回与用户更相关的结果而补充和提高现有的搜索技术。图 8 示出了采用本发明一实施例的示例性和说明性的搜索引擎。该搜索引擎例如可以配置为企业网搜索引擎或者包括在企业网搜索引擎中,或者配置为网络搜索引擎。搜索的可以是文档、数据库条目、依附在图片上的标签或文本、词、字符、图片或者任何描述的数据对象,在这里总称“文档”。如

图 8 所示的用户界面的实现方式所示,可以通过(例如)标准浏览器 800 在 802 处提示用户输入如上所述的需要的非罗马字母表字符的发音(例如可以是语音发音)或音译。也可以提示用户输入一个或多个相关条目到文本框 804 中。在这种情况下,用户所输入的音译是“ken”,用户在文本框 804 所输入的相关条目是“Sword”、“Blade”和“Jet Li”。了解到虽然这里所示出的相关条目是英语的,但是相关条目可以用任何语言输入,包括需要的字符的语言(该情况中为日语)。一旦用户完成了音译和相关条目两者的输入,用户可以单击或者选择具有示例性文本“查找汉字字符”的按钮 806。假设所输入的音译和所输入的相关条目与单个候选字符相对应,则可以出现图 9 的示例性屏幕。响应于用户的输入,搜索引擎可以搜索关联项数据库以查找包含在一些记录中的汉字字符,这些记录包含用户所输入的音译和相关条目。在这种情况下,返回唯一的候选汉字字符 906。然后用户可以接受该字符并且在 902 处指示搜索引擎在网络(例如因特网)上查找字符的示例,或者可以决定通过选择按钮 904 而在文本框 804 中增加另外的相关条目或者用新的相关条目替代先前包含在其中的相关条目。或者,用户可以被呈现以多于一个的候选字符,并且如果用户可以识别他或她所寻找的字符,则可以简单地指向需要的字符或者从那些所呈现的字符中选择需要的字符,而不是提供另外的相关条目。如上所提及的,很多用户识别的字符远远多于他们能够再现的字符。本发明的实施例正是支持了这样的能力。

[0047] 或者,用户可以反复使用这里所描述的方法来构建这样的(汉字或其它)字符的搜索串以用于搜索例如基于字符的网页和/或公司内联网,而不是搜索单个字符。本发明实施例的其它使用可以包括用于识别基于网络的拍卖的项目以高效地匹配买方和卖方。

[0048] 本发明的实施例提供为了创建文档、指定搜索和/或控制应用程序并对应用程序有所贡献而访问特定非罗马字母表字符(例如日语和汉语字符)的容易掌握的、高性能的和灵活的方法和系统。本发明的实施例可以容纳各种数据库,所有的数据库都具有相关字符。例如,虽然汉语是具有很多口语的单一书面语言,但是关联项数据库可以被构建以使得公共字符可以与多于一种的方言相关联。例如,克里奥尔语(来自两种或多种语言的方言组合)可以用于访问所构成语言的通常含义相同但发音不同的字符。与字符访问的传统手段不同的是,本发明的实施例可以嵌入在卖给软件购买者的那些应用程序中,并且不必链接到任何特定的硬件。

[0049] 搜索项目的“环境”和该环境可以如何建立是所有高级搜索算法的基础。本发明的实施例可应用于任何书面语言的搜索引擎中。实际上,本发明的实施例的效用不止是基于字符的文档的产生和搜索。例如,通过向搜索引擎提供这样的环境,本发明的实施例允许搜索引擎找到与搜索者的搜索目标最接近的项目并且排除所有其它项目。将本发明的技术方案应用于搜索引擎,用户可以通过提供上下文以便提高一些命中相对于其它命中的分数而提高搜索的效力和效率。提供上下文也可以在排除无用的项目中有效,从而筛选搜索结果使其减少为用户最可能感兴趣的那些搜索结果。如下详细所述,本发明的实施例可应用于任何语言的搜索引擎,而不管在其书面形式中使用罗马字符还是非罗马字符。

[0050] 了解到可以使用以下所描述的实施例而不是现有的搜索引擎和技术,或者除了现有的搜索引擎和技术之外还可以使用以下所描述的实施例,不管这些是否涉及搜索网络、搜索数据库或者简单搜索文本文档中成功搜索的历史的编译、链接频率或其它已知手段。

[0051] 根据本发明一实施例,为了开始搜索,用户可以口头输入、打字输入或者另外提供诸如英语(和/或例如他或她自己的母语)之类的任何语言的一个或多个搜索项目。然后所提供的搜索项目可以被分辨成一系列同音异义词并且上面所详细描述的方法可以用于将所提供的一个或多个搜索项目分辨成词、字符或一组书面项目。通过参考数据库 402 以找到用于搜索项目指定的语言,不管是什么语言,口头响应和键入响应的交替都可以快速将一个或多个搜索项目的初始指定分辨成一系列字典含义。

[0052] 所提供的搜索项目的字典含义可以包括数种字典含义。例如,英语词“court”具有多个不同的英语含义。在 Webster 字典中,词“court”具有 5 种主要含义和 5 种主要含义内的 12 中次要含义。用户规定这些含义中的一种用于所搜索的词的能力将允许由搜索引擎进行明显更加精细的搜索。甚至在没有搜索项目匹配的情况中,所检查的目标源中的上下文项目也可以允许体现本发明技术方案的搜索引擎返回有意义的搜索结果。在 Webster 字典中,词“court”被如下定义:

[0053] Main Entry :court

[0054] Pronunciation: ' k0rt, ' kort

[0055] Function :noun

[0056] Usage :Often attributive

[0057] Etymology :Middle English, from Old French, from Latin cohort-, cohorsenclosure, group, retinue, cohort, from co-+-hort, -hors(akin to hortus garden)--more at YARD

[0058] 1a :the residence or establishment of a sovereign or similar dignitary 1b :asovereign's formal assembly of councilors and officers 1c :the sovereign and officers and advisers who are the governing power 1d :the family and retinue of a sovereign 1e :a reception held by a sovereign

[0059] 2a(1) :a manor house or large building surrounded by usually enclosed grounds 2a(2) :MOTEL 2b :an open space enclosed wholly or partly by buildings or circumscribed by a single building 2c :a quadrangular space walled or marked off for playing one of various games with a ball (as lawn tennis, handball, or basketball) ;also :a division of such a court 2d :a wide alley with only one opening onto a street

[0060] 3a :an official assembly for the transaction of judicial business 3b : a session of such a court < court is now adjourned > 3c :a place (as a chamber) for the administration of justice 3d :a judge or judges in session ;also :a faculty or agency of judgment or evaluation < rest our case in the court of world opinion -- L. H. Marks >

[0061] 4a :an assembly or board with legislative or administrative powers 4b : parliament, legislature

[0062] 5 :conduct or attention intended to win favor or dispel hostility : homage < pay court to the king >

[0063] 因此,词“Court”具有多个不同的字典含义,可以在搜索期间通过向搜索引擎提供

一个或多个相关条目而分辨这些含义。为了允许这样,这些不同字典含义中的每一个都可以被提供为关联项数据库的表中的一行,其中每一行包含一个或多个相关条目所述相关条目与字典含义有关一或者在用户头脑中与需要的字典含义相关联。根据本发明一实施例,除了搜索项目之外,用户还可以输入相关条目,其例如可以包括含义的领域和 / 或整体上下文(例如技术区域、工业、政府部门等等)。这样的相关条目允许搜索引擎通过将搜索范围限制或扩展为所提供的一个或多个搜索项目的那些字典含义而缩小搜索范围,这些项目的相关条目与用户所提供的一个或多个相关条目相对应。或者,搜索引擎可以部分地基于关联项对“分数”的贡献而为每个方案奖励点数。也可以基于用户所输入的相关条目以及基于用户的先前成功搜索而建立分级类别。而且,当根据本发明一实施例的搜索引擎进行搜索时,其可以提示用户选择一个或多个已知的分级类别。这些已知的分级类别可以在关联项数据库 402 中被预先播种并且 / 或者可以被预先输入先前成功的搜索的相关条目。然后由用户所提供的使得产生成功搜索的这些相关条目可以被增加到关联项数据库 402 中并且对下一搜索变得可用。也可以提示用户创建新的分级类别,这些新的分级类别将被记为用户概况的一部分并且可以被将来使用,如果搜索成功的话。这是用于罗马字符搜索引擎的有效策略,就像日语或汉语字符的情况一样。注意到用户概况可以将先前的个人搜索的结果存储在如上所述的数据库 402 中,并且可以被配置为使用先前的相关条目来帮助确定用户的目标。也可以给用户允许或排除这些过去的相关条目和分级类别的机会。

[0064] 可以根据标准来分析搜索结果,该标准例如可以包括相关性(在页、文档或引用命中中找到多少相关条目)和强度(最高优先级的相关条目的质量)。

[0065] 如在先前所述的实施例中,除了搜索项目之外的由用户所提供的相关条目可以是词和 / 或声音,或者与搜索项目有关并且 / 或者在用户头脑中与其相关联的其它输入。(一个或多个)搜索项目和相关条目的组合中的每一个都可以被并行搜索。例如,可以提示用户输入或者提供用于需要的搜索项目的五个相关条目。或者,可以提示用户提供类别(例如动物、蔬菜或者矿物)作为他或她的用于需要的搜索项目的相关条目。例如,当搜索项目是“工作犬”时,用户所输入的相关条目例如可以包括“犬齿(canine)”、“狩猎(hunting)”和“獾(spaniel)”。

[0066] 基于用于用户的关联项数据库 402 的内容,搜索引擎也可以被配置为向用户提示一个或多个问题的答案,这些问题涉及所提供的搜索项目和 / 或相关条目。例如,搜索引擎可以被配置为向用户提示另外的信息,该信息关于搜索项目的某种特性,例如其尺寸“大于面包盒?”,或者提示用户例如确定搜索项目是否具有任何物理结果。基于搜索项目和相关条目,搜索引擎可以上下搜索分级类别,不管这样的类别是由用户在当前会话期间提供的还是由用户在先前会话期间提供的。

[0067] 搜索引擎可以被配置为向用户提示搜索项目的分类和 / 或成员资格以聚焦搜索。例如,根据本发明一实施例,当前的搜索引擎可以提示用户输入搜索项目的特性作为相关条目或者进一步特性化预先输入的相关条目。例如,如果搜索项目是“大”,相关词是“破坏”和“损坏”,则可以提示用户输入搜索项目和 / 或相关条目的一个或多个特性。例如,怪物、风暴和恐龙都是“大”的并且可以被取得,即使它们仅略微与其相关(它们都在经过之后留下损坏痕迹)。

[0068] 根据本发明一实施例的搜索引擎也可以被配置为排除某些关联项和 / 或词组。例

如,当前搜索引擎的用户可以规定某些排除的相关条目不特性化正在搜索的项目。在这种方式中,当存在分级类别时,用户可以使得搜索引擎夹断(即不搜索)某些分级路径并且/或者不搜索包含和/或与关于排除的含义有关的那些含义。例如,用户可以指定搜索项目“工作犬”并且指定相关条目,相关条目例如是“寒冷天气”、“雪橇”并且包括相关条目“萨莫耶德犬”不包括在内的规定。响应于此,搜索引擎将搜索与工作犬、寒冷天气和雪橇相关联的页和文档,而不会包括指定萨莫耶德品种犬的搜索结果。

[0069] 在同一方式中,搜索引擎可以被配置为向上或向下搜索分级路径,以查找匹配搜索项目的结果和相关条目。类别层级的全部部分可以被完全包括或者从搜索中排除。

[0070] 图 10 示出了本发明的实施例可以在其上实现的计算机系统 1000 的框图。计算机系统 1000 包括总线 1001 或用于传送信息的其它通信机制,以及一个或多个与总线 1001 耦合的用于处理信息的处理器 1002。计算机系统 1000 还包括随机存取存储器(RAM)或其它动态存储设备 1004(称为主存储器),其耦合到总线 1001 并且用于存储信息和将由(一个或多个)处理器 1002 执行的指令。主存储器 1004 也可以用于在处理器 1002 执行指令期间存储临时变量或其它中间信息。计算机系统 1000 还包括只读存储器(ROM)和/或其它静态存储设备 1006,其耦合到总线 1001 并且存储用于处理器 1002 的静态信息和指令。诸如磁盘或光盘之类的数据存储设备 1007 耦合到总线 1001 并且用于存储信息和指令。计算机系统 1000 还可以通过总线 1001 耦合到显示设备 1021,显示设备 1021 用于向计算机用户显示信息。包括字母数字键和其它键的字母数字输入设备 1022 通常耦合到总线 1001 并用于将信息和命令选择传送给(一个或多个)处理器 1002。另一种类型的用户输入设备是光标控制器 1023,例如鼠标、跟踪球或光标方向键,其用于将方向信息和命令选择传送给处理器 1002 并且用于控制光标在显示器 1021 上的移动。如 1025 处所示,话筒可以用于提供口头输入,摄像机可以用于输入用户手势或手语。

[0071] 本发明的实施例涉及计算机系统 1000 的使用并且/或者涉及多个这样的计算机系统以实现用于输入和搜索罗马字符或非罗马字母表字符的方法和系统。根据一个实施例,可以由一个或多个计算机系统 1000 响应于(一个或多个)处理器 1002 执行包含在存储器 1004 中的指令序列而提供这里所描述的方法和系统。这样的指令可以从诸如数据存储设备 1007 之类的另一计算机可读介质中读入存储器 1004。包含在存储器 1004 中的指令序列的执行使得(一个或多个)处理器 1002 执行这里所描述的步骤并且具有这里所描述的功能。在可替换的实施例中,硬件实现的电路可以被代替软件指令使用或者与软件指令组合使用以实现本发明。因此,本发明不局限于硬件电路和软件的任何特定组合。

[0072] 虽然前面的详细描述描述了本发明的优选实施例,但是要了解到上述描述仅仅是说明性的,不限制所公开的发明。本领域技术人员可以识别其它可替换的实施例,并且所有这样的实施例都被认为落在本发明的范围内。

102 罗马字音译	104 汉字字符	106 关联项
ken	刀	Sword/Sabre/Blade/Bayonet/Sting/Clock Hand/Kill Bill/ Jet Li
ken	件	Matter/Case/Item/Forms/Fact/Application
404 ken	406 券	Ticket/Coupon/Bond/Certificate/Speeding/Form/ Concert/Traffic
ken	县	Prefecture/Official/Government/Region/Administration/ Kyoto/Department
ken	108 权	Right/Morally Good/Power/Authority/Influence/Wrong/ Privilege/Moon
ken	験	Test/Exam/School/Anxiety/Case/Grades

图 1



图 2

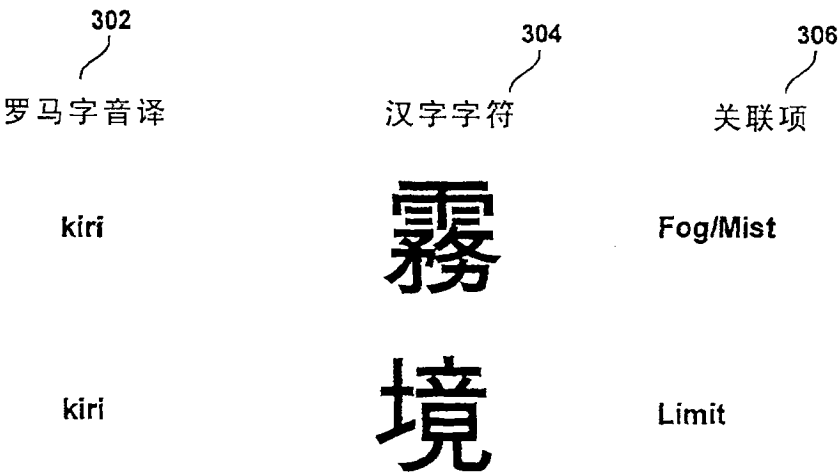


图 3

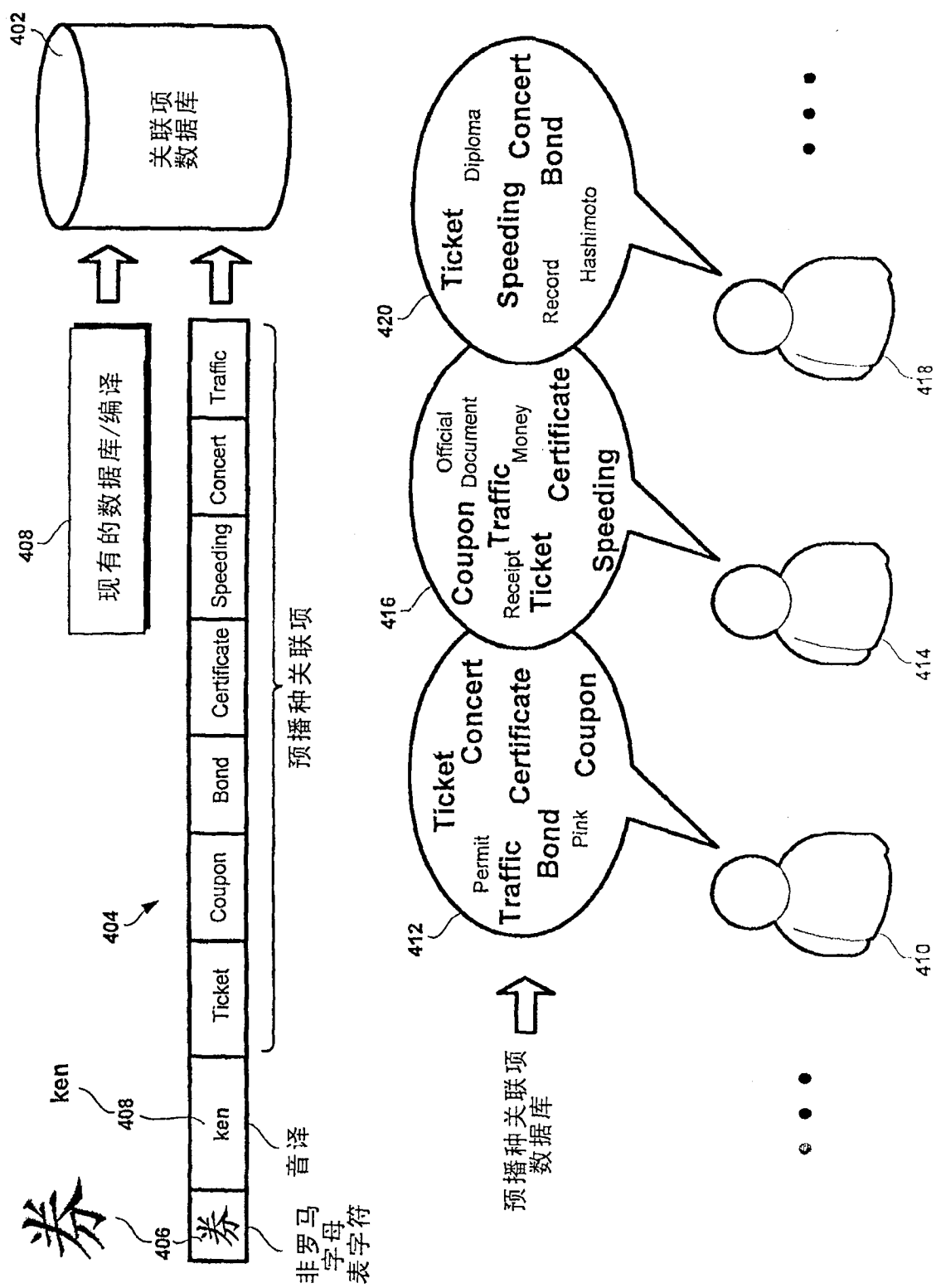


图 4

上海话汉语											508
广东话汉语											506
普通话汉语											504
日语											502
符号	罗马字	关联项1	关联项2	关联项3	关联项4	关联项5	关联项6	关联项7	● ● ●	关联项n	
望	kitai	Hope	Anticipation	Expectation	Wish	Spring	Dream	Cancer			
危	kitai	Danger	Jeopardy	Distress	Fear	Keep Out	Godzilla				
体	kitai	Fuselage	Body	Airplane	Airplane	Wreck	Airframe				
气	kitai	Gas	Vapor	Air	Car	Explode	CO2				
刀	ken	Sword	Sabre	Blade	Bayonet	Sting	Clock Hand	Kill Bill	Jet Li		
件	ken	Matter	Case	Item	Form	Fact	Application				
券	ken	Ticket	Coupon	Bond	Certificate	Speeding	Form	Concert	Traffic		
景	ken	Prefecture	Official	Gov't	Region	Admin	Kyoto	Department			
权	ken	Right	Power	Authority	Influence	Wrong	Privilege				
验	ken	Test	Exam	School	Anxiety	Case	Grade				
●											
●											
●											

图 5

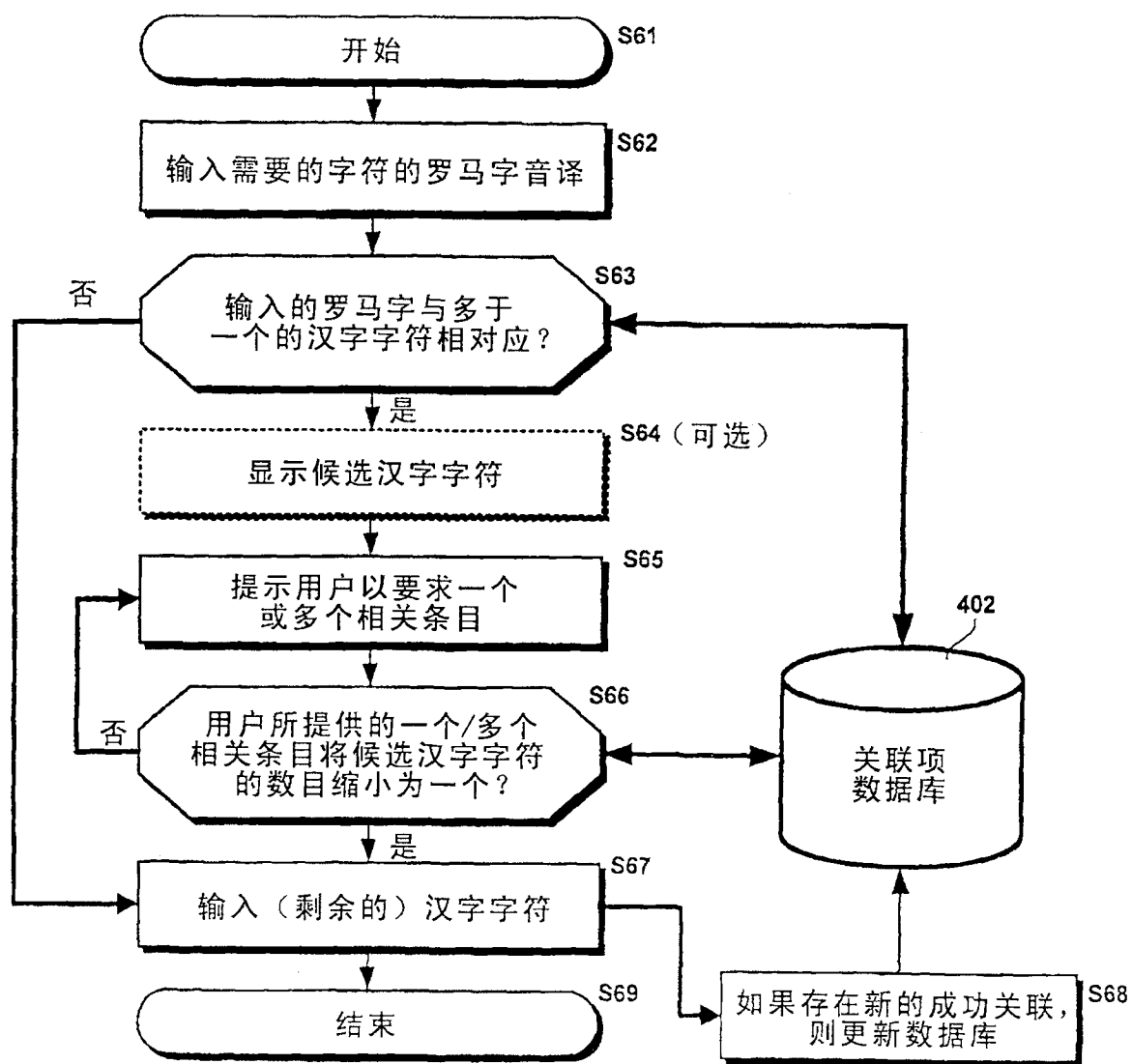


图6

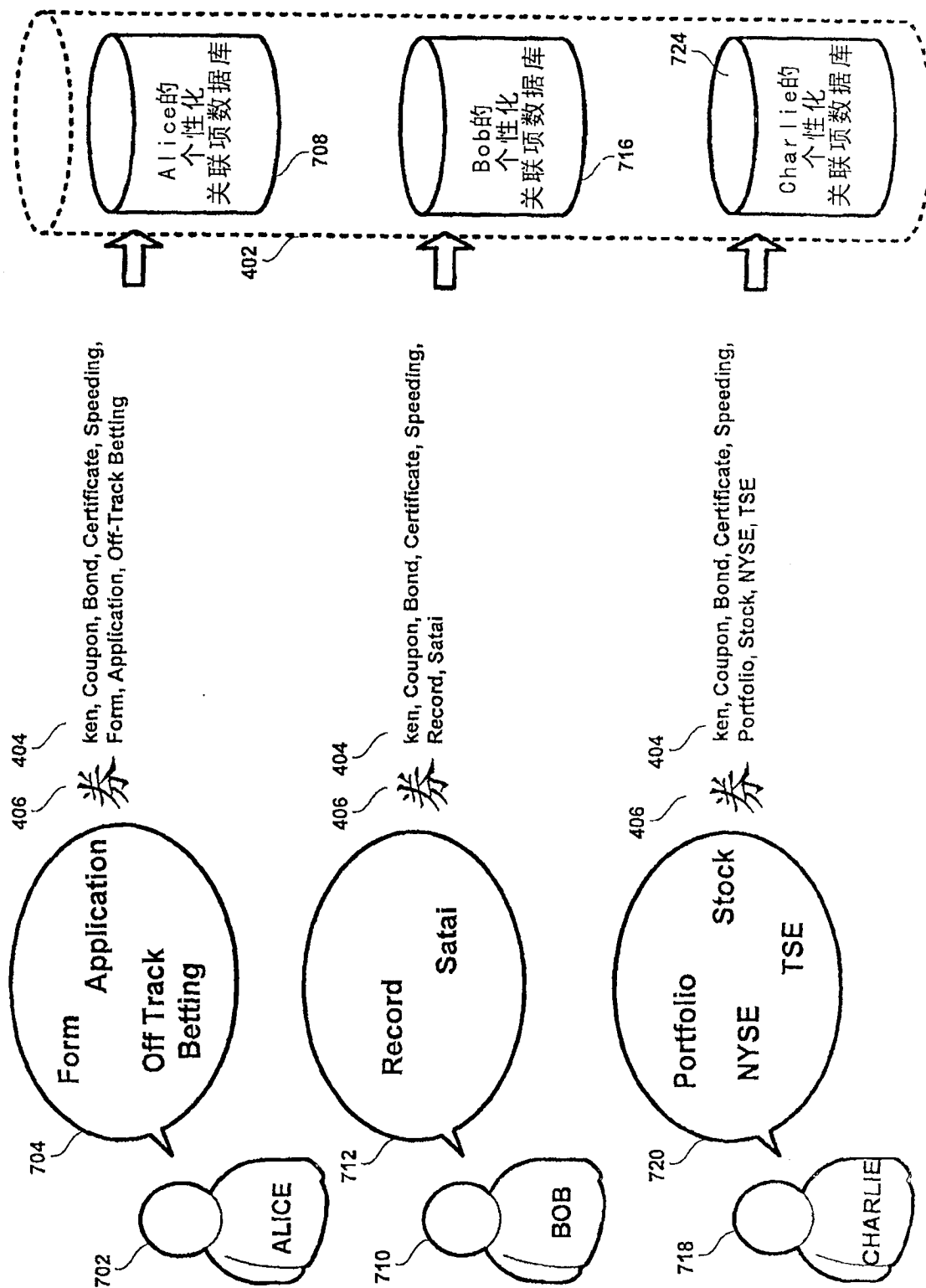


图 7

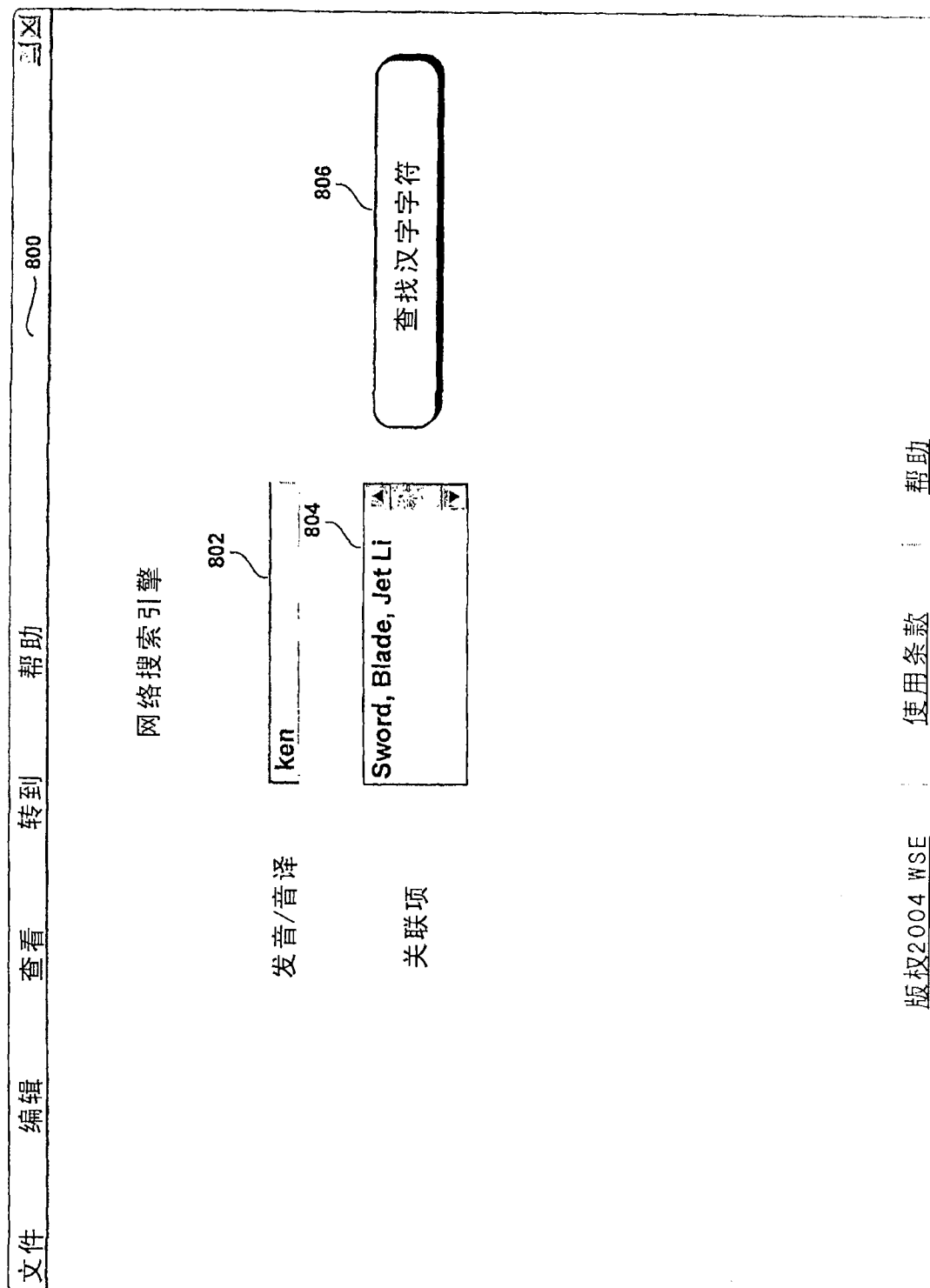


图 8

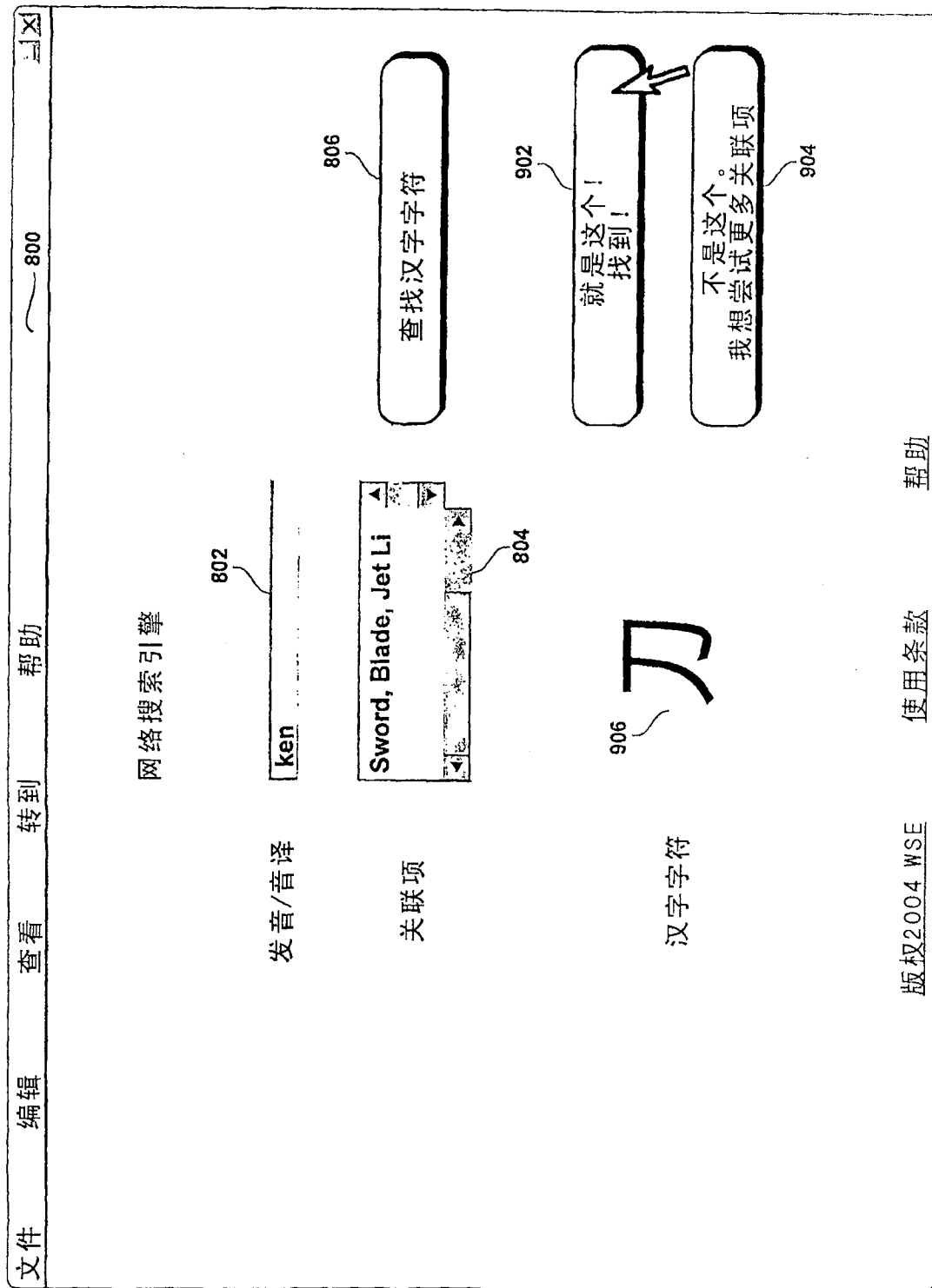


图 9

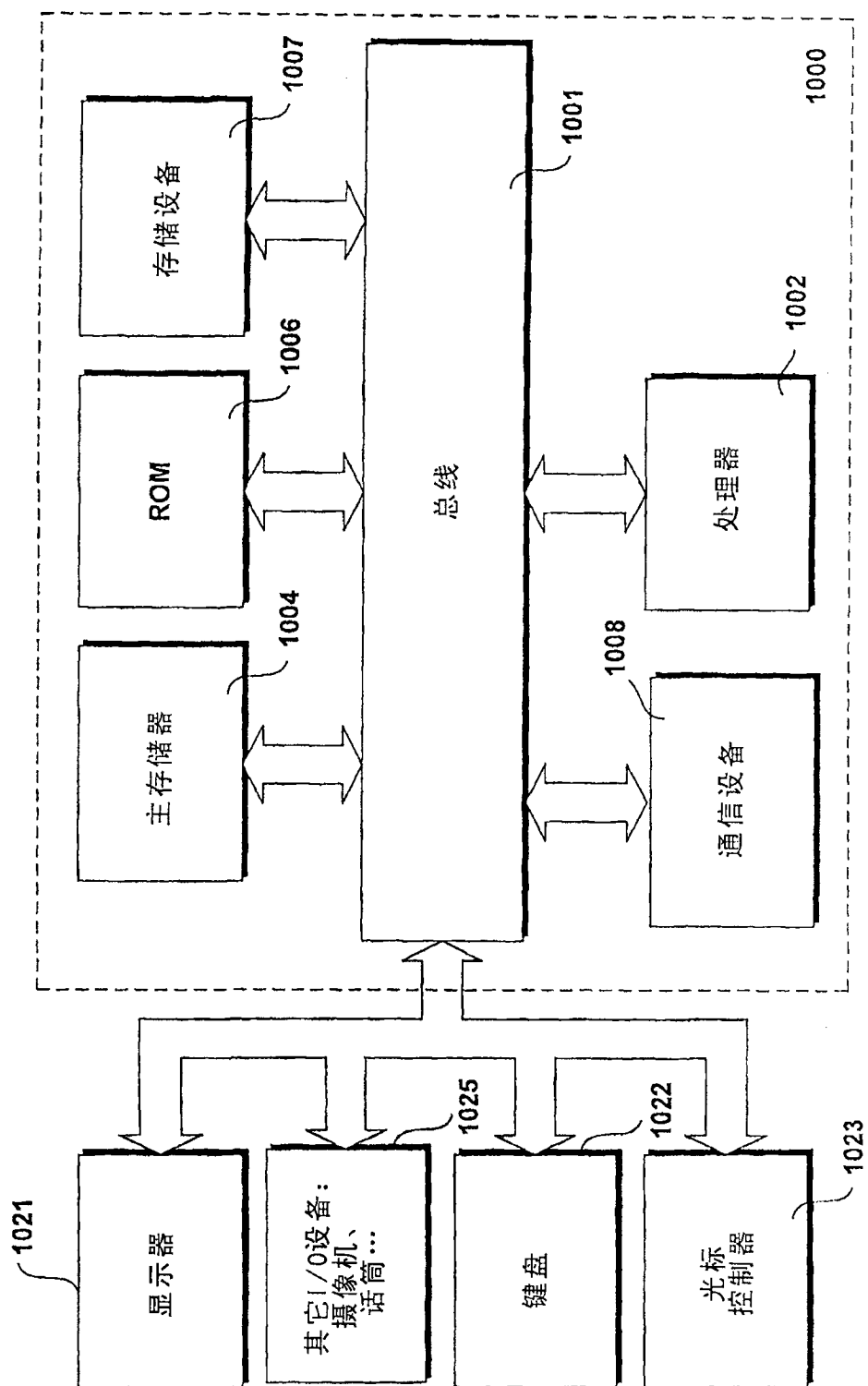


图 10