



(12) 发明专利

(10) 授权公告号 CN 101137980 B

(45) 授权公告日 2016.01.20

(21) 申请号 200580046617.6

代理人 宁晓 郑霞

(22) 申请日 2005.12.28

(51) Int. Cl.

(30) 优先权数据

G06F 17/30(2006.01)

60/640,872 2004.12.29 US

G06F 15/173(2006.01)

11/319,928 2005.12.27 US

(85) PCT国际申请进入国家阶段日

(56) 对比文件

2007.07.16

CN 1096038 C, 2002.12.11, 说明书第3页5行至第5页23行.

(86) PCT国际申请的申请数据

CN 1468403 A, 2004.01.14, 全文.

PCT/US2005/047235 2005.12.28

US 6493703 B1, 2002.12.10, 说明书第4栏62行至第8栏11行、图4-6.

(87) PCT国际申请的公布数据

W02006/071931 EN 2006.07.06

CN 1527976 A, 2004.09.08, 说明书第6页22行至第9页27行.

(73) 专利权人 贝诺特公司

审查员 魏峰

地址 美国加利福尼亚州

(72) 发明人 斯科特·布雷夫 罗伯特·布拉德肖

杰克·贾 克里斯托弗·明森

(74) 专利代理机构 北京安信方达知识产权代理

有限公司 11262

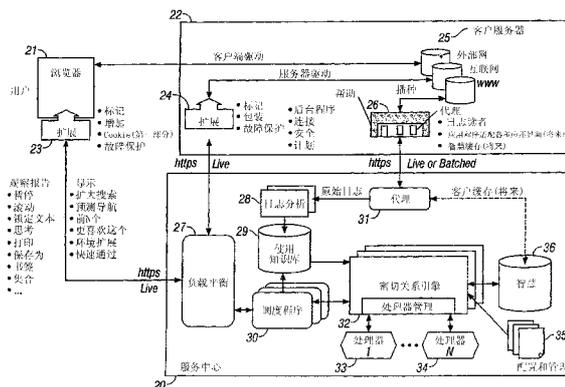
权利要求书4页 说明书32页 附图12页

(54) 发明名称

识别、提取、捕获和均衡专业技术和知识的方法和装置

(57) 摘要

本发明包括一组显著改进企业搜索和导航结果的补充技术。本发明的核心在于被称为跟踪网站访问者的行为的用户排序的专业技术或知识索引。专业技术索引设计为集中在企业属性的四个关键发现：主题权威、工作模式、内容新鲜和组关键技术。本发明产生有用、及时、交叉应用、基于专业技术的搜索和导航结果。相比之下，例如反向索引、NLP 或分类学的传统信息检索技术用与企业需要相反的一组属性：内容群体、字模式、内容存在和统计趋势来处理相同的问题。总的来说，本发明包含 Baynote 搜索 - 对现有 IR 搜索的加强、Baynote 引导 - 一组团体驱动的导航以及 Baynote 观点 - 访问者兴趣和趋势以及内容差距的总和观点。



CN 101137980 B

1. 一种用于自动确定在线团体中每个在线资源的主题 / 环境的计算机实现方法, 并且所述方法执行以下步骤:

通过所述在线资源的用户的目标团体或子组观察所述在线资源的使用模式;

通过观察与所述在线资源的所述使用模式相关的用户隐式动作以及通过从所述观察提取行为模式, 识别在线资源的有用性;

根据环境提炼所识别的在线资源的有用性;

根据观察的使用模式, 给每个术语和短语分配一个针对所述资源的术语向量项, 该术语向量项描述了识别的资源主题与每个所述术语和短语的密切关系到何种程度;

根据从团体导航和使用模式得到的资源间的得知的关联性, 提炼所述术语向量项的排序;

使用所述术语向量项的排序用于以下任一项:

确定和描述所述资源的主题;

确定和描述主题间的固有关系;

确定和描述资源间的固有关系;

确定和描述用户的当前兴趣;

确定和描述用户的当前兴趣和每个资源间的相似性;

所述观察的使用模式包括用户在线搜索、导航和交互行为, 所述行为包括以下任一种: 执行的搜索及在用户踪迹中定位; 查看的资源及在用户踪迹中定位; 在资源上暂停、搜寻、滚动、思考时间以及鼠标移动; 在资源文本上使用的锚定和线; 虚拟书签和虚拟打印; 以及显式下载、收发电子邮件、打印、保存和移走。

2. 根据权利要求 1 所述的方法, 其特征在于, 还包括以下步骤:

基于所述观察的使用模式而在搜索或导航环境中对文档排序。

3. 根据权利要求 1 所述的方法, 其特征在于, 还包括以下步骤:

采用向量空间模型和使用频率——反向文档频率权重表示资源主题。

4. 根据权利要求 3 所述的方法, 其特征在于, 还包括以下步骤:

在每个用户的基础上保持资源主题; 以及

通过组合个人的用户评估而确定组主题评估。

5. 根据权利要求 4 所述的方法, 其特征在于, 还包括以下步骤:

采用向量空间模型和使用频率——反向文档频率权重表示主题评估; 以及采用线性算法而组合用户主题评估。

6. 根据权利要求 1 所述的方法, 其特征在于, 还包括以下步骤:

采用向量空间模型和使用频率——反向文档频率权重表示资源主题; 以及基于余弦相似性度量而对所述资源主题给出相似性分数。

7. 根据权利要求 1 所述的方法, 其特征在于, 还包括以下步骤:

采用基于时间的现象来贡献有用性的确定, 所述现象包括资源的新鲜性、资源的重新发现、基于趋势的使用和资源的周期性使用中的任一个。

8. 根据权利要求 7 所述的方法, 其特征在于, 还包括以下步骤:

通过检测和消除吸引人的没用资源而提炼有用性, 这些资源即使没有用也表现大量的活动性。

9. 根据权利要求 8 所述的方法,其特征在于,还包括以下步骤:

吸引人的没用资源的检测以已经接收的用户对资源的重新访问次数以及用户访问的次数和对所述资源采用的所有活动为基础。

10. 根据权利要求 1 所述的方法,其特征在于,还包括以下步骤:

为隐式观察报告的重要性提供超过提供给显示观察报告的权重的非常高的权重。

11. 根据权利要求 1 所述的方法,其特征在于,还包括以下步骤:

分析所有观察报告;以及

经由所述分析,产生一组包括从用户的团体提炼的经验的推荐;

其中所述推荐随着时间老化并且如果他们具有相对小的价值则被丢弃;以及

其中将以重复的使用为基础的最有价值的推荐存储在长期存储器中。

12. 根据权利要求 1 所述的方法,其特征在于,还包括以下步骤:

对于可能是匿名的给定用户,所述用户访问特定网站,以及对于包括用户在什么页面和用户如何到达该页面中的任一个的给定环境,向所述用户提供允许所述用户更有效地导航到所述网站的推荐。

13. 根据权利要求 1 所述的方法,其特征在于,还包括以下步骤:

产生一组可以应用于搜索的推荐;以及

对于可能是匿名的给定用户,以及对于给定的搜索查询,采用所述推荐来提炼和扩大产生的搜索。

14. 根据权利要求 13 所述的方法,其特征在于,还包括以下步骤:

不仅由个人的使用而且由团体的使用驱动所述推荐,均衡群众的智慧和团体出现行为。

15. 根据权利要求 1 所述的方法,其特征在于,还包括以下步骤:

基于信息环境而识别包括对等体组和专家组中任一个的团体;

其中团体是嵌套的并且由不同级别的环境限定。

16. 一种用于自动确定在线团体中每个在线资源的主题/环境的密切关系引擎装置,包括:

处理器,用于通过所述在线资源的用户的目标团体或子组观察所述在线资源的使用模式;

所述处理器通过观察与所述在线资源的所述使用模式相关的用户隐式动作以及通过从所述观察提取行为模式,识别确定在线资源的有用性;

所述处理器根据环境提炼所识别的在线资源的有用性;

根据观察的使用模式,所述处理器给每个术语和短语分配一个针对所述资源的术语向量项,该术语向量项描述了识别的资源主题与每个所述术语和短语的密切关系到何种程度;

所述处理器根据从团体导航和使用模式得到的资源间的得知的关联性,提炼所述术语向量项的排序;

所述处理器使用所述术语向量项的排序用于以下任一项:

确定和描述所述资源的主题;

确定和描述主题间的固有关系;

确定和描述资源间的固有关系；

确定和描述用户的当前兴趣；

确定和描述用户的当前兴趣和每个资源间的相似性；

所述观察的使用模式包括用户在线搜索、导航和交互行为，所述行为包括以下任一种：执行的搜索及在用户踪迹中定位；查看的资源及在用户踪迹中定位；在资源上暂停、搜寻、滚动、思考时间以及鼠标移动；在资源文本上使用的锚定和线；虚拟书签和虚拟打印；以及显式下载、收发电子邮件、打印、保存和移走。

17. 根据权利要求 16 所述的装置，其特征在于，还包括：

所述处理器用于基于所述观察的使用模式而在搜索或导航环境中对文档排序。

18. 根据权利要求 16 所述的装置，其特征在于，还包括：

所述处理器用于采用向量空间模型和使用频率——反向文档频率权重表示资源主题。

19. 根据权利要求 18 所述的装置，其特征在于，还包括：

所述处理器用于在每个用户的基础上维持资源主题；以及

所述处理器用于通过组合个人用户评估而确定组主题评估。

20. 根据权利要求 19 所述的装置，其特征在于，还包括：

所述处理器用于采用向量空间模型和使用频率——反向文档频率权重表示主题评估；

以及

所述处理器用于采用线性算法组合用户主题评估。

21. 根据权利要求 16 所述的装置，其特征在于，还包括：

所述处理器用于采用向量空间模型和使用频率——反向文档频率权重表示资源主题；

以及

所述处理器用于基于余弦相似性度量而为所述资源主题给出相似性分数。

22. 根据权利要求 16 所述的装置，其特征在于，还包括：

所述处理器用于采用基于时间的现象来贡献有用性的确定，所述现象包括任何资源的新颖性、资源的重新发现、基于趋势的使用和资源的周期性使用。

23. 根据权利要求 22 所述的装置，其特征在于，还包括：

所述处理器用于通过检测和消除吸引人的没用资源而提炼有用性，这些资源即使没有用也表现大量的活动性。

24. 根据权利要求 23 所述的装置，其特征在于，还包括：

所述处理器用于使吸引人的没有资源的检测以已经接收的用户重新访问资源的数量以及用户访问的数量和对于所述资源的所有活动为基础。

25. 根据权利要求 16 所述的装置，其特征在于，还包括：

所述处理器用于为隐式观察报告的重要性提供超过提供给显示观察报告的权重的非常高的权重。

26. 根据权利要求 16 所述的装置，其特征在于，还包括：

所述处理器用于分析所有观察报告；以及

经由所述分析，所述处理器用于产生一组包括从用户的团体提炼的经验的推荐；

其中所述推荐随着时间老化并且如果他们具有相对小的价值则被丢弃；以及其中以重复的使用为基础的最有价值的推荐被存储在长期存储器中。

27. 根据权利要求 16 所述的装置,其特征在在于,还包括:

对于可能是匿名的给定用户,所述处理器用于所述用户访问特定网站,在包括任何用户在什么页面和用户如何到达该页面的给定环境中,向所述用户提供允许所述用户更有效的导航所述网站的推荐。

28. 根据权利要求 16 所述的装置,其特征在在于,还包括:

所述处理器用于产生一组可以应用于搜索的推荐;以及

对于可能是匿名的给定用户,以及对于给定的搜索查询,所述处理器用于采用所述推荐来提炼和扩大产生的搜索。

29. 根据权利要求 28 所述的装置,其特征在在于,还包括:

所述处理器用于不仅由个人的使用而且由团体的使用而驱动所述推荐并且均衡群众的智慧和团体出现行为。

30. 根据权利要求 16 所述的装置,其特征在在于,还包括:

所述处理器用于基于信息环境识别包括对等体组和专家组任何之一的团体;其中团体是嵌套的并且由不同级别的环境限定。

## 识别、提取、捕获和均衡专业技术和知识的方法和装置

### 技术领域

[0001] 本发明涉及对信息的电子访问。尤其是,本发明涉及一种用于识别、提取、捕获和均衡专业技术和知识的方法和装置。

### 背景技术

[0002] 多年来,企业一直在与低效率的搜索技术进行斗争。相比经由例如 Google 和 Yahoo 的服务在公网上可以获得的内容,对于企业内部内容仍然缺乏高相关的搜索解决方案。在各公司中具有十几到上百的独立应用程序和知识信息库,找到重要的商业信息每年要花费几千亿美元 [来源:A. T. Keamey]。目前, CIO 和商业执行官再次将企业搜索作为在接下来几年的商业 /IT 业中的首要挑战之一。

[0003] 企业搜索需求的满足很糟糕。已经开发了各种搜索技术来应对搜索网络、搜索个人用户计算机 (PC/ 桌面) 和搜索内部商业文档 (企业) 的挑战。所有这些方案都是独特的,但是没有一个方案能为企业提供充分的解决方案。PC (桌面) 搜索

[0004] PC 或桌面搜索可以比作在凌乱的车库中找东西。你知道将它放在了某个地方但是就是找不到。因此确定位置是唯一的目标。并且当你找到它的时候,你是唯一的鉴定人来决定是否你真正找到了正确的内容或文档,因为是你首先收集或写下该内容。这件事上你是唯一的专家和权威。

[0005] 微软的传统 PC 搜索基于在搜索时解析文档。这样很慢而且只能找到在例如文件夹或电子邮件目录等独立地方中的内容。最新的 PC 搜索从 Google 中引入反转索引技术,并且很快 Yahoo、Ask Jeeves 和 Microsoft 也会变得可用。他们开始解决速度和存储空间问题,从而用户可以在个人文档系统、Outlook 或电子邮件系统、日历和其它桌面环境中找到信息。

[0006] 网络搜索

[0007] 搜索的另一系列是网络搜索。这里,该故事更像第一次在波士顿驾车。你不必是你正在寻找主题的专家并且你不必学习新的课题。有时,你搜索得到例如天气、旅游或购物等新的服务。通过网络搜索,你依靠网上数百万人的帮助你并且你不需要知道或在乎谁是真正的专家或权威。因此,有时你会得到糟糕的建议或可能在错误的地方购物。

[0008] Google 之前的网络搜索仅仅依靠例如反转索引、自然语言处理 (NLP) 和数据库索引的技术。它们都还不错,但是当计算指向网页的链接数目时它就不够好。随着多个网址链接到你的网页,你的网页变得很重要,仅仅是因为网站后面的网络管理员经历了将那些额外的链接添加到你的网页的麻烦。因此,诞生了网页排序 [tm] 和 Google 的商业成功。

[0009] 企业搜索

[0010] 然而,企业不像 PC 或网络环境。想像你在寻找书籍来学习 Java 编程,你知道你的最终目标但是有许许多多关于 Java 的书籍,我应该读哪一本;这必须非常正确。因此,发现处理找到正确的参考内容或者公司的其它专家知道的信息。对于企业,好的搜索结果的最终判断不仅仅在你自己。好结果的这些仲裁者可以是你的对等者 (peer) 或你依靠来做你

的工作的专家。

[0011] 图 1 示出了企业搜索问题的一个示例,其示出了在企业搜索领域的状态流程图。在图 1 的示例中, Dave 在搜索特定信息并且找到 2800 个文档。Dave 在返回的前十个结果中没有发现有用的结果,因此 Dave 给 Sam 打电话。Sam 接着搜索并且也什么都没找到,于是给市场部发电子邮件。市场部的 Mark 和 Tina 搜索并且也没有找到什么。Mark 给 Eric、Nancy、和 Ganesh 再打电话,并且在 Ganesh 的设计文档中找到了答案。Tina 再给 Eric、Nancy 和 Ganesh 打电话,而现在每个人都心烦意乱。明显的,如果 Dave 在他的开始搜索中就发现了 Ganesh 的设计文档,这将对 Dave 非常有用。事实上,该文档可能已经在 2800 个文档之中,但是 Dave 不可能识别该最有用的文档。

[0012] 传统的企业搜索技术采用反转索引、NLP 和数据库索引方案(见图 2)。主要问题是目前的引擎对每个查询都会返回成百上千个搜索结果给用户。任何看上去像 Java 或编程的都混合到一起给你看。很像垃圾邮件,搜索引擎用大量、过时、无关、非正式、冗长的(silo)、矛盾的和未授权的结果打扰用户。用户快速放弃而采取包括电话、电子邮件、聊天等更昂贵的方式来得到信息,或者更糟,开始重新产生、补充或放弃已经存在的信息。

[0013] 企业搜索显示出一组独特的特征

[0014] 通过比较企业搜索和网络或 PC 搜索的关键问题,可以得出结论:企业搜索与网络搜索的对照是独特而直接的。事实上,网络搜索的原理不会也不能用于企业搜索,反之亦然。在这点上要考虑五个关键属性:搜索引导、用户行为、内容的新鲜性和可信性、用户同质性和保密考虑。

[0015] 主要引导

[0016] 例如,在网络上,Google 的成功取决于以网页排序作为主要引导。虽然网页排序对于在网络中提供一些充分有力的判断很有效,但是对于企业内容搜索不会发生相同的效果。首先,企业内容缺少需要提供网页排序引导效果的大量链接,并且企业也没有动力在持久性的基础上产生这些链接。第二,网页排序的真正目的是找到人类努力的足迹以间接地显示主题权威,因为几乎不可能在巨大的网络世界中找到真正的专家。对于企业,你应该不需要间接地猜测谁可能是专家,你知道可信任的专家是谁,你雇佣他们,他们天天在公司作为他们领域内的专家而进行工作。对于相关的引导和排序,企业搜索应该依靠作为主题权威的他们。

[0017] 用户行为

[0018] 用户行为在企业和网络之间是完全不同的。作为网络上的个体,我们具有比我们可能知道的更多的脸孔。我们可能同时是男人、父亲、儿子、丈夫、兄弟、打高尔夫球的人、旅行者、摇滚音乐家、投资者和上百的其它身份。当我们在网络上搜索时,搜索是一次性的并且到处进行。同时,我们输入的关键词趋向搜索目标本身。当我们输入“天气”时,我们在寻找天气信息。网络上的用户反馈并不可靠,因为只有很小一组吵杂的用户有时间给出反馈,因此搜索结果被他们非代表性的偏见所歪解(这最后一句话如何联系到本段的其余部分?也许应该建立一小段来解释用户反馈中的偏见)。

[0019] 然而,企业搜索趋向于基于用户的角色和他所处的情况而快速重复其本身。当一个销售人员正在寻找一些销售担保时,其它在相同区域中负责相同产品的销售人员也很可能需要该相同的信息。同样重要的是,这个在其个人生活中可能具有 300 个角色和身份的

人具有很少的工作角色,例如,最多 6 个。他可能是在巴黎办公室工作的工程师,同时他是交叉功能文化委员会的成员。同样重要的是,需要注意到在企业搜索中的关键词更像是一个人在寻找的文档的暗示,甚至是鱼饵。据认为,百分之八十的人寻找他以前见过的信息。给出企业用户的可预见性,我们可以安全的依靠自激发的动作和行为来收集没有偏见的反馈。

[0020] 新鲜和可信性

[0021] 网络搜索对更旧的内容给予重视或排序更高。内容存在的时间越长,就越有可能被找到,这是因为其他人有时间发现并且连接到这项内容。

[0022] 企业需要表现不同。新鲜内容反映新的商业状况,并且因此必须被排序更高从而让更多人看到。通过对新鲜内容的快速响应,确保了商业敏捷性。一条一周前的内容可能比一年前的内容更好,除非这种情况,如果今天的内容可用并且表示与一周前的内容不同,则该一周前的内容完全不好。企业搜索用户不需要足够好的内容,他们需要最合适的搜索结果。

[0023] 同质性

[0024] 网络或消费者世界是非常异质的,而企业相反:其为同质的,或者更准确的说是分段同质的,这表明在一个公司中不同的部门或组(销售对市场对工程)可能不同(分组),但是在一个组内,无论人们的身份如何不同,他们工作的方式非常类似或者同质。

[0025] 这种分离属性的含意是深刻的。在一个包括上百万人的大的异质世界里,在理解人们喜欢什么、需要什么等等的努力中,统计是解决该问题的唯一的已知技术。网络搜索正确地依靠统计为用户找到不那么精确的信息。企业又是不同的。对于小的样本人口和同质组,统计不起作用。为了理解他们,你需要知道他们的喜欢和不喜欢。没有预测(我们的‘预测’是什么意思),只有知晓。

[0026] 通过对企业特征的理解,可以看到,企业搜索需要集中在主题权威性,重复的基于角色的工作模式、新鲜和正式的内容和组关键技术(一个组去做一件工作的集体知识和专家技术)。重新审视传统的基于 IR(信息获取)的搜索,我们认识到其集中在相反的方面。其依靠整个内容群体(爬行并且索引它),而不是主题权威,单词或语言学模式而不是工作模式,更早存在的内容而不是新鲜或正式的内容,统计趋势预测而不是组相似性了解。从而,需要集中在企业的正确关键特征的技术。

[0027] 企业搜索技术的问题对于很多 CIO 和商业执行官已经成为关键的技术。在发明人自己对十几个 CIO 和商业执行官的有限调查中,人们排序企业搜索优先权问题为 10 个中 9-10。传统的全文本引擎的挑战是相关性差。他们同时对一切(所有内容)都好也对什么都不好(无关结果)。NLP 技术通过集中于人类语言变得更确定的一个应用程序和一个域而获得了更好的相关性。当企业在运行成百上千的应用程序时,NLP 的问题是解决方案处于存储中并且仅在特定的应用程序中令人满意。雇员不可能逐一地登录到这么多系统以寻找信息。两类的解决方案也具有一旦使用就不适于改变的困扰。分类和结构在企业中随着时间会快速改变。

[0028] 目前的搜索软件也受到具有继承的昂贵的产品体系结构、设计和市场以及销售模式的传统企业模式的困扰。虑到软件许可、服务、培训和其它相关费用,典型的企业搜索研发将花费 \$500K 到几百万。

[0029] 因此,通过成本和搜索质量的改进,转换如何购买和研发企业搜索技术是有利的。

## 发明内容

[0030] 本发明通过均衡企业搜索所依赖的:公司内和公司外一个人的对等者和专家,解决现有技术企业搜索的状态局限性。本发明提供一种识别、提取、分析和使用专业技术排序来为用户产生个性化的、精确的搜索结果,从而他们不需要打电话、电子邮件等的系统。

[0031] 发明人已经公开了一组不同于例如 Verity、Autonomy、Endeca 和 Google 应用的所有现有的基于 IR(信息检索)的企业检索的唯一方法。发明人仔细分析了企业相比 Web 搜索环境的特征,并且从技术发展、学术研究和社会行为方面在相关学科中应用一组方法。本发明提供一种可以独自工作或经由 plug-n-play 接口通过最小的努力嵌入到其它应用程序中的技术。这个结果在搜索有用性、相关性、应用程序之间的搜索联合以及节省成本方面是巨大的改进。本发明的优选实施方式也均衡传统的搜索技术。

[0032] 本发明通过采取与传统搜索原理和技术完全相反的方法而提供了相关信息发现。如上所述,传统内容搜索技术和产品使用内容作为引导搜索的基础。其采用例如信息检索(IR)算法、自然语言处理(NLP)技术和规则、产品或结构分类学、或由链接数进行页面排序的技术。传统的数据搜索依靠在关键字或数据库行或列的号码而建立数据库索引。它沿用(crawl)和索引内容和数据,用字标签或短语产生反向的全文本索引或数据库索引,潜在地辅以分类学和页面排序以改进搜索结构。采用传统搜索技术的搜索结果是糟糕的,具有大量的低相关性的结果。对于很多商业过程,当搜索失败时,用户不得不寻找代替的、昂贵的获得信息的方法,这样或者花费用户大量的时间,或者更糟地涉及其他人来帮助寻找信息(见图1)。

[0033] 代替使用内容作为信息发现的开始点,本发明提供一种从人所在的企业开始的系统。毕竟,企业由具有专业技术和关键技术(know-how)的专门人才和专家构成。他们指导工作并且在逐个角色地频繁重复他们的工作模式。本系统检测和捕获存储在人们大脑中并且展现在他们日常行为中的专业技术和工作模式,并且产生基于行为的知识索引。然后知识索引依次用于产生专家引导的个性化信息。这个过程对专家本身是透明的,并且因此使用起来很有效而且非常经济。

[0034] 按照一个方面,本发明提供一种用于确定知识和/或专业技术的计算机实现方法,包括以下步骤:识别用户的目标团体或子组;通过所述用户的所述目标团体或子组观察网络文档和其它在线资源的使用模式;以及基于所述观察的使用模式,执行任何以下步骤:自动确定所述网络文档和其它在线资源的主题;以及检查哪个所述网络文档和其它在线资源对所述用户的目标团体或子组使用最多和有用。

[0035] 按照另一方面,本发明提供一种用于确定知识和/或专业技术的装置,包括:用于识别用户的目标团体或子组的设备;用于由所述用户的目标团体或子组观察网络文档和其它在线资源的使用模式的设备;以及基于所述观察的使用模式,用于实现任何以下功能的设备:自动确定所述网络文档和其它在线资源的主题;以及检测哪个所述网络文档和其它在线资源对所述用户的目标团体或子组使用最多和有用。

## 附图说明

- [0036] 图 1 示出了现有技术企业搜索的当前状态的流程图；
- [0037] 图 2 示出了传统的基于 IR 的搜索模型的流程图；
- [0038] 图 3 示出了根据本发明的系统结构的示意性方框图；
- [0039] 图 4 示出了根据本发明由嵌入式应用程序捕获的行为相关性的流程图；
- [0040] 图 5 示出了根据本发明的内嵌用户界面的屏幕快照；
- [0041] 图 6 示出了根据本发明采用 JavaScript 标签呈现的内嵌用户界面的屏幕快照；
- [0042] 图 7 示出了根据本发明的弹出式用户界面的屏幕快照；
- [0043] 图 8 示出了根据本发明的专家引导的个性化搜索交叉应用程序的示意性方框图；
- [0044] 图 9 是根据本发明的用户库的屏幕快照；
- [0045] 图 10 是根据本发明的用户库的第二屏幕快照；
- [0046] 图 11 是根据本发明的用户库的第三屏幕快照；
- [0047] 图 12 示出了根据本发明的文档推荐的流程图；
- [0048] 图 13 示出了根据本发明的增大搜索的流程图。

### 具体实施方式

[0049] 对信息发现采用专业技术和行为

[0050] 本发明包括一组显著改进企业搜索和导航结果的补充技术。本发明的核心在于专业技术或知识索引,也称为专业技术知识库,其产生网站和网络应用程序访问者的观察报告。专业技术索引的设计集中在企业搜索的四个关键发现:主题权威、工作模式、内容新鲜和组关键技术。本发明产生相关的、及时的、交叉应用的、基于专业技术的搜索结果。相反,例如反向索引、NLP 或分类的传统信息检索技术用与企业需要的内容群体、单词模式、内容存在和统计趋势相反的一组属性来处理相同的问题。

[0051] 本发明的另一实施方式使得新技术在现有的企业应用程序和知识库环境中明显地起作用,从而不需要用户培训或采用新的界面。其还支持所有遗留的全文本或 NLP 搜索技术,例如 Verity、Autonomy、Endeca 和 Google 应用。事实上,其就是在那些技术之上工作并且采用他们的基本结果作为提炼的基础。

[0052] 本发明的第三实施方式来源于调整开放资源技术,例如 Lucene,用于建立可升级网络查询引擎,将所有信息源的空间和索引绑定到一组有意义的结果中。

[0053] 嵌入应用程序 UI 以捕获行为相关性

[0054] 本发明本身经由对搜索结果界面的简单改变而嵌入到任何现有的网络应用程序中,例如 www、CRM、ERP 和入口等。对于非网络应用程序,通过插入 SOA(面向服务体系结构)桩代码,从而可以由专业技术索引对搜索交通进行检查和重排序而完成类似的工作。此外,可以通过本发明配置不仅是搜索结果页面的任何网络页面以提供主动的引导,而不需要用户输入查询。

[0055] 依靠自身兴趣

[0056] 本发明不需要用户明确的表决、提供反馈或利用通常导致合作过滤的其它机制。其依靠人们独自的做其正常的工作并且留下他们需要和偏好的证据的踪迹以完成他们的工作。依靠自己比传统的用户被指导为其它人表决的合作过滤完全不同并且为更可靠的引导。当要求用户给出反馈时,多数人由于缺少时间或它不是优先的而不去做。当被迫时,人

们不加考虑地快速勾选复选框,因此会误导使用这些数据的人们。存在一小组人们喜欢填写调查表并给出反馈,但是他们经常是用户群组中自由表达的、评论的和不具有代表性的样本。Amazon 和 eBay 二者在采用传统的合作过滤技术来完成相似性排序的过程中都有负面的经历。

#### [0057] 隐式相关性动作

[0058] 本发明允许用户动作追踪的逐个应用程序的配置。隐式动作按键(下面讨论)作为搜索结果的一部分进行嵌入以捕获在用户做其工作时的用户意图和偏好的重要暗示。例如,当发现相关内容时,公共入口可以给出用户“查看”、“下载”、“打印”和“电子邮件”按键作为动作来反映他们的意图。“查看”可能是较弱的指示,而其它是偏好的较强指示。本发明开发了额外的用强烈的信心预测访问者意图的隐式观察报告。这些观察报告包括检测考虑时间、虚拟书签、虚拟打印和虚拟电子邮件的能力。在所有情况中,访问者没有针对内容执行书签、打印或电子邮件,但是他们将该内容保持在计算机屏幕上很长时间,即,足够长的时间来用这些内容作为工作的参考。这些观察报告在其真正对团体有用之前,在对等者和专家当中被交叉检查。

#### [0059] 显式相关性动作

[0060] 至少可以添加两个额外的显式按键以跟踪用户行为的清楚暗示。“保存到库”指示给定查询的内容的强烈、明显认可,而“移走”或“降级”表示对给定查询和角色的内容的强烈不喜欢。库是虚拟的并且不会物理地存在于浏览器或者甚至 PC 上。其为主要用户行为跟踪目标或日志。此外,显式相关性排序高于隐式相关性,但是二者都由一个每用户库对象管理。

#### [0061] 搜索垃圾控制

[0062] 人们熟悉垃圾邮件。当前的搜索显示类似于垃圾,响应于简单的查询会返回成千上万的结果,并且很多结果与用户所要寻找的完全无关。在考虑到雇员扮演的角色之后,通过使用“移走”动作,本发明的系统可以以类似于垃圾邮件报告机制的方式,对与一组相同角色的用户相关很小或不相关的结果进行识别并且降级。例如,如果三个工程师移走他们对一个文档的兴趣,则其它类似的工程师应该看不到这个较高级别的文档,而销售人员可以仍旧对该文档给出高的级别。

#### [0063] 应用程序之间的专家引导、个性化搜索

[0064] 在整个企业的环境中考虑本发明。虽然用户动作通过其 PC 上的商业应用界面完成,但是被称为“我的库”的动作日志会在系统服务器中存储和维护。该日志对 PC 和涉及的应用程序都不需要关心。

#### [0065] 我的库 / 行为日志

[0066] 这是每个用户对象。其对于用户一般都是不可见的,除非有权利用户或应用程序想要直接使用它们。可以在系统生效之前挖掘或者学习库中的数据。其不断改善自己,调整并且适应内容和其查询的真正商业使用。该库存储用户身份和属性、所有查询、相关内容 URI、对所有相关内容和数据的一个或多个索引、查询缓存、内容和数据缓存、访问时间、个性化的排序公式、近似散列法和对考虑的保密政策的装载比率控制。

[0067] 可以添加我的库的桌面版本以提供内容缓存、内容推进 / 更新 / 警告,并且断开内容访问。

[0068] 域专业技术网络

[0069] 本发明的一方面涉及检查多个个人库或行为日志。当企业从对等者或专家范围开始分析很多人的日志时,会出现大量关于信息消耗和雇员生产力的观点。

[0070] 对等者在这里被定义为一组具有共同兴趣的用户,例如产品、主题组、事件、工作角色、位置等等。

[0071] 专家在这里被定义为一组比查询或浏览的人具有不同知识和技巧的访问者,但是查询和浏览的人依靠这些专家而有效做工作。

[0072] 例如,一个工程师具有其它工程师的对等者组,还有由产品经理、销售人员、一些消费者、人力资源 (HR) 人员和办公室助理组成的专家组。当环境改变时,对等者和专家可以改变。工程师约翰可以扮演超出他所在组织的角色。他可以是交叉功能的委员会成员,并且物理上在伦敦办公室工作。因此约翰有其作为一部分的三个环境关系。这种环境这里被称为域专业技术网络或 DEN。一个雇员可以属于几个 DEN。下面将详细讨论各种类型的 DEN。

[0073] 体系结构

[0074] 专业技术索引系统也被称为专业技术和行为知识库。该要素对本发明很关键。其是具有采用 Web 服务、XML、J2EE 和其它基础技术的面向服务体系结构 (SOA) 的基于服务器的系统。

[0075] 以下构件是本发明的关键部分。

[0076] 行为仪器:也被称为工作监视器,该要素负责灌输和记录在各种商业应用程序上的用户行为。该应用程序搜索形式为本发明实现的很多观察记录张贴之一。浏览器和应用程序导航、文档上传、Web 插件、页面标签、电子邮件服务器和客户集成、内容管理、文档管理、记录管理以及合作系统都是仪器的共同位置。本发明还及时返回并且解析公共日志文档、例如 Web 服务器日志、查询日志、例如 LDAP 的目录文档,以构建和提取历史或基本水平专业技术。

[0077] 实时行为日志:该要素为上述的每用户对象。

[0078] 域专业技术网络:该要素为上述的工作关系对象,其连接个人、对等者和专家协会,并且记录重复的基于角色的企业工作模式。

[0079] 不一致的网络索引 (NUNI):该要素提供最相关的、及时而权威的搜索结果。这将在下面详细讨论。

[0080] 环境映射和动态导航:在个人、对等者和专家日志的帮助下,NUNI 索引不仅可以产生好的搜索结果,而且提供用户没有经由其搜索查询或关键字直接询问的额外的环境信息。该环境结果可以在搜索结果侧条中,或者作为个性化动态导航的一部分而呈现给用户。下面将详细讨论动态导航。

[0081] 生产力报告和解决方案:该要素基于行为日志和 NUNI 索引而产生各种报告。

[0082] 总之,专业技术索引系统集中于企业主题权威、工作模式、内容新鲜和组关键技术以传递专家引导的个性化信息。

[0083] 技术概览

[0084] 图 3 示出了本发明优选实施方式的系统体系结构的示意性方框图。下面提供该体系结构各方面的详细讨论。该体系结构包括服务器中心 20、客户企业 22 和用户浏览器 21。

用户浏览器设置有扩展 23,其访问客户企业 22 处的客户服务器 25 并且经由负载平衡器 27 访问服务器中心 20。目前采用 HTTP 协议实现与服务器中心的通信。用户按照企业协议访问客户服务器。下面将详细讨论浏览器扩展 23。

[0085] 本发明的显著之处在于提供故障保护。构建扩展 23 以及企业扩展 24 从而如果服务器中心 20 没有以成功的方式响应,则关闭扩展而企业和浏览器以正常方式互相作用。本发明的特征仅仅在服务器是活动的并且正确执行其操作的情况下提供。因此,对于企业的用户,服务器的失败不以任何方式损害企业的运行。

[0086] 如上所述,也为与服务器中心 20 的负载平衡器 27 经由 HTTP 协议通信的企业提供扩展 24。

[0087] 企业还包括经由代理 31 采用 HTTP 协议与服务器中心通信的帮助器 26。代理从企业获取日志信息并且将其提供给日志分析器 28,日志分析器 28 产生提供给使用知识库 29 的结果。在密切关系引擎 32 和浏览器以及企业之间经由各种发报机 30 交换信息。浏览器本身响应于来自其的搜索查询而向服务器提供观察报告并且接收显示。在下面将详细讨论这些观察报告和显示。

[0088] 本发明的一个关键特征是密切关系引擎 32,其包括多个处理器 33/34 和配置管理设备 35。在本发明的运行期间,也被称为知识的一条信息被收集在知识数据库 36 中。下面将详细讨论密切关系引擎 32 的操作。

[0089] 本发明系统通常作为基于由卖主,例如 Verity、Autonomy、Google 等提供的传统引擎而对现有的搜索系统的加强而构建。基于传统技术的内容和数据搜索系统被称为现有搜索机制。

[0090] 本发明系统实施为对现有搜索机制的包装。当用户发出搜索查询时,由系统首先处理该查询。该系统通常向现有搜索机制转发该查询。其还可以对其内部的索引和数据库的一个或多个搜索或相关操作。一旦从各种搜索中获得结果,将他们合并为单一一组结果。这些结果的实际表示在于客户的判断力,客户既可以从系统中取得原始结果数据并采用 JSP、CGI 或类似的机制表示,也可以使用系统提供的默认搜索结果页面,该默认搜索结果页面可能采用级联样式单或其它类似技术进行了客户化。

[0091] 结果中的各文档一般与用户对该文档采取的各种可能动作一起表示。可用的动作是站点配置,并且可以包括,例如,“考虑”、“查看”、“下载”、“电子邮件”或“打印”。当用户对于特定文档选择这些动作之一时会通知系统。然后使用该数据推断该特定文档与产生该文档的查询的相关性。这样,如果用户对于文档选择“查看”动作,该系统可能推断该文档对用户的那个查询具有某些实际价值,而如果用户选择例如“打印”或“下载”的更持久的动作,该系统可能推断该文档与用户高度相关。该系统可以检查虚拟打印或下载来给出与物理打印、下载或书签的精确近似性。该技术依靠于在一定时间内,例如一分钟,检查用户对浏览器的动作,其中文档保持长时间打开,即,长时间停留。另一方面,如果用户对查询结果根本没有执行任何动作,系统可能推断该结果与用户不相关。该数据被保留并且用于影响将来用户查询的结果,并且产生质量度量。

[0092] 库

[0093] 该系统保留各用户的内容参考和 / 或使用的库。该库还被称为行为日志。虽然该库不需要对用户可见,但是其在某种程度上类似于 Web 浏览器中的书签。事实上,用户可能

甚至意识不到它的存在。取决于系统如何配置,当从搜索结果中选择对于文档的某些动作时,该文档名及其位置会自动添加到用户的库中。文档也可以通过从搜索结果显式的“添加到库中”的选项而添加到用户的库中。用户的库中的文档参考的存在一般表示该文档对用户具有特殊兴趣。从而,如果查询的结果产生也出现在用户的库中的文档,则其排序通常被提高。

[0094] 在某些配置中,可能向用户的库中直接添加文档,而不用先在搜索结果中遇到。这样的文档不需要由现有搜索机制索引或者甚至访问。然而,因为其出现在用户的库中,如果其与查询匹配,则仍旧可以合并到最终的搜索结果中,并且因此在由系统产生的结果中可用。以这种方式发现的内容通常非常有价值并且因此在结果排序中被给予特殊的偏好。

[0095] 关系

[0096] 商业中的人们按照多个不同方式互相关联。例如,在一个组的对等者之间、上级和下属之间或者主题专家和搜索者之间存在关系。当对这些不同种类的关系进行建模和观察时,其显示出可以用于影响和提炼搜索结果的观点。例如,如果一组对等者的几个成员都发现一个特定的文档有用,则由于一个对等者组的成员通常具有类似的兴趣而使一个相同组的其它成员也发现该文档有用。类似地,如果某人在寻找关于特定主题的信息,则该主题的知名专家发现有用的文档将可能也对搜索者有价值。

[0097] 该系统为各用户保持一个或多个指定的关系以表示一个用户(主体)和其他用户(相关的用户)之间的这类关系。关系是形式上与主体具有特定关系的一组用户。关系可以是双向或单向。双向关系同等地适用于在主体和相关用户之间的两个方向。从而,如果用户A与用户B具有双向关系,则用户B与用户A具有相同类型的关系。一个示例可以是对等者关系,其可以说明在同一组织部门或具有类似工作说明的两个用户:如果用户A是用户B的对等者,则用户B也是用户A的对等者。另一方面,单向关系是定向的:如果用户A与用户B具有单向关系,用户B不一定与用户A具有相同种类的关系。一个示例可以是上级一下属关系:如果用户A是用户B的下属,则用户B不是用户A的下属。

[0098] 因为相关的用户是系统的用户,他们具有自己的库。根据配置,该系统可以搜索一些或所有相关用户的库作为查询的一部分并且将任何搜索结果合并到结果中。来自相关用户的库的结果与基线结果的偏差程度可以在例如专家比对等者具有更大偏差的关系水平和例如一些对等者可能比其它对等者施加更多影响的用户水平进行配置。

[0099] 在默认配置中,系统为各用户保持两种不同的关系:

[0100] 对等者:这是双向关系,用于表示具有共同兴趣、工作角色、位置和其它因素的用户。人们可以基于不同环境而属于多个对等者组。该系统通过学习而开发对等者组。对等者组根据团体和商业变化而改变和调节。

[0101] 专家:这种关系表示一个人拥有的技巧或知识。该系统通过检查具有发现和收集具有最大影响的最多和最有用的文档的能力的团体和个人而检测专家。专家是相对的。今天的专家如果停止与最有用的内容的联系则可能变得没用。

[0102] 虽然这种典型的配置仅仅为各用户提供单一的对等者关系和单一的专家关系,但是高级的配置可以为两个或多个个人提供各对等者对和专家对,其被称为域专业技术网络(DEN)。多个对等者关系或DEN允许用户识别多个在不同时间各自相关的不同对等者组,例如,天天操作的部门组,代表委员会成员的特殊兴趣组等等。多个专家组允许用户具有多个

集中在不同主题区域的不同专家组。

#### [0103] 监控用户活动

[0104] 图 4 示出了由嵌入式应用程序捕获的行为相关性的流程图。在图 4 中,信息搜索者在使用商业应用程序。在进行搜索中,例如“sales preso on bonds”,服务器执行各种数据挖掘活动并且为信息搜索者产生结果。在进行搜索的过程中,本发明产生隐式相关动作的观察报告,例如“查看”、“下载”、“打印”和“电子邮件”。服务器还产生显示相关动作的观察报告,例如“保存到库中”以及类似垃圾控制的动作,例如“移走”。下面将详细讨论这些术语。由系统产生的观察报告用于确定特定文档对搜索者的价值。如下面详细讨论的,该系统累积关于文档价值的信息并且然后为该文档开发有效性的度量。

[0105] 图 5 示出了根据本发明的内嵌用户界面的屏幕快照。由于系统中使用的标签可配置和可定制,所以对于特定的企业,可以将用户界面混合到现有的网站。图 5 给出的示例是公共网站。

[0106] 图 6 示出了根据本发明采用 JavaScript 标签所呈现的内嵌用户界面 (UI) 的屏幕快照。该特定示例示出了“最受欢迎的”标签,其为终端用户给出了最受欢迎的文档列表。采用 JavaScript 标签呈现该 UI。其它标签,例如“下一步”、“类似文档”、和“优选的”以类似的方式呈现。

[0107] 图 7 示出了根据本发明的弹出式用户界面的屏幕快照。与内嵌用户界面相同,该界面也采用 JavaScript 标签呈现。该特定示例示出了“下一步”。该标签渐现,并且当关闭时渐隐以提高用户体验。与内嵌标签相同,弹出式对话也可配置以混合到任何现有的 Web 页面风格。

#### [0108] 动作监控 / 观察

[0109] 该系统对知道用户发现哪个文档有用或相关具有活动的兴趣。然而,当认为特定文档有用或相关时,一般不能依靠用户对系统明显表示。相反,该系统不得不从用户对文档采取的任何动作而推断该信息,可以使用或不使用该系统,

[0110] 完成这种推断的一个方法是在搜索结果中对于每个文档为用户提供一个或多个对于典型动作的方便的按键或链接。由于这些动作可由鼠标单击实现,与采用正常浏览器控制通常需要执行多种动作的多次点击相比,用户更愿意使用它们而不是标准浏览器控制来执行这些动作。此外,因为这些按键或链接在系统的控制之下,该系统能够注意到用户对一个文档所采取的行动。这样,为用户提供了方便的机制以执行他们要对一组搜索结果中的文档执行的任何动作。

[0111] 在实际中,阻止这些用户动作是简单的。与 HTML 的通常情况一样,表示动作的每个按键或链接具有与其相关联的 URL。通常,该 URL 直接指代关联的文档。然而,对于该系统,这些 URL 改为指代 CGI、servlet 或与该系统相关联的类似机制。该 URL 包括关于用户、文档和用户想要执行的动作的信息。该系统记录动作和相关信息,然后,在简单“查看”类型动作的情况下,将请求重发到原始文档或者重发到一些其它类型的 Web 资源以完成所请求的动作。

[0112] 大多数内容搜索系统在表示搜索结果时,时文档的标题作为到达该文档的活动链接。该系统也采用这种标准惯例,除了该活动链接被认为是“查看”动作并且以与上述其它动作相同的方式被监控。

[0113] 可选的“添加到库中”动作也可用于文档。如名称所暗示,该动作将文档添加到用户的库中。这是用户明显地通知系统文档非常有用的方法。用户采用该动作的主要动机是确保该文档在将来的查询中被认为是有用的,因为在用户库中的文档一般有提高的排序。

[0114] 当用户选择负责手动地表示搜索结果时,连同在结果中各文档的其它通用数据一起提供用于配置动作的 URL。确保这些 URL 是用于用户可能对结果文档采用的各种动作是客户的责任,否则系统的价值减小。

[0115] 搜索和导航的一般隐式观察报告

[0116] 该系统采用更普通的设备在搜索和导航期间观察对于所有内容的用户行为。该观察报告没有用户参与而隐式进行,除了用户完成他们正常的浏览和搜索。观察报告与搜索或者导航统一,并且然后用于改善将来的搜索和导航。

[0117] 查询监控

[0118] 该系统还得益于知道产生用户采取动作的文档的原始查询。例如,如果系统注意到之后用户提出了相同的查询,或者如果注意到多个不同用户进行相同或类似的查询,它可以在原始查询中发现兴趣的新查询中增加文档的排序。然而,由于查询字符串可能很麻烦,不总是实际包括在动作 URL 中。相反的,系统为各查询保持查询字符串的数据库并且给出唯一的 ID。然后该唯一的 ID 可以包括在表示在搜索结果中的动作 URL 中。当用户对特定结果文档采取动作时,该系统可以确定通过查找该查询 ID 产生特定文档的查询。

[0119] 混合搜索

[0120] 该系统采用混合搜索来加强搜索结果。在混合搜索中,将单个查询发送到两个或多个独立的搜索处理器,各处理器产生一组零个或多个以某种方式与查询匹配的文档,其被称为结果集。取决于配置和境况,相同的文档可以在来自这些各种搜索的一个或多个结果集中出现。一旦所有的搜索处理器已经完成所请求的查询,将其结果集合并到一个单一的结果集。在合并的结果集中,采用考虑例如产生该文档的搜索处理器的数量和 / 或类型以及文档在各单独的结果集中的排序等因素而使用可配置的公式排序为单独的文档指定排序。

[0121] 两个不同的搜索处理器不需要是不同的软件实体。例如,运行两个不同索引和 / 或具有不同配置参数的相同的搜索引擎可以构成两个不同的搜索处理器。更重要的是,两个不同的搜索处理器通常应该为相同的查询产生不同的结果。人们可以认为各搜索处理器为查询提供不同的观点。

[0122] 可以为各搜索处理器分配权重,确定其影响在合并的搜索结果中排序的程度。该权重可以或者是静态的常数,或者根据查询、结果或其他情况改变的动态计算的值。

[0123] 搜索处理器可以,但不是必须,彼此独立运行。某些搜索处理器可以配置为将不同搜索处理器的结果集作为输入并且按照某些方式进行操作以产生其自己的结果集。这类搜索处理器被称为过滤器。过滤器对于例如缩小来自不同搜索处理器的结果的任务很有用,例如,移除太大、太旧等的文档,或者以某种方式修改,例如由文档内容计算摘要或标题、由修订日志添加注释、操作级别分数等。不过滤其它搜索处理器的输出的搜索处理器被称为独立搜索处理器。将第一搜索处理器是独立搜索处理器而第二和后面的搜索处理器作为其前面的搜索处理器的过滤器的搜索处理器的这种排列顺序称为流水线。构成流水线的单独的搜索处理器也被称为级。

[0124] 混合搜索的结果集通过合并一个或多个流水线的输出结果集形成。作为规则,各流水线为各文档在其结果集中产生用于排序文档的相关性的分数。当两个或多个独立搜索处理器的结果混合到一起时,这些分数被归一化到相同的范围,然后乘以缩放因数。如果相同的文档出现在多于一个流水线的结果集中,则来自各结果集的分数加到一起以形成混合结果中的单个分数。这些累计的分数确定该文档在混合的结果中的最后排序,其中最高分数被给予最高排序。

[0125] 实际情况中,为了效率不同的流水线可以并行运行。在概念级别上,单个流水线的各个级顺序运行,尽管在实践中在级能递增地产生他们结果集的部分时仍旧可以实现一些并行。用在混合搜索中的流水线的成分以及他们运行的方式,例如,串行或并行,由管理员配置和/或由终端用户动态地操作。

[0126] 在实际配置中,由该系统包装的现有的搜索机制被称为基线处理器。任何其它的搜索处理器被称为辅助处理器。基线处理器通常建立于传统搜索技术之上并且因此能够作为一个虽然不是最理想但是适当的文档搜索机制而独立运行。其中,这表示它应该访问企业中的大多数公共文档,具有能够处理来自大多数商业用户的典型请求的查询处理器,并且其不作为流水线中的过滤级。另一方面,辅助处理器没有这些要求,他们可以仅仅访问少数文档,他们可以使用或不使用传统搜索引擎来完成他们的目标,而且他们可以事实上作为流水线中的过滤级。

[0127] 注意该系统可以事实上配置有两个或多个基线搜索处理器。这有时被称为联合搜索,其中合并其它独立搜索引擎的结果。虽然这不是该系统的必要目标,但是这是其混合搜索技术的有利的特殊情况。

[0128] 图 8 示出了根据本发明的专家引导的个性化搜索交叉应用程序的示意性方框图。在图 8 中,服务器包括关于用户的库“我的库”的信息。用户的浏览器 21 显示具有“我的库”的视图。该视图的源包括搜索商业应用程序、Web 搜索和其它商业应用程序信息。这产生了网络效果从而其它应用程序也可以使用该服务器。用户的库是行为日志。其可以嵌入其它应用程序中,因此不仅是新用户界面或应用程序。内容由用户搜索和观察报告产生并且一般对用户不可见。系统的分析允许改善质量并且提供桥知识库。如在此讨论的,在本发明的操作中存在隐式的垃圾控制形式。该系统提供动态的个人导航支持。也实现近似哈希值、装载率和保密 C 政策。本发明以浏览器的形式和装面插件的形式实现,并且包括内容更新和缓存。本发明相关访问的信息依照域专门知识网络,该网络包括个人信息、对等者信息、专家信息和团体信息,这将在后面进行详细讨论。

[0129] 图 9 是根据本发明的用户库的屏幕快照。

[0130] 图 10 是根据本发明的用户库的第二屏幕快照;而图 11 是根据本发明的用户库的第三屏幕快照。

[0131] 样本搜索处理器

[0132] 该系统可以由与密切关系引擎(图 3)关联提供的不同搜索处理器实现,其可以按照不同的方式结合以实现不同的目标。下面的讨论说明多个可用的公共搜索处理器。

[0133] Lucene 基线搜索

[0134] 该搜索处理器是通过对现有的 Lucene 索引(参见:<http://Lucene.Apache.org>)提出查询而产生其结果的独立的基线处理器。其产生的结果集包括内容定位器和相关性分

数,该分数是 0.0 到 1.0 范围内的浮点数。

#### [0135] 库搜索

[0136] 该搜索处理器是独立的辅助处理器,其对于与指定的查询匹配的文档搜索特定的用户的库。在典型的实现中,为各用户的库保留 Lucene 索引,因此该搜索处理器本质上是相对不同索引用不同缩放因数运行的 Lucene 基线搜索处理器的特殊情况。

#### [0137] 我的库 (My Lib)

[0138] 库搜索处理器的这种特殊情况相对调用原始查询的用户的库运行。其通常以相对大的缩放因数运行。这样,其中用户之前表示兴趣并且与当前的查询匹配的文档趋向于接收提高的排序。

#### [0139] 关系搜索

[0140] 该搜索处理器是独立的辅助处理器,其搜索给定关系中相关用户的库。概念上其类似为各相关的用户调用库搜索处理器并且然后合并结果。实际上,可以按照多个不同方式进行优化,例如,通过并行执行各库搜索,或者通过为整个关系维护单独的合并索引。

#### [0141] 我的对等者

[0142] 该搜索处理器是已经为主体的对等者关系之一专业化的关系搜索处理器的情况。如果用户具有多于一种这样的关系,则可以以多种不同的方式为给定的搜索确定采用的特定关系。

[0143] 例如:

[0144] ●可以通过用户对这部分的显式动作设置,例如,用户可以指出该工作正在在特定的对等者组中进行;

[0145] ●可以通过当前搜索环境隐式设置,例如,用于开始查询的实际搜索形式可以选择特定的对等者组;

[0146] ●可以计算,例如,通过分析查询本身。

[0147] 该搜索处理器背后的原理是用户的对等者趋向于和该用户具有类似的兴趣,因此如果一个文档对于一个对等者有特别的兴趣,即,该文档在该对等者的库中,则它对于该用户也可能感兴趣。该搜索处理器一般以相对高的缩放因数运行,从而升高既与查询匹配又处于对等者的库中的文档的排序。

#### [0148] 传递关系搜索

[0149] 很多,但不是所有的单向关系是传递的:如果用户 A 与用户 B 具有特定的单向关系,并且用户 B 和用户 C 具有类似的单向关系,则如果该关系是传递的,可以推断用户 A 和用户 C 具有相同的单向关系。如果给定的关系表示传递的单向关系,则这种关系的传递闭环 (closure) 是原始关系的成员与对于各相关用户的相同关系的成员的合并。在完全闭环中,该过程对于各相关的用户和他们的各相关的用户等连续递归,直到计算了传递关系的完全树。在部分闭环中,递归局限于特定的深度。

[0150] 传递关系搜索处理器是独立的辅助处理器,其搜索属于特定单向关系的完全或部分闭环的所有用户的库。对于整个关系可以指定单一的递归深度,或者对于开始关系的各成员指定单独的递归深度。一旦已经计算了闭环本身,传递关系搜索处理器非常类似于普通关系搜索处理器,概念地在闭环中对各用户执行库搜索并且合并结果。为此,可以与常规关系搜索相同的方式对其进行优化。

**[0151] 我的专家**

[0152] 该搜索处理器是已经为主题的专家关系之一专业化的传递关系搜索处理器的特殊情况。如果用户具有多于一个的这种关系,可以以多种不同方式确定用于给定搜索的特定关系,如对于我的对等者搜索处理器所概述的。

[0153] 该搜索处理器背后的原理是,当进行一个主题的搜索时,如果用户可以识别在特定主题中的专家,则那些专家发现感兴趣的文档,即,出现在专家的库中的文档,对于该用户大概也有兴趣。此外,假设专家关系是传递的。从而,如果用户 A 认为用户 B 是某主题的专家,并且用户 B 认为用户 C 是相同主题的专家,则用户 A 也认为用户 C 是该相同主题的专家,即使用户 A 并不认识用户 C。

[0154] 如同我的对等者处理器,该搜索处理器以高缩放因数运行,从而使得由专家选择的内容被给予提高的排序。

**[0155] 新鲜**

[0156] 企业中的内容搜索相比于更普通的 Web 搜索的一个重要方面是结果中新鲜或崭新的重要性。在 Web 上,稍微老一些的数据一般被认为更有价值,因为它们有被全世界用户评价和审查的机会。然而,在企业中是相反的情况:大多数用户已经看到较老的数据,因此当执行搜索时,较新的数据通常更有用。

[0157] 新鲜搜索处理器是简单的辅助过滤处理器,其通过增加更最近的文档的分数并且减少较老文档的分数而捕获这种差别。文档的分数改变程度根据其年龄变化。从而非常近的文档可以使它们的分数增加得多于不太近的文档,而非常老的文档可以使他们的分数减少得多于中等年龄的文档。各种类型缩放的门限和范围都是可配置的,例如,可以设置一个仅处理老文档而不加强新文档,或相反,仅处理新文档而不加强老文档的过滤器。

**[0158] 显式偏差**

[0159] 一些文档对某些查询是规范而正确的答案。例如,在必须特别注意调整事件的组织中,例如, HIPPA、SOX 等等,关于特殊程序的查询,理想的回答是该程序最近、最正式の説明,可能排除所有其它文档。

[0160] 显式偏差搜索处理器是辨识某些查询或查询关键字并且在那些查询的结果中插入一组固定文档的辅助处理器,每一个文档具有通常非常高的固定分数。这一般不需要正式的搜索索引完成。典型的,其使用将关键字映射到文档的简单的表配置。可以配置为独立的处理器或过滤器。当其配置为过滤器,其可以被进一步配置为替换或者取代输入结果。当显式偏差搜索过滤器没有发现匹配的关键字时,其不修改输入结果。

**[0161] 受欢迎性**

[0162] 一些搜索题目在任何给定的企业中趋向于规则的重现,通常结果中只有少量的文档吸引每一个人。该系统可以通过注意何时相同的查询被提出多次然后观察响应这些查询哪些文档被操作的最频繁而检测这些受欢迎的结果。

[0163] 受欢迎性搜索处理器是利用该知识的辅助过滤处理器。其检查受欢迎的查询然后增加已经由进行相同查询的之前用户历史选择的结果中的文档的排序。实际上,其类似于显式偏差处理器,除了关键字与文档的表是通过系统通过分析查询和动作日志得到的数据而自动产生。

**[0164] 质量度量**

[0165] 由于系统观察查询和对查询结果采取的动作,其可以动态监控其结果的质量。然后其用于例如返回投资 (return-on-investment, ROI) 报告或网站设计反馈的目的。

[0166] 关于搜索质量反馈的简单形式可以通过比较查询日志和动作日志发现。如果用户查询没有产生对应的动作,或者也许仅对较差排序的文档产生动作,则系统可以推断该查询产生不良结果。另一方面,一个特别是对高排序的文档产生多个不同动作的查询可以被认为是良好的。

[0167] 质量反馈的其它形式是将对通过基线搜索处理器发现的文档采取的动作和对仅仅通过系统发现的文档采取的动作进行比较,或者也许是将其与仅仅通过系统更容易发现的文档进行比较。该系统可以通过注意到当实际对那个文档采取动作时,哪个搜索处理器对文档的相关分数贡献得显著来完成。如果对文档分数的唯一显著的贡献者是基线搜索处理器,则该系统可以推断其不给该结果增加任何特定的值。另一方面,如果一个或多个搜索处理器对文档的分数贡献显著,则系统可以推断其确实为结果增加值。

[0168] 通过结合这两种度量,该系统可以动态的产生感兴趣的和有价值的 ROI 报告。例如,一个报告可以比较查询的好对差的搜索结果的比率,该比率对于由系统提高的查询和没有由系统提高的查询的比率相同。如果将一美元成本分配给差的查询,则可以计算由最初的搜索系统呈现的差的搜索结果的成本和由本系统实施成本的差别。另一个报告可以集中在系统节省其用户的时间量。例如,仅仅通过搜索处理器而不是基线搜索处理器找到一个文档可以假设节省用户两个小时的手动搜索,而通过搜索处理器从低排序升高到高排序的文档可以假设节省用户 30 分钟的搜索。如果将每个小时的成本分配给用户的时间,则可以计算采用该系统节省的成本。

[0169] 非核心技术

[0170] 由于该系统设计用于包装现有的文档搜索机制,其需要采用大量不是其自己固有的技术。下面的讨论说明由本系统采用的这类非核心技术。熟悉本领域的技术人员可以理解下面仅仅是本发明优选实施方式的示例,也可以选择其它技术实现本发明。

[0171] 语言

[0172] 本系统大部分采用 1.5 版本的 Java 语言实现,并且所有的类采用由 Sun 微系统在其 1.5.04 Java 软件开发包中提供的 Java 编译器进行编译。假定在 Java 语言的任何 JVM (Java 虚拟机) 支持版本 1.5 中正确运行。如果客户不提供他们自己的 JVM,则默认使用 1.5.04 版本的 Sun JVM。

[0173] 应用程序服务器

[0174] 本系统的多数核心功能采用 Java servlets 和 Java 服务页面 (JSP) 实现。目前的实现写到 2.4 版本的 Java Servlet 规范和 2.0 版本的 JSP 规范。原则上,它应该在支持这些规范的任何应用程序服务器中运行。如果客户不提供他们自己的应用程序服务器,则默认使用 5.0.28 版本的 Apache Tomcat 应用程序服务器。

[0175] 搜索引擎

[0176] 本系统采用 1.4.1 版本的 Lucene 搜索引擎来管理用户库。目前的实现包括支持 1.4.1 版本的 Lucene。如果客户没有提供他们自己的基线搜索引擎,则提供采用 1.4.1 版本的 Lucene 的基本实现。

[0177] Web 服务器

[0178] 任何传统的 Web 服务器可以用于本系统来提供通用的内容。本系统的参考实现采用 Apache2.0.52。

[0179] 用于 Web 应用程序的虚拟配置的工具

[0180] 配置目标网络应用程序具有新的性能的工具,从而可以证明 Web 应用程序中具有新的性能,虽然 Web 应用程序不能以任何方式修改。

[0181] 用于概念的虚拟证据的方法

[0182] 提供了一种能够通过引导评价者通过一系列步骤评价现有 Web 应用程序中一组软件性能并且自动提供需要的构造来支持该评价的自动化处理。该处理是虚拟的,因为其需要目标 Web 应用程序没有变化并且没有软件安装。

[0183] 本发明的一方面涉及一种虚拟 Web 销售工具。在该实施方式中,本发明包括采用代理技术实现的虚拟环境。由预期的客户使用本系统来访问系统 Web 网站。这允许预期的客户看到“他们知道之前”和“他们知道以后”系统对预期客户活动的应用程序的影响。没有复制预期客户的活动的应用程序的内容而产生模拟活动的应用程序的虚拟环境。本系统虽然在代理环境中执行阻止和增加,但是没有物理地处理任何内容或干扰活动的应用程序的结构。这样,当预期客户进入该虚拟环境时,他们感觉好像实际在他们活动的应用程序中。本实施方式的一个好处在于本发明可以完成操作,而不必物理地进入客户的应用程序环境、得到日志或者客户涉及的 IT 部门。这样,客户没有真正知道存在变化,但是可以看到影响。

[0184] 在本实施方式中,可能经历不需要在传统意义上的软件安装的过程,其使得客户进入网站以具有与传统的软件提供者的同类的经历,除了本发明允许人们一直在线完成。该方法在该种意义上来说时虚拟的,除了与 Web 浏览器交互以利用服务而不需要做任何事情。本实施方式提供了一种为系统自动操作销售过程的概念的虚拟证据 (POC)。其目的是通过访问者进入感兴趣的产品的网站而捕获兴趣。用户通过点击 POC 而访问系统。然后系统自动操作通过 POC 的过程。一旦他们操作 POC,服务被开启,现在他们就是付费的客户。

[0185] 例如,为了通过网站捕获用户的兴趣,允许用户“试一下”。他们输入自己的电子邮件地址。系统用一级屏幕验证电子邮件地址。然后,系统在他们尝试之后发送一封电子邮件,并且可能是他们可以看到系统如何工作的屏幕的链接。

[0186] 在第二阶段,系统为他们产生示范空间。这是基于他们在第一步骤中对系统给出的一些信息,并且此外,系统现在需要他们上传一些信息,例如一个日志文件,以为系统提供一些关于如何使用他们的网站的历史信息。然后本系统得到日志文档,自动产生一组给他们解释该系统可以提供什么预期值的增加的报告。然后本系统经历一个自动过程并且产生他们的网站在本系统之前和之后看起来的样子的“之前和之后”的图片。

[0187] 做一定量的后端支持来使其工作。一旦他们在第二阶段实际进行,本系统向他们解释他们需要上传他们的日志文档。然后本系统可以为他们提供他们可以打印出来的报告并且采用其围绕该系统建立内部动力。然后本系统允许他们以 POC 形式在某一段时间使用该系统,并且然后将其转换为真实的顾客。例如,该时间周期可以是 30 天,也可以是 90 天。这样,本发明的这一方面通过需要很少人为干预的在线方法处理而发展,以允许他们体验本系统的价值,其不需要直接与销售人员接触而感到任何压力,并且也不需要将销售人员到他们的地址。

[0188] 有用性的算法的方法

[0189] 本发明的这方面涉及推导出说明电子资源有效性的分数的方法。与公知的相关性算法相反,其帮助发现与一个查询上下文关联的文档,有用性的计算基于用户关于任何电子资源的行为而测量该资源的实际有用性。给定一个题目,可能有成百上千的相关文档,但是只有少部分是有用的。有用性测量一个文档对于给定的用户如何有用,而相关性测量与该内容匹配的关键词。为任何电子资源和范围从数百万到单个用户的任意用户群体大小而计算有用性分数。

[0190] 从而,相比集中于相关性检测的传统搜索技术,本发明检测有用性。对于相关性,例如,如果一个人在学习 Java 编程,存在数百种可以用于学习 Java 的相关的 Java 书籍。他们都有用吗?不是。如果一个人真正想要学习 Java,他应该问 Java 老师读什么书,他们可能会推荐两本或三本书,而不是数百本 Java 书。这样,相关的书包括所有这数百本书,而有用的书是两本或三本非常有用的书。这种有用性以专家、团体、对等者的知识为基础。

[0191] 专家、对等者和团体知识由本发明基于观察的用户群体的行为而自动提取并且组合。由于用户行为随着时间变化,系统适应于专家、团体、对等者知识的代表。可以实时记录用户行为(通过应用程序中其它地方说明的观察的各种方法)或者可以从用户行为的现有日志文件中提取用户行为。在正在进行的基础上,基于正在进行的实时观察和获取日志文件的连续更新,本系统可以继续改进性能。更新总计日志文件中的差别,例如在逐月的基础上。这是除本系统基于观察捕获的信息之外。存在观察报告没有追踪的 Web 日志中的某些东西,并且存在现有网络日志不追踪的而可以被实时观察的某些东西。存在比运行时间或用户时间需要的更多的活动,但是通过从这些日志文件中捕获的海量数据中得到的事实和总结概括之后思考这些很有意思。

[0192] 系统的自我学习和适应的方法

[0193] 本发明的这方面涉及一种能够使系统确定其用户群体的行为关于电子资料 and 用户群体本身的成员中的改变,并且自动使其操作适应对变化的自我校正。自我校正使系统在变化明显之前能够主动识别并且适应变化,从而最小化对于维持系统的管理干预的需要。

[0194] 从而,本发明的这方面涉及本系统的属性,即,系统的继承性质,因为其观察人们的行为。当人们的行为改变时,他们的偏好变化,并且他们的有用内容改变。本系统自动适应这些改变。从而,默认地,本系统是可以校正自己的自我学习系统,因为当人们开始校正自己时本系统也跟随他们。

[0195] 本发明的技术为内容类型独立或内容不确定

[0196] 本系统对例如音频/视频文件、诸如 RDBMS 数据的数据类型以及诸如“购买”按钮的应用程序节点的任何内容/信息类型起作用。从而,本发明的优选实施方式包括独立的和内容不确定的系统,由于本系统不考虑内容本身。这不像传统的搜索技术,该传统的搜索技术解析内容、在内容中挑选出关键字并且用那些关键字选择结果。相反,本发明不关心内容是什么,而是关心资料的位置以及人们如何与该资料交互。本发明不关心那部分内容是什么。在简单情况中,其可以是文本文件,但是也可以是视频文件,其在传统技术的意义上,没有文本去解析并且没有索引可以建立。

[0197] 本发明的技术从 Web 服务器日志、搜索引擎日志、Web 分析服务器日志和其他日志

文件中播种系统,从而它可以从操作的第一天产生价值。

[0198] 通过基于专家和对等者的角色、名誉和专业技术等对其进行分配的管理者完成监督指导,虽然这不是必须的步骤。还可以从历史日志文件中推断和提取这样的信息。因为本系统是学习系统,随着人们使用本系统,可以得到更多的价值。本发明的这方面涉及使本系统从操作的第一天有用的播种技术。它也许不是 100%有用,因为其可能会有些偏离 (down the road),但是它至少有 50%到 80%的价值。在本实施方式中,采用实际是企业中发生什么的历史记录的 Web 服务器日志。它没有最终需要的详细信息,但是其具有粗略的信息。日志文档提供历史性信息。优选的实施方式依赖网站的业务模式而使用数周到数月的日志文件。这样,本发明提供了一种得到用户已经具有的某些东西的方式,即日志文件,并且使其转变为用于播种系统的资源。然后,如在此讨论的,随着时间过去,由于本发明利用浏览器扩展或正在运行的脚本进行观察,本系统会学习更多。本系统不仅利用基本日志,而且利用从各公司可以商业获得的对本领域的技术人员公知的较高层次的分析理论的那些日志产生的分析。

[0199] 本发明没有沿用或索引应用程序而组合多种应用程序、网站和知识库。

[0200] 通过用户实际使用那些应用程序而完成组合。组合是这里的核心技术的一种属性和自然结果。搜索的传统方法是具有多个索引,每一个链接到不同的知识库或不同的应用程序。为各个不能交叉搜索的知识库针对具有不同索引的各知识库执行搜索。

[0201] 在本发明的系统中,因为当人们在应用程序的环境中使用资源时,他们不在乎在哪里使用资源,所以自动提供组合搜索。他们可以在一类知识库中使用一条特定的内容,然后下一分钟可以移到不同的知识库,例如从 CRM 系统开始然后移到 ERP 系统中。这样,用户产生踪迹,即各种系统的虚拟链接。当再次搜索这条查询时,本发明的系统可以从多个不同的数据资源推荐信息,从而由于用户会产生该组合而使组合是自动的。即,用户从各种数据资源并且跨越各种数据资源使用信息的模式产生该组合。

[0202] 这里本发明不需要沿用网站或应用程序,或者应用程序的索引或者其内容。此外,本发明尊重任何已经适当的安全。在建立组合搜索系统中的重要挑战是组合的搜索系统必须理解并且与这些应用程序潜在的安全一起工作。很难做到这样,因为各应用程序一般具有他们自己的安全模式。一般来说,安全模式不在不同应用程序之间共享。当保护安全的同时,搜索的组合是巨大的挑战。在这种意义上,本发明是唯一的,其没有任何特定的适配器而自然完成,并且其保证可以完全保留那个应用程序的潜在的安全机制。这以完全唯一的方法完成。本系统通过浏览器代替实现私有模块来保护安全。

[0203] 传统上,为了解决组合问题,将有某些搜索应用程序进入各应用程序,并且概念上包括专门的安全模块。问题是搜索引擎实际上在建立所有内容的索引。当执行搜索时,人们不能简单返回搜索结果列表然后阻止没有权利访问的其他人点击该列表。因此,有效地,在一个索引中搜索引擎复制多个安全模块。发明人已经意识到不需要这样做,因为本系统具有通过系统进行用户查询的浏览器。然后本系统访问其内容数据库,并且作为回应,提供结果列表。这时本系统不过滤掉所有内容,而是相反地过滤所返回的结果。本系统在浏览器内部提供技术,如果在线用户可以访问这些知识库,本系统则实时检查这些知识库。如果答案是否定的,则阻止用户查看内容。该内容被保持在列表之外。用户甚至不知道发生。主要操纵是否浏览器具有访问权利的是这时登录浏览器的人,基于该人在系统中的特权,

其决定是否这个人可以看到结果。如果这个人不能看到一些结果,则系统不显示这些结果。这样,本系统实时询问应用程序是否一个特定的用户,例如目前登录的用户,现在可以访问内容。如果是真,则允许用户看到文档。本系统实时的每次询问应用程序是否用户具有足够的特权。此外,使用什么机制并不重要,因为根据他们是基于组、基于身份或基于外形,这个人可以具有不同的访问权利。

[0204] 完全由个体、对等者组和专家组在非预定的级别依靠例如查询条件、导航模式的环境使用而驱动的个性化。

[0205] 本发明的这方面通过知道用户是谁而完成个性化搜索,当使用系统时用户展示一定行为时,该用户是自识别的,例如通过 cookie、登录等。甚至用户是匿名用户,系统在用户的浏览器中放置 cookie。这样,当用户使用系统时,他留下个人踪迹,然后系统基于用户是谁而使信息个性化。在该系统中,没有人基于个性化而预定关系,因为该系统以用户的行为为基础。用户和其他人的密切关系产生一个空间,被称为俱乐部。这样,用户可以通过展示在一个领域的兴趣而暗中形成他自己的俱乐部。实际上没有人监控该用户。俱乐部完全通过用户的行为建立。

[0206] 本发明用于风险管理和验收测试的可控部署。本系统通过控制在活动运行的系统中可以看到产品特征的人们的数量而减少产品部署风险。设置特殊的 cookie 并且将其发送到受控的测试人群。通过该 cookie,网站的用户可以看到本发明的特征,而一般用户看不到这些特征。这是在企业中部署新技术的理想方法。

[0207] 扩大的搜索。本发明的这个特征将传统的全文本搜索与来自全球、对等者、和专家群体的偏好的积极信息进行混合,以为用户给出精确的答案。本发明的这方面规定如何使用索引。使用团体的观点,本发明可以扩大对于更好结果的搜索。在扩大的搜索中,对客户 Web 服务器提出搜索请求并且得到结果。然后,搜索的请求连同 Web 服务器结果、产生的查询以及用户标识符被发送到搜索服务器。响应回来。然后系统以搜索服务器格式发送回扩大的结果并且呈现给客户 HTML。

[0208] 前 N 个是基于环境和由团体的使用驱动的最有用信息的列表。例如,环境可以由用户查询、题目的显式说明或特定网页或网页组上的存在进行设置。本发明还产生一个被称为前十位的观点,例如,对于给定题目或给定环境的前十条最重要、最相关、最有用的信息。用户可以基于团体的环境驱动的信息使用而看到信息。前十位是受欢迎的结果。为用户给出与查询术语(环境)相关的十个最受欢迎的链接,或者也许没有术语。如果没有术语,则返回前十个最受欢迎的页面。为了所有这些查看,人们可以应用过滤器,例如仅仅查看落入技术范围的前十位,或只查看前十个 PDF 文件。

[0209] 更喜欢这个 (More-Like-This) 基于内容使用和团体身份扩大了类似的内容。如果用户喜欢一条内容,系统观察它,并且具有基于团体的兴趣可以显示给用户的其它内容。

[0210] 更喜欢这个是一种概念,其适用于当用户正在读这个页面时但是想要找到另一个非常类似的页面。更喜欢这个基于团体说的是更喜欢这个,意味着基于使用模式。这样,团体看到这个页面类似于用户正在读取的页面。

[0211] 预测导航提供基于在应用程序上类似用户在哪里开始和结束而为浏览导航提供捷径。

[0212] 在该实施方式中,如果具有用户身份的人们来到应用程序中的特定节点,则该用

户更喜欢去另一个相关的地方。本发明的这方面基于由对等者和专家的先前导航,包括他们在哪里开始和在哪里结束,而预测导航并且简化传统的导航。这样,开始点和结束点对于预测用户的导航、尝试简化一系列导航步骤的中间部分并且没有在大量其它地方不浪费时间地直接将用户送到目的地很关键。

[0213] 预测导航也被称为“下一步”,并且取决于人们想要显示哪个计算或结果。预测导航使用导航踪迹。有一个导航踪迹的概念;系统跟踪用户最后去过的N个页面。如果适用,系统保留用于开始该踪迹的查询的历史。从而,用户搜索并且结果出现。用户可以点击该结果。用户可以点击该结果。用户可以再次点击以到其它地方。用户保持点击。系统累计这些历史页面。这样表示整个历史是由于查询。系统基于用户的历史和过去的其它观察报告,尝试指出用户将要去哪里。返回的推荐是基于历史,即用户在该对话中已经累积的踪迹,该用户之前其它人们已经发现有用的页面。这样,系统尝试匹配用户已经去过的地方和指出他将去哪里。

[0214] 本发明的这方面指出如何使用索引。使用团体的观点,本发明可以扩大搜索以得到更好的结果。

[0215] 快速浏览(Zip through)涉及将内容预加载到系统上的概念。随着用户接收结果,系统为用户显示链接是什么的预览。这样,用户只是快速浏览而不用进入该页面并且改变该页面。如果他看到内容是他想要的,则点击进入。

[0216] 在用户团体中动态和适应性识别对等者和专家

[0217] 对等者和专家不难划出一个范围,但是网络用集线器集中并且在各种集线器之间连接。信息和知识组合到一起,自然和自动的产生智慧融合效应。

[0218] 本发明本质上使用相同的信息来确定用户团体。谁是对等者?谁是专家?本发明不仅确定用户愿意看什么内容,而且可以确定实际是用户的对等者的人。自然形成对等者组。他们没有明显的分界线。在这些人们之间有密切关系。非偏见和关键的隐式投票

[0219] 本发明的这方面提供比传统内容投票或调查更准确的内容有用性的预测。由使用驱动文档等级反映了内容的真实价值。对于例如在此公开的系统以高信心可靠的工作,隐式观察报告非常重要。如果你要求人们对内容投票,你可能得到有偏见的结果。你也得到非常歪曲的样本,因为大多数人没有时间投票、调查、做任何明显的事情。投票的人们趋向有偏激的意见。他们手上有大量的时间。他们坦率直言并且固执己见。这样,他们趋向于不代表整个群体。样本不匹配群体。他们也趋向于投否定多于肯定。这样,本发明优选的不采用显式投票。其考虑隐式动作。请求打印的用户是隐式的,因为当他进行请求时他在做其它事情。用户没有给出反馈也没有被要求反馈。本发明开发被动的观察报告而不是主动的反馈。虽然,某些实施方式可以包括主动反馈,例如否定反馈。用于计算智慧的方法(用户团体随着时间对资源的相对价值)

[0220] 本实施方式涉及观察关于电子资源的信息的方法,用户群体关于电子资源的行为以及资源和行为随着时间的改变,以辨别资源对用户群体随着时间的相对价值。该方法确定价值的范围从短期的信息到中期的知识,到长期的智慧。最终,本系统为企业提供正在进行的、内容不确定的和适应制度的存储器。

[0221] 计算的智慧意思是该智慧是团体行为关于一组资源的一种形式,其不随着时间而改变。根据人们如何频繁的改变意见而具有上述四项。例如,内容是最不可靠的东西,因为

内容可以改变。信息是第二级别的信任。如果由于人们的意见而使信息保持稳定,则这组信息可以成为知识。如果知识可以经过时间考验并且继续被使用和支持,则其成为智慧。因此,智慧不能年年改变。知识可以每个月改变,而信息可以天天改变。内容本身没有任何意义。这样,如果内容没有随着时间变成没用,但是随着时间它成为不变的并且有用性保持不变,则它从内容变为信息的级别,然后到知识的级别。随着一段时间过去,如果有用性保持较高或者继续增加,则其成为智慧。然而,如果随着时间有用性有波动,则该变化表示其可能不是真正的智慧,而只是感兴趣的当前信息。

[0222] 本发明通过使用内容、数据和应用程序而不是依靠内容所有者对什么是缺少的、热点的、好的和坏的认识而提供内容差距分析。

[0223] 人们不知道什么内容是缺少的,或者人们在寻找什么。也不知道可以产生什么类型的内容包括人们需要消耗的。由于从本系统可以知道趋势是什么,人们要求什么,可以回答人们是否满意某些内容的问题。即,知道差距在哪里。许多人在请求这件事并且没有找到任何有用的。存在差距。

[0224] 差距分析报告提供检测内容中的差距的能力。假设内容在那里,他们就是找不到。通常存在差距。对于这种情况,笨系统确定人们实际在寻找什么和什么是缺少的。在传统搜索或导航情况中某些人可能搜索什么东西,然后或者失败,或者如果他们不甘心失败,他们可能再次搜索,或者他们可能潜在地尝试导航。通过这些机制,他们可能成功或者失败。当人们开始显示搜索或导航行为时,本系统解决这个问题。可以精确地知道他们在寻找什么,并且在搜索和导航中随着时间开始浮出表面的内容是团体本身认为有用的内容,而不考虑开发商或商人,或者什么商人考虑这些内容。这样,某些人在寻找一些东西,他们认为将会有用,但是他们找不到。本发明的这方面允许人们评价需要该信息的程度。

[0225] 在部门、公司或行业级别应用信息差距

[0226] 本系统提供随着时间的信息流,并帮助公司管理信息。本发明的这方面使用应用的信息差距作为随着时间的分配流来确定信息如何流动。本系统使得人们理解人们要求什么和什么内容可以用来满足那些需要。此时,可能看到信息流从部门到部门和从区域到区域进去出来。可能看到哪个位置表示什么信息,或者什么组或什么号码的人们。

[0227] 识别专家和对等者使公司能够在全局范围内定位其专业技术的能力

[0228] 可以测量用于公司或行业效率针对信息消耗的仪表盘 (dashboard),并且可以第一时间得知白领工人的生产力。本发明的这方面涉及在全局范围内确定具有专业技术的并且由专家和对等者使能的公司的能力,其允许本系统为找到谁知道什么以及什么可以在哪里做而提供仪表盘。

[0229] 通过赞助广告建立自动市场列表的方法

[0230] 该方法识别在公共网站上用于广告的目的的购买关键字和标题广告空间的公司。本发明考虑通常的和罕见的关键字以及给定标题广告的上下文,并且自动产生通过其网站寻找改进的领先一代的公司列表。本发明使用在线广告本身中发现的信息并且将其与其它公共信息资源结合以产生精确的公司列表。本系统然后追踪广告的买家,并且自动在候选期望列表中包括该信息。

[0231] 用于改进赞助广告变换率的方法

[0232] 本发明的这方面帮助想要保留客户或增加领先性的公司。这通过增加例如Google

和 Yahoo 广告的赞助广告的变换率来完成。

[0233] 基于上下文和用户从哪里来,例如来自具有给定搜索术语的 Google,本系统可以将用户引导到给定网站上最有用的信息或相关网站的集合。如果没有这种能力,到达网站的用户将不再具有这些公共搜索引擎指导他们到最相关信息的优势。

[0234] 本发明通过观察从公共搜索引擎给定的初始查询上下文,团体在哪里找到最有用的信息,而将这些用户路由到最有用的信息。在这个过程中存在两个步骤:1) 首先,本系统捕获在给定的查询环境下有用信息位于那里的集体智慧;2) 第二,当用户来自 Google 或 Yahoo 时,本系统调整集体智慧以将人们准确引导到正确内容。

[0235] 工程特征

[0236] 使用驱动的连接分析

[0237] 如果用户在特定的网站点击了链接,则捕获该链接中的文本以扩大在该网站中将来的搜索和导航。只有在导航踪迹指向成功的文档发现,即,用户隐含地找到有用的文档时,才注意该文本。

[0238] 本发明的这方面,使用驱动的连接分析,涉及锚定文本(anchor text)。这与 Google 非常不同,因为当 Google 为页面排序时,其解析每一个单独页面和多少链接等。本发明解析由人们使用的链接。一个链接除非被使用,否则是死链接。因此,如果某人点击链接,则该链接由于这个人而有用。

[0239] 此外,除了个人用户的使用,使用的链接由很多对等者和专家的使用交叉检查。对等者和专家隐含的确认和认可减少了来自个人行为的噪声并且加强了信号—噪声对噪声的比率。

[0240] 链接的成功使用通过相对于该链接而捕获和分析个人用户行为、对等者组行为和专家组行为进行确定。

[0241] 链接的使用和文本的重要性通过用户、对等者、专家、查询上下文和时间的混合向量确定。本发明的这方面,链接的成功使用,确定了单个用户如何行为。除了考虑链接本身,如果用户点击它,则它是有用的,本系统还对类似于用户的多少其它对等者点击这些链接做额外的分析。例如,多少其它专家不同于该用户,但是该用户依靠他们完成他的工作,以及谁也点击该链接。这样,存在两级价值:内容的个人使用,以及对等者组和专家组的使用。这些方面给出了数据的总价值。

[0242] 环境术语的隐式建立

[0243] 为内容建立元数据已经是历史性挑战。本发明的系统部署使团体隐式产生说明内容是关于什么的环境术语的唯一方式。通过观察进行查询并且经由查询找到有用结果的用户而学习环境术语。还包括关联各个使用具有链接文本的导航踪迹。本系统通过观察访问者如何使用各种术语来说明内容以及网站如何使用链接来说明内容而建立它的词汇表。此外,用的越多的查询和链接越重要并且越关联内容,而在下游踪迹中产生没用的内容的连链文本与内容无关。

[0244] 经由三种方法之一完成捕获该信息

[0245] JavaScript 标签。在这种方法中,页面用一段 JavaScript 实现。这种 JavaScript 减色链接的使用和文本并且向服务器发送这条信息。

[0246] 浏览器外加。在这种方法中,浏览器用软件实现。该软件检测链接使用和文本并

且向服务器发送这条信息

[0247] 日志分析器。在这种方法中,经由特殊程序,日志分析器,分析对于网站的访问日志,该日志分析器检测链接的使用并且向服务器发送这条信息。

[0248] 将以上所有被捕获的信息称为观察报告。上述捕获的观察报告的分析在服务器中发生。

[0249] 客户

[0250] 系统客户包括三个通用区域:UI、观察者和代理。

[0251] 客户端包括承担客户 UI(参见图 5-7)的 Web 浏览器。客户包括 JavaScript 观察者以在客户端完成 Web 页面的使用的观察报告。本发明的一个实施方式包括表示来自系统引擎的推荐的工具条 UI。本发明的这方面表现为产生显示 UI 所需的 JavaScript 的 JavaScript 标签。在该实施方式中,在页面上并且沿着存在系统产生的推荐的位置显示企业 Web 内容。不同的 UI 提供弹出,用户点击页面上的例如图标的东西,其调用系统代码以在环境中来显示弹出。

[0252] UI 还包括用于从系统服务器获取结果的 API。这代替了从企业安装的搜索服务器直接获得结果。在典型的企业网站上,用户输入搜索术语,点击搜索,则搜索到达他们的 Web 服务器。然后 Web 服务器返回给他们的搜索服务器。搜索服务器以某种格式返回结果,通常是 XML,然后他们在前端具有表示代码以通过 XML 解析并且表示搜索结果。本发明以类似方式操作,除了当他们的 Web 服务器返回到搜索服务器时,不是直接返回到搜索服务器,而是返回到服务器侧扩展。然后扩展从他们的搜索服务器获得原始结果,将其反馈到系统,而系统或者重新排序结果或者,无论如何,加强结果,可能增加更多条目。这被提供回到扩展,而扩展报告回他们的 Web 服务器。他们的 Web 服务器继续做他们以前做的,通过 XML 解析,重定格式,并且将其发回到客户。

[0253] JavaScript 观察者是一段为用户给出的封装为标签的 JavaScript 代码,而用户使用该标签实施他们的页面。JavaScript 标签驻留在客户页面上并且完成观察报告。例如,滚动或暂停观察报告。例如,如果终端用户在读取页面,他将遵守什么被定义为“暂停”。一旦暂停发生,即,一旦 JavaScript 观察者观察到暂停,则他将该信息发回到服务器。服务器累积这些观察报告。

[0254] 这些观察报告的每一个在例如关系密切引擎的后端都对应于特定的计算,扩大的搜索涉及重新使用用户的 UI 的观念,代替位于表示层和搜索服务器之间并且扩大来自那里的搜索。预测导航、前十位、更像这样、环境文扩展和快速通过都是用户可以放入页面的所有类型的标签,并且他们在产生建议中都使用不同的算法。

[0255] 基于历史,即用户已经在该对话中累积的踪迹,回来的建议是用户之前的其他人已经发现有用的页面。这样,本系统尝试匹配用户已经去过的地方并且尝试指出他将要去哪里。

[0256] 观察报告

[0257] 存在来自浏览器扩展的八个方向的观察报告,其是观察用户的对话和收集信息的脚本。

[0258] 代理说明系统没有访问那些页面源代码的权力的用户页面上的系统 UI。这些用户是不想给出对他们页面的访问权利的预期客户。本系统需要实时页面,但是通过将我们的

标签插入页面而向用户显示通过我们的系统页面看起来将是什么样。为此,本系统使用代理从 URL 进入或得到页面,然后基于配置规则,改变那个页面的 HTML,并且然后将那个页面发送回浏览器。代理位于浏览器和目标 Web 服务器之间,即,潜在客户的 Web 服务器。代理本身有它自己的 URL,并且它在目标 URL 中作为该 URL 的一部分通过。用于本发明实施方式目的的 URL 包括两个 URL。实际使用的 URL 指向代理,但是嵌入在 URL 本身的是你需要代理去往的目标,即,客户 URL。URL 首先到达本系统,并且然后重建 URL,带你到客户页面。然后,代理为你建立 HTTP 链接以获取客户网站的页面。其着眼于页面并且向其施加一组规则,并且然后将该页面发送回用户。

[0259] 关于客户浏览器,页面首先由标签实施。目前优选的标签格式是 JavaScript (JS)。客户在他们的 Web 服务器上结合 JS 文件。然后他们用脚本标签在 HTML 中引用。一旦 JS 文件被装载,该文件在系统中建立对象。然后在页面中建立显示 UI 的地方。在系统侧,本系统向 UI 发回 HTML。客户端的管理员指定标签上的样式单。例如,即使是相同的 HTML,由于样式单不同,用户得到不同的颜色图案和字体。

[0260] 可以提供插件用于与代理类似的目的,因为它也修改 HTML 并且返回搜索结果。然而,与代理不同,用户配置插件。当在用户的内部网站执行 URL 搜索时,插件接收搜索请求,为结果执行搜索,将其发送回系统,然后其扩大结果并且发送回插件。插件完成显示的 HTML 的交换。这样,代替显示那个 URL 的 HTML,插件程序显示来自系统的修改页面。插件还可以完成观察报告。因为它具有比 JavaScript 更多的对浏览器功能的访问权利,它具有更好的能力来捕获更宽范围的观察报告。

[0261] 不同的 UI 以相同的方式工作。例如,前十位和预测导航对于 UI 以如上所述的相同方式工作。唯一的不同在请求中。例如,当用户要求系统扩大时,系统被要求进行特定的计算。

[0262] JavaScript 观察者在页面上保持、等待并且观察用户。观察报告由用户动作构成并且观察报告被发送到系统,其包括关于观察报告的所有信息,例如它是什么页面,是否存在用户信息。

[0263] 当用户在一个特定页面上已经花费了 N 秒或 N 分钟时观察暂停。

[0264] 范围可以作为选择的事件而被选择,但是典型地为大约 30 秒到 5 分钟,取决于被检查的文档的复杂性。过量的时间意味着用户离开计算机。本系统优选的不捕获上限,因为用户可能正在读取文档。

[0265] 在用户读文档的地方存在虚拟书签 / 虚拟打印特征,并且发现它有用,但是不能记住文档读过的一切,因此他将该窗口保持在其它窗口后面的地方打开。然后用户进行其它任务,而当他需要参考该文档时,他再把页面弹出来;用户可以对文档设置标签。这样,即使用户没有标签文档,表示该信息可能在下面几个小时或几天有用,而用户不需要标签,用户将该文档保持在窗口上很有用。在这种情况下,用户没有显式地标签一个文档,但是由于某些原因它保持文档打开。如果一个人在任何给定的时间看一个典型的计算机,在计算机上打开的东西没有开放,因为用户在用它们而且没有关闭它们。它们趋向于被打开因为它们是虚拟书签。这样被保持打开长时间,例如,两分钟、五分钟、十分钟的东西被认为是虚拟书签。

[0266] 滚动涉及屏幕的滚动

[0267] 锚定文本是超链接文本,其使用户到达当前页面。其可以像用户点击新闻术语然后产生一些关于该主题的当前新闻一样简单。

[0268] 思考是一种使用模式,即,暂停和滚动,或者某些动作,鼠标移动等的结合,其表示用户在考虑该页面。这样,思考是一些动作之后的静止周期。

[0269] 邮件是当用户向其它用户发送或转发内容时,或者以目的在于发送邮件的虚拟书签的类似方式发送内容。

[0270] 密切关系引擎和团体的智慧

[0271] 图 12 示出了根据本发明的文档推荐的流程图。在搜索 110 的开始,存在各种术语向量 T1、T2 和对等者 / 专家群体。术语向量可以利用并且与每一个其它文档的术语向量进行比较,从而选择前 N 个匹配。找到 N 个最好的文档的最受欢迎的术语并且将这些加到术语向量中。在这点上,为每个文档可以得到积极信息并且新的术语向量可以与每一个其它文档的术语向量进行比较。这两者可以是结合的 118 然后前 N 个可以是选择的 119。下面详细讨论术语向量和文档搜索的概念。

[0272] 本发明对在文档中执行搜索的方法使用类似的策略。公知的方法以向量空间模型表示一切。这种方法得到文档然后制作向量,其包括术语空间中的所有文档术语。为每一个文档完成。搜索表示为术语空间中的另一向量。

[0273] 在这点上,本发明使用向量空间模型,但是其建立向量以表示文档的方式与文档中是什么无关。其与其它人已经搜索了什么喝已经找到该文档有关。例如,个人执行关于术语“芯片设计”或许其它词的搜索。他可能已经结束找到文档。可能已经查看返回结果的列表,但是他可能已经结束找到文档。当他一实行,并且他找到,则本发明将他所搜索到的内容与他那文档关联,并且贡献给向量。还有其它方法填充术语向量,例如通过用户导航行为(后面说明),或者通过用户的显式输入。这样,本发明基于其他人如何使用文档而建立文档的代表。

[0274] 代替给文档单个术语向量,其是在搜索空间中所发生的,本发明为每一个个人用户对于一个文档给出他们自己的术语向量。因此,每一个用户开始说关于一个特定文档是什么的他们的意见。某些人对某些文档可能没有意见。然而,例如,如果某些人曾经执行过搜索并使用该文档,则他们的意见登记在关于那个文档的他们的向量中。知道这些,本发明允许执行各种功能,例如“我想匹配所有文档,但是我不想看到每个人的意见”;或者“我只想看到我的对等者的意见或专家的意见”。在这种情况下,本发明对于不同的人采取多种术语向量,将它们加到一起,得到群体认为那篇文档关于什么观点,并且使用该结果将人们连接到正确的文档。

[0275] 这样,本发明的这方面提供搜索扩大,其依靠基于使用的题目检测的新颖技术。术语向量提供向量空间模型来表示使用确定的题目。

[0276] 活动性。除了术语向量,本发明还包括考虑每个用户使用什么文档的向量。每个用户具有这些之一,并且将其称为活动向量。每次它们使用特定的文档,本发明注意到他们已经使用它。活动向量中的每一块(bucket)是特定的文档或资料,并且块累计使用。某些块可能是零。某些可能是很大的数字。用户对资源的不同动作,例如读取或打印,对活动向量都有不同的贡献。可以基于群体中每一用户的活动向量而产生该群体的活动向量。例如,对于特定的群体,例如七个用户,将使用向量相加以确定特定文档在该全体中被使用多少。

本发明结合这两部分信息,即术语向量和活动向量,以帮助推荐文档。例如,可能有一个文档其按照题目非常匹配但是使用的不太多。即使两个向量都涉及基于使用的信息,一个向量涉及使用的数量而另一个向量涉及使用的环境。本发明将这两个数量结合到一起来建议文档。如果我们需要,本发明还可以将来自现有搜索引擎(其基于资源的内容而达到题目匹配)的结果与术语和活动向量相结合。

[0277] 每个个体有他自己的库(使用的资源的集合)并且对个体的库中的每个文档,系统有用户的术语向量,其表示他们认为文档是关于什么。本系统还具有他们的活动向量,其表示他们已经在任何环境中使用那个文档多少。现在可以结合任何给定组的用户并且要求他们关于文档的意见集合。还有全球使用向量,其是每个人的向量的总和。还有特殊使用向量,用于匿名用户的组。每一个不认识的人都对相同的使用向量贡献。当结合向量以产生集体的观点,对于不同人的向量可能有不同的权重,但是每个人的总和是全球向量。本发明还包括用户的对等者和专家。

[0278] 对等者。本发明确定对等者的方式类似于在合作过滤器应用程序中确定对等者的方法,例如 Amazon.com,但是在这种情况下,确定是基于文档使用。本发明考虑两个用户的文档使用(活动)向量,例如一个人已经使用这个文档、这个文档和那个文档;另一个人已经使用相同的三个文档,因此他们是类似的。即,在这种情况下,当本发明比较两个用户的活动向量时,存在文档使用类似性。在优选实施方式中,在明显的一组使用的资源中重叠的两个用户被认为是对等者,无论是否其它资源仅由一个用户或另一个用户使用。即,一个用户使用另一个使用的子集意味着他们共享公共兴趣或角色的事实。在这点上,本发明还可以考虑人们使用的实际术语:他们搜索类似的术语?类似的文档?类似的搜索术语,或混合?这样,本发明考虑术语使用。另一个考虑是用户的专业领域。目前考虑两个人具有专业技术向量,并且他们的专业技术向量也是术语向量。它是代替表示文档的术语向量,表示一个人和他知道的东西。它可能来自身份或可能基于他们所使用的而自动检测。

[0279] 专业技术。给定用户知道最多的是他/她的专业技术。本系统可以基于他们使用的资源和他们趋向花费时间的主题而自动确定用户的专业技术。专业技术被验证。本发明考虑一个人的收集。当用户搜索文档时,我们问全球群体他们怎么认为那些文档,不是用户说它们是什么,这是用户的术语使用,但是全球群体说这些文档是什么是大约。当考虑在给定用户的集合中与资源关联的全球术语向量的结合时,对于用户的专业技术是什么产生图片(企业可以由术语向量表示)。用户不能自己声称他的专业技术是什么。其他用户的群体最终确定专业技术向量。为了识别专家,本系统考虑专业技术向量。例如,如果用户在搜索芯片设计,本系统考虑每个用户的专业技术向量,并且找出谁是关于芯片设计的专家。本系统选择,例如前30位专家。然后本系统为每个文档找到那30位专家的术语向量和活动向量,总和到一起,然后执行比较。

[0280] 一个可选择和称赞的确定用户专业技术的方法首先识别在集合中对于给定兴趣的题目具有高影响因素的那些文档。资源影响因素是资源对特定的群体多有用的均衡。一旦在整个集合中为资源计算影响因素,可以按照包括的资源的影响因素而估计每个用户的库。采用这种方法,在他们使用的资源的集合中具有相对很多高影响资源的用户被认为是给定题目的专家。这样的用户也可以被分配专家影响因素,反映对给定题目的给定群体,该用户的资源集合的影响。

[0281] 如果用户是关于特定的文档,基于或者用户的全球群体、对等者群体、或专家群体而为该文档采用术语向量,该用户可以询问什么文档类似于此,例如询问“更像这样”。本系统可以比较该文档向量和每一个其它文档的术语向量。在这种情况下,本发明考虑术语向量,其由组关于术语对特定空间的相关性确定。因此,在术语向量之上存在第二测量,即,相关性的测量。因此可以说这个文档在这个空间是相关的,不仅是它具有这些共同的词。

[0282] 当采用术语向量和活动向量执行搜索时,用户得到返回给他的最有用的文档。本系统还可以说既然找到这些文档,向量是已知的,可以离开然后找到额外的可能有兴趣的文档并且建议那些。这是扩展搜索结果列表的方法。其中经常发生这样的环境是当用户在导航并且发现某个页面时。用户打开最接近这个的文档。用户点击他喜欢的文档,然后说这是最接近的,所有表示我喜欢这个。事实上,得到返回的已经在最初的搜索列表中,但是可能还有一些新的东西。

[0283] 本系统以用户特定的方式基于使用而执行导航追踪。这样,本发明还可以追踪人们去哪里,从一个文档到另一个。用户可以结束去特定文档寻找信息,但是然后可以在找到有用的文档之前点击多个文档。如果一个或多个用户跟随这相同的模式,本系统可以推荐登陆最初文档的用户立即到有用的文档。即,本发明在被发现有用的文档之间建立关联,即使其他文档可能由用户沿着导航路径已经遇到。这样,本发明推荐直接到有用的文档,没有使用户导航通过可能正常干扰的文档。这是基于使用:对于各用户,存在一个矩阵表示访问的文档和从那个位置到达的最有用的文档之间的连接。如本系统的其它方面,我们结合用户的意见以得到集体的意见,例如用户的对等者、专家或全球群体的意见,在这种情况下,考虑哪里是从当前位置去的最有用的地方(资源)。对于每个用户,例如,我们可以得到每个对等者的矩阵并且将其加到一起,然后提出建议。以这种方法,本发明追踪用户对等者的导航模式。除了提供建议,还可以在全球或群定特定的基础上,使用确定的导航模式来提供在资源集合中用户活动的可见性。例如,这样的可见性或使用地图对于网站设计者理解他们网站的效力和理解用户的兴趣将非常有益。

[0284] 人们关心例如在此公开的系统是当系统开始推荐文档时趋向于增强他们的有用性。本发明采用确认技术解决此问题。例如,如果存在本系统加强的流行(fad),例如人们突然去特定文档,但是他们不再回来,例如流行电影。每个人都去是因为他们听说它很好,但是事实上,它很糟而且人们讨厌它。很多人去该文档,但是没有人回来。本发明通过观察人们回来的百分比而调整文档的有用性,并且确定是否有足够的人回来以证实它是合理有用的文档。关于做决定,在一实施方式中,如果某些东西是新的,并且它确实在开始得到一些注意,在某种意义上,本系统鼓励这种注意,因为它可能是某些新的而且重要的东西。但是,如果随着时间过去,人们不开始回来,它开始快速衰退。因此,本发明同时考虑新的意见和确认的意见。

[0285] 除了连接用户从文档到文档,本发明还使用导航以发现识别文档是关于什么的信息。当某人点击链接而直接到文档并且使用它时,这告诉本系统用户点击该链接认为他得到什么并且然后使用它。无论链接文本是什么,它是关于文档是什么的相当好的反映。然后本系统使用链接文本并且现在那个链接文本也对术语向量有贡献。好像该链接是查询的替换,实际上,如果用户已经点击了去往该文档的所有十个链接,则本系统以某个权重使用该链接文本,因为不是每个链接具有相同的权重。这取决于链接多么接近该文档。如果用

户点击文档中的一个单词并且点击另一链接,然后下一文档中的另一链接,则最近的链接得到最高的权重,而较早的链接得到较小的权重。这样,在本实施方式中,权重基于接近性的形式。

[0286] 本发明提出的导航的另一方面涉及用户以特定文档开始的地方,例如文档 1,并且本系统基于这个开始点而进行各种推荐。如果用户先在不同的文档,例如文档 13,而不是文档 1,则本系统可以推荐一组不同的文档。在这种情况下,本发明使用附加模型,其考虑本系统从例如文档 13 推荐的文档并考虑从文档 1 推荐的文档,并且本系统一起均衡它们。这样本系统可以使用用户的导航踪迹(包含一个或多个导航点)来建议下一个要去的最好位置。熟悉本领域的技术人员可以理解,存在可用于处理本系统向量信息的许多选择。

[0287] 本发明的一个方面涉及确定什么是成功使用的文档。一个方法考虑某人花费在该文档多少时间。关于此有两个概念:一个是文档处理时间的概念。我们能够推断的是某人实际花费在查看和阅读该文档的时间。这与根本没有阅读文档相对比,它可能只是在屏幕上,但是用户没有看它;或者用户在寻找什么、搜索,但是没有找到。文档处理时间是成功使用的最简单均衡,因为本系统只需要考虑某人在一个文档上花费多少时间。另一个考虑是寻找时间和处理时间是不同的。这样,用户在周围滚动而没有找到他们想要的,即,他们没有处理它。本系统可以得到花费在文档上的时间,减去用户在文档上滚动的时间。本系统还可以使用滚动作为用户实际在该文档上的显示。本系统可以应用滚动和暂停的结合并且用其得到用户实际处理文档多长时间的感觉。因此,如果某人滚动,然后他们停止,并且待在那里 30 秒钟,然后他们又开始滚读,则本系统可以猜想用户读那个文档 30 秒钟。任何猜想中的错误是总计的,因为本系统很少考虑一个人的意见,而是在一组用户中总计该信息。

[0288] 术语向量是每个可能术语的大向量。在目前的优选实施方式中,每次术语通过用户的行为而与文档关联时,本系统增加术语的向量条目。不与文档关联的术语得到零。另外,通过使用与很多文档关联的术语在术语向量中得到少的权重,因为它们非常普通。这样,本系统考虑所有公知的术语并且考虑那些术语与多少文档关联。与很多文档关联的术语具有术语与文档相关的文档的数目。本系统基于术语与很多文档的关联而应用公式来降低术语的级别。如果有一个单词在集合中的每一个单个文档中存在,则它的级别等于零。某些词,例如“该”和“和”是标准停止词,在它们甚至开始系统分析之前就从搜索中移除。

[0289] 这样,本系统基于数据结构从初始术语向量的产生开始。对于系统中的每一个用户,存在被称为术语文档矩阵的文档的矩阵和相关的术语。这是对于系统公知的各文档的用户的术语向量的集合。换句话说,每个用户有代表各文档的用户认为该文档关于什么的术语文档矩阵。例如,一个人认为一个文档关于油和关于提炼厂,而不是其它东西。本系统服务已经选择的特定群体,例如其可以包括对等者或专家,例如这个人的前 30 位对等者。为了执行比较,本系统知道这 30 个用户,并且每个用户基于他们对于当前用户是多重要的对等者有一个权重。具有最高权重的对等者是第一对等者,具有次高权重的对等者是下一个最好的对等者,等等。本发明考虑所有这些对等者的术语文档矩阵并且将其加到一起,基于他们的对等者分数而给出权重,并且产生单个术语文档矩阵,其是对于系统中每个文档该群体的意见。然后本系统得到该矩阵并且计算表示查询的术语向量或当前环境与为各文档表示术语向量的矩阵中的各行之间的余弦。余弦计算的结果表示根据用户的对等者,该文档如何接近地与环境匹配。

[0290] 一旦本系统已经决定群体中所有的术语向量以及对于各文档分配了数字,然后本系统选择最前的文档。这样做的一个方法是选择前十个文档然后将其加到一起以得到单个向量,其总体说明这些文档关于某个主题。然后,本系统得到那些已经使用的搜索术语并且考虑其它的顶点在哪里。那些是本系统或者想要建议或者自动进入核心的额外的术语。现在有一个新的术语向量,并且本系统经历相同的过程。然后本系统可以将该新的术语向量与每个其它的文档匹配,得到一组新的分数,并且选择那些文档的前十个。本系统具有包括每个人对这些文档的意见的矩阵。本系统现在可以比较这个新的术语因数并且为他们的全部得到一组新的分数。本系统还进行,我们得到用其它分数表示这个文档对题目如何相关的各文档的单个向量,该分数表示这个文档多有用。

[0291] 一旦本系统已经决定群体中所有的术语向量并且给各文档分配数字,则本系统选择最前的文档。这样做的一个方法是选择前十个文档然后将其加到一起得到单个向量,其总体说明这些文档是关于某个主题。然后,本系统得到那些已经使用的搜索术语并且考虑其它的顶点在哪里。那些是本系统或者想要建议或者自动进入核心的额外的术语。现在有一个新的术语向量,并且本系统经历相同的过程。然后本系统可以将新的术语向量与每个其它的文档匹配,得到一组新的分数,并且选择那些文档的前十个。本系统具有包括每个人对这些文档的意见的矩阵。本系统现在可以比较这个新的术语因数并且为它们的全部得到一组新的分数。本系统还进行,我们得到用其它分数表示这个文档对题目如何相关的各文档的单个向量,该分数表示这个文档多有用。

[0292] 考虑非常受欢迎的文档,非常不受欢迎的文档、有些受欢迎的文档和从来没有用过的文档。每个用户有一个文档向量或他们的活动向量,其识别什么文档已经被使用。每个用户基于对等者群体还有一个相关的权重。为每个文档给定这两个数字,即它匹配的如何和它如何受欢迎,本系统结合这两者以产生一个分数。结合它们的一个方法是计算权重的和。另一个方法是保留这些数字,但是移除任何低于门限的数值。在第一种方法中,当系统将它们加到一起,文档的排序被改变;在第二种方法中,保留了作为相关性的排序,但是本系统移除不符合一定准确性门限的内容。在另一个方法中,不是直接组合,而是应用趋向于比喜欢极端更喜欢平衡数字的变换。例如,直接线性组合喜欢极端但是平方根组合可以产生更平衡的结果。一旦本系统组合存在组合的向量,就存在排序,并且系统可以推荐例如前 10 个文档。一旦文档返回用户,下一个部分观察用户对文档做什么。例如,本系统考虑这样的动作是用户以某种方式的导航,在链接要去的文档保留一段时间。本系统收集该信息以学习文档的有用性。用户从进行搜索一个题目开始,例如油。本系统通过推荐某些文档而响应。用户点击一个文档并且打印它。在用户的术语文档矩阵中,本系统为连接到这些单词的文档加上一个数字。如果其他人进行搜索并且包括该术语文档,则根据某个人为某种目的,该矩阵表示该文档是相关的。这样,如果这个人是对等的,并且文档是相关文档,该术语匹配良好,则该文档被推荐。如果用户不是密切的对等者或该文档没有很好的术语匹配,或者该文档不是很活动,则它不被推荐。

[0293] 如上讨论,有两种向量,起表示基于使用的相关性和活动性。当这两种向量与搜索引擎结果组合时还可以产生第三种向量。这样,这三种事情组合到这种向量中。有很多种方法可以将他们组合以从搜索引擎确定如何权重搜索结果相对如何权重其它结果。一个方法是得到 IR 结果的列表并且从该列表中移除准确性在确定门限以下的所用内容。这产生

扩大的搜索,其中通过移除不准确的结果而扩大搜索。

[0294] 另一方法包括涉及对等者、专家和全球群体的单独的向量,然后将其与不同的权重结合,例如专家得到最高权重,权重可以基于身份或者可以基于用户对一系列问题的反应。

[0295] 如上所述,每个用户具有捕获用户认为每个文档关于什么的术语文档矩阵,并且他们具有表示用户使用这些文档多少的活动向量。这个活动向量不仅通过搜索使用。它可以通过导航使用并且基于搜索术语或链接术语建立。确定对等者和专家按照如下方式进行:

[0296] 对于给定用户,建立那个用户的专业技术的图片。通过全球群体或通过合适的对等体组而验证专业技术。在第一种情况中,全球群体具有术语文档矩阵,其表示全球群体关于每个文件的意见。这本质上是每个单个用户关于该文档的相等权重的意见总和。这是全球的意见。对于每个用户,考虑那个用户使用过什么文档,以及他们使用这些文档的数量。这一步包括确定这个用户的专业技术。例如,这个用户已经以例如四的权重使用了文件一,因此当本系统进入文件一时,它决定全球群体认为这个文档是关于什么。得到那个之后,将其乘以四,然后将其加到用户的专业技术向量。

[0297] 如果全球群体认为那是重要的文档,并且如果用户已经使用它很多次,则该用户具有更多的专业技术。本系统在这个用户的集合中为每个文档做这些。用户使用最多的东西根据它们的专业技术会得到最高的权重。每次本系统添加全球群体考虑的用户已经使用的文档。这样,专业技术是用户的集合表示什么专业技术的均衡。本系统不知道用户实际的专业技术是什么。可能是某人已经完成收集关于这个题目所有正确的文档的极好的工作,但是如果他已经那么做了,则在某种意义上,他实现作为专家的目的。即,如果用户有关于那个题目所有好文档,则那个收集是专家收集。权重的量来给定文档的受欢迎性是个问题。权重的量给定一个文档由这个用户如何使用并且权重的量给定该文档在群体中如何受欢迎。例如,每天晚上本系统组合这些数字并且重新计算专业技术。这样,由于可能改变,本系统在一些基础上例如每天或每月,重新计算每个人的专业技术的领域。本系统检查和计算每个人的专业技术领域,然后如果理想的指出专家是谁,给定特定的查询,则本系统得到查询向量并且将该查询向量与每个用户的专业技术向量进行比较。然后,本系统可以产生前 N 个专家,这就是专家群体。另一种情况发生,其中本系统没有查询而有一个文档,但是用户想知道专家是谁。在这种情况下,本系统可以使用文件本身来确定专家是谁。这样,文档本身具有向量,并且本系统可以将这个文档的向量与每个人的专业技术向量进行比较,并且给出这个文档的题目,确定由这个文档表示的这个题目的专家是谁。

[0298] 对等者。每个用户有一个术语文档矩阵和活动向量。有三种事情系统可以考虑并且组合来确定对等者。一个是比较对例如两个用户对等者值是什么。本系统为每个人做这个决定,但是现在集中在两个用户。考虑一个用户的活动向量和另一用户的活动向量,并且考虑他们如何类似。使用类似文档类似次数的两个人是类似的,在相同的地方看并且相对成比例的量。这样,一个用户的活动向量和另一个用户的活动向量之间存在类似的度量。确定对等者的另一个方法是看他们对什么题目感兴趣。为了完成这,对他们的术语文档矩阵加和,其给出他们已经搜索和过去使用的什么题目的感觉。总和表示这个人什么感兴趣,并且将其称为兴趣向量。本系统比较兴趣向量。第三,本系统可以比较每个用户的计算的

专业技术向量来确定对等者。作为选择,本系统可以采用这些方法的结合。因为用户头脑中具有特定的主题并且用户具有某个数字,根据他们如何接近匹配那个特定的人而权重对等者。某些对等者可以具有接近用户的数字的数字。某些将会有更小或更大的数字,取决于系统使用的符号。最后,有一个数字表示用户有多像这个人。用户可能想要 30 人的对等者组,然后在组中数字 30 可能具有更小的权重,而组中的数字 1 可能具有更大的权重,并且之间每个人具有之间的权重。本系统还可以具有不产生任何少于门限的对等者的门限。

[0299] 图 12 示出了根据本发明的扩大的搜索的流程图。在扩大的搜索中,有客户库的客户进行搜索请求。搜索被发送到搜索服务器而扩展对于扩大的信息进行请求,例如从 Google。扩大的结果被返回到服务器并且接收的结果被添加到服务器信息中,该服务器信息然后以搜索服务器形式被发送回搜索服务器。然后客户接收该搜索所呈现的 HTML。

#### [0300] 基于时间的有用性

[0301] 如在该应用程序中的其它地方所述,本系统的每个方面随着进行新的观察报告而随着时间的适应和改变,处理新的日志文件,并且出现新的模式。这种适应的一个方面包括为最近发生的使用模式给出比那些过去发生的优先权。在本系统的优选实施方式中,基于时间衰退函数通过衰退过去使用模式而完成崭新性偏见。例如,活动向量可能以每天 0.01% 的速率衰退,从而发生在过去的活动对活动向量的影响小于最近的使用。类似的,术语向量可以设置为以特定速率衰退,从而关联资源的术语朝鲜更最近的使用模式偏移。这样,资源的有效性可以以时间敏感的方式计算。

[0302] 在过去有用的资源不一定在现在有用。存储在系统中的所有信息,包括对等体分数和专业技术分数,可以以类似的方式被设置为时间衰退。考虑活动向量,非常新的或重新发现的资源可能需要向上推进并且超越崭新性偏见,使得他们由群体的发现优先于有机会出现的强使用模式。这样,非常新的资源,被定义为非常最近的活动在所有时间构成大部分他们的全部活动的那些资源,可以被给予附加的新鲜性偏见。对于管理员,也可以显式的给某些资源或资源的集合分配新鲜性偏见。这种新鲜性偏见使得非常新的资源在短时间内出现的比他们实际上更活跃。例如,当它们在一年特殊时间再次出现时,还可以确定资源的周期性使用并且给予资源以活动性偏见。

#### [0303] 基于使用的术语和短语估计

[0304] 本发明的这方面涉及本系统基于捕获的使用数据推断的术语之中和术语和短语之中的关系。首先,可以建立术语密切关系(相似性)矩阵,其互相关联术语和短语。互具有高密切关系的术语和短语被认为是单个题目的表示,并且可以甚至是互相的同义词。例如,可以基于术语在用户的查询或使用的链接中术语共同出现的频率,或者在资源的术语向量中术语共同出现的频率而构造术语密切关系矩阵。这个矩阵,结合术语和短语本身的语言特征,可以用于自动识别同义词、缩写词和原短语。原短语是排列的两个或多个经常一起出现的单词,表示他们应该被认为是单个多词短语而不是多个独立的单词。术语密切关系矩阵结合导航使用模式和资源的术语向量甚至可以用于检测其它术语和短语的子标题的术语和短语。由于术语/短语之间所有这些识别的关系和同义词、缩写词以及原短语的自动检测基于团体的使用,识别的关系是固有地适合特殊的团体。

#### [0305] 本发明的目标应用

[0306] 除了上述讨论,本发明在用于商务网站的市场前端、销售和频道合作外延网、客户

支持网站；垂直保健应用程序，例如医生入口和患者搜索网站；垂直政府应用程序，例如市民入口；和金融服务及保险垂直应用程序，例如代理和顾问入口也有用。

[0307] 虽然这里说明的本发明参照优选实施方式，熟悉本领域的技术人员应该理解在不背离本发明的精神和范围的情况下其它应用程序可以代替这里说明的那些。因此，本发明应该仅由下面包括的权利要求书限定。

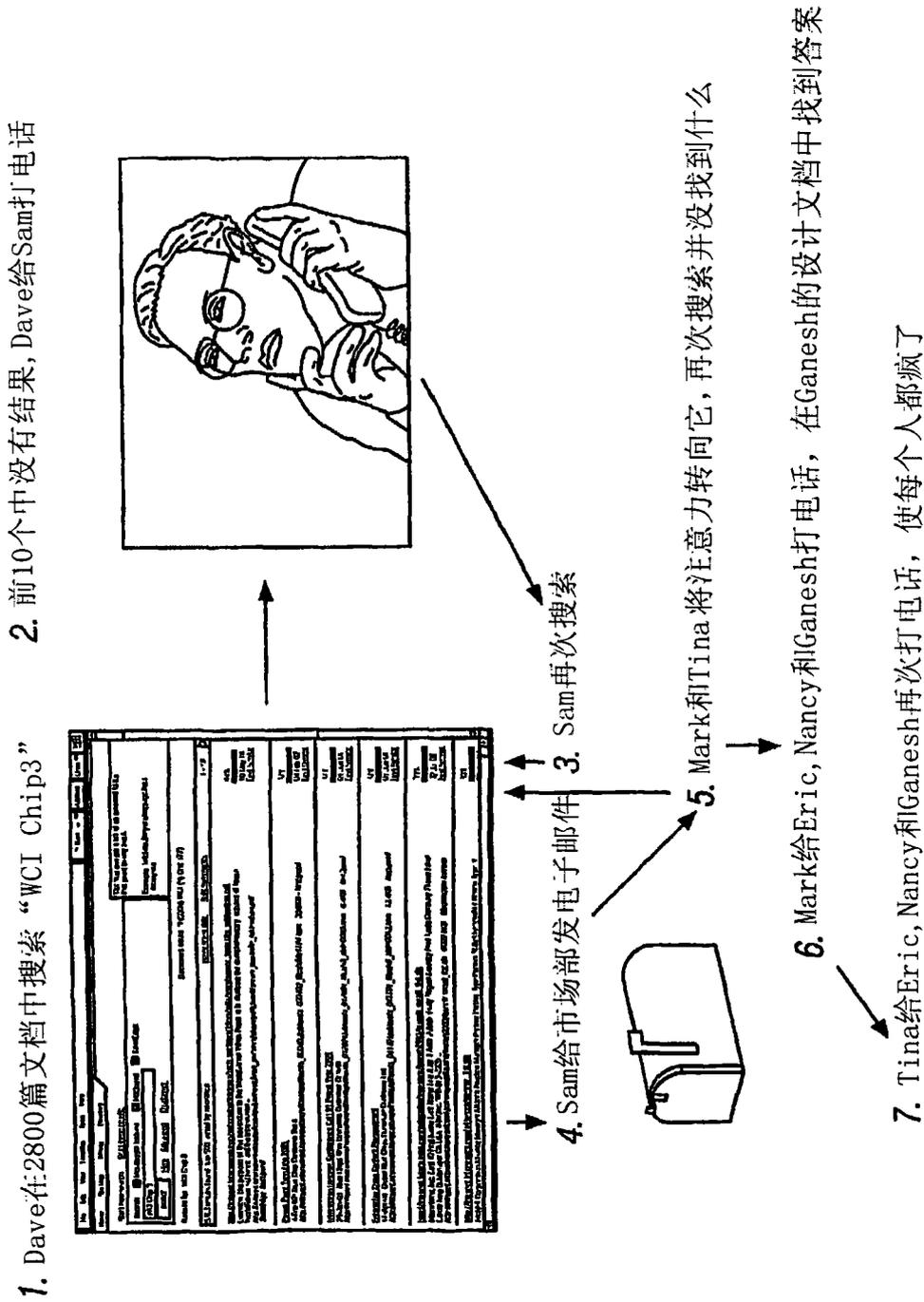
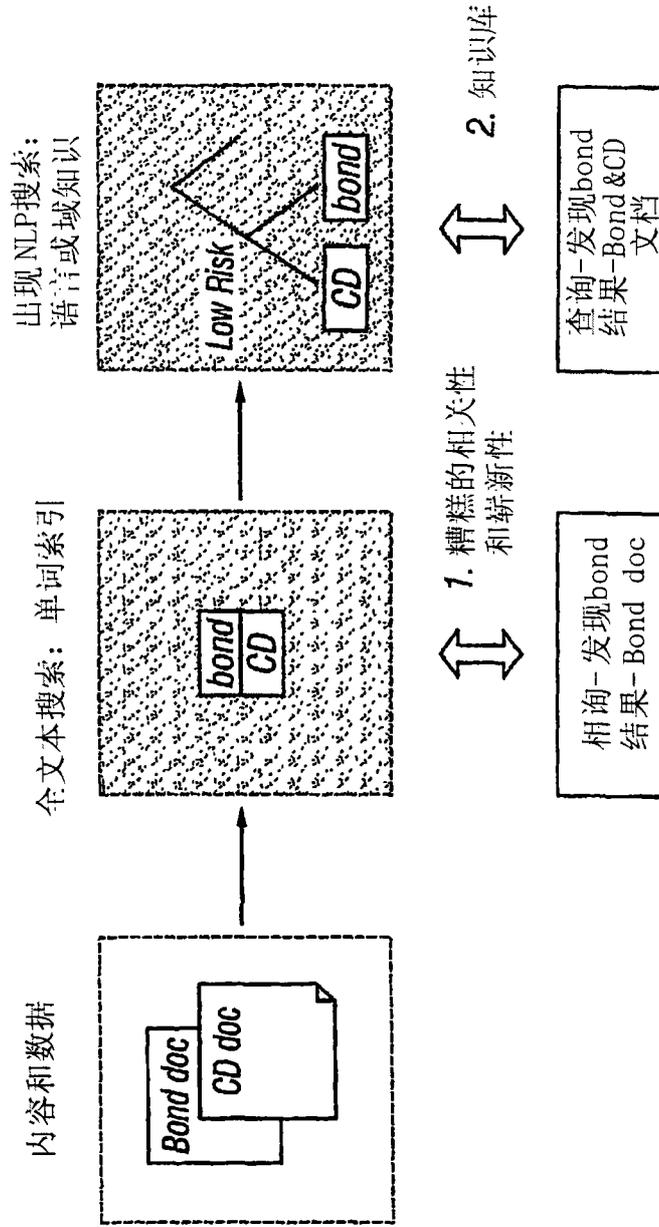


图1  
(现有技术)

今天的基于信息检索的搜索模式的问题



信息寻找者

图2

(现有技术)

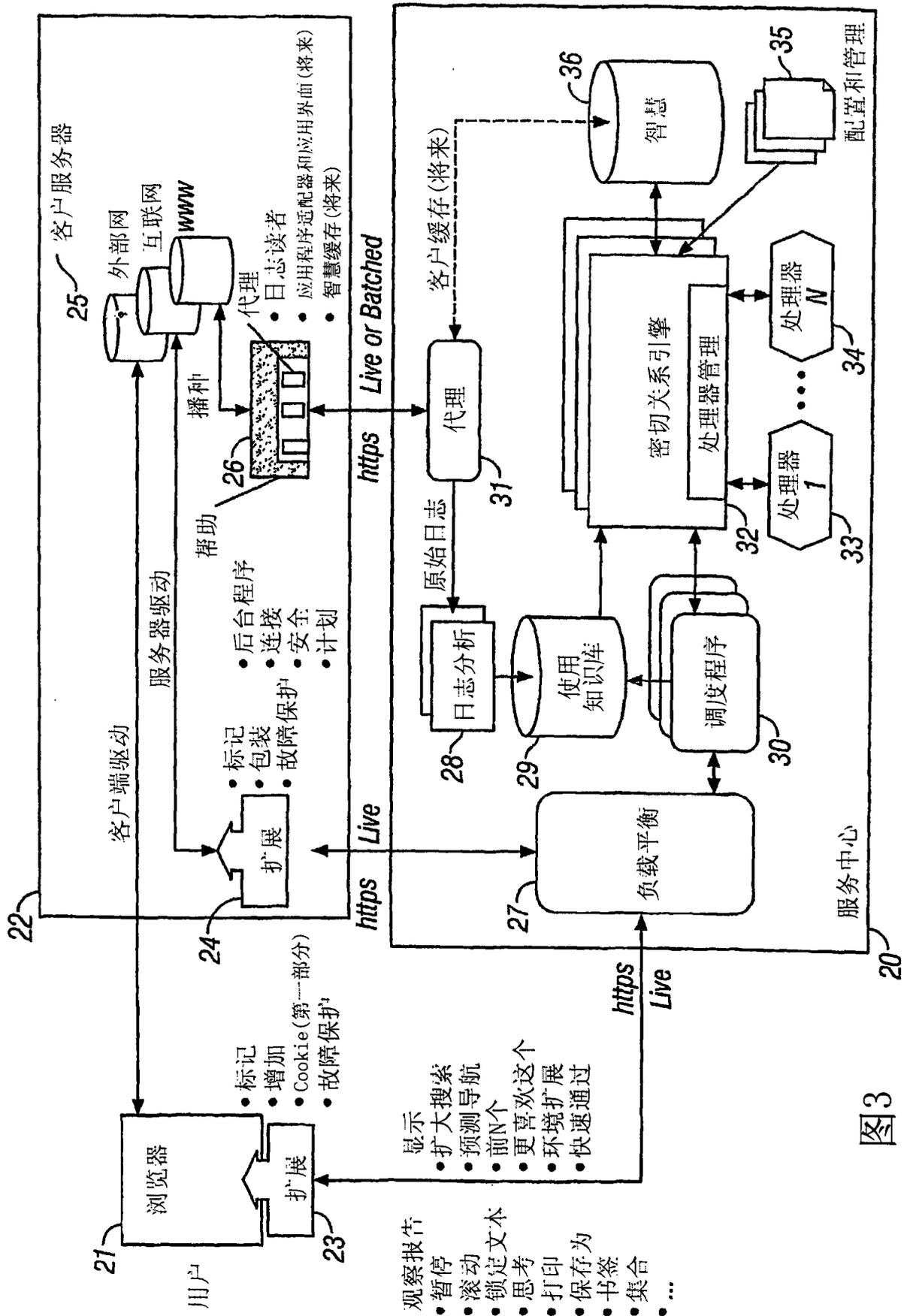


图3

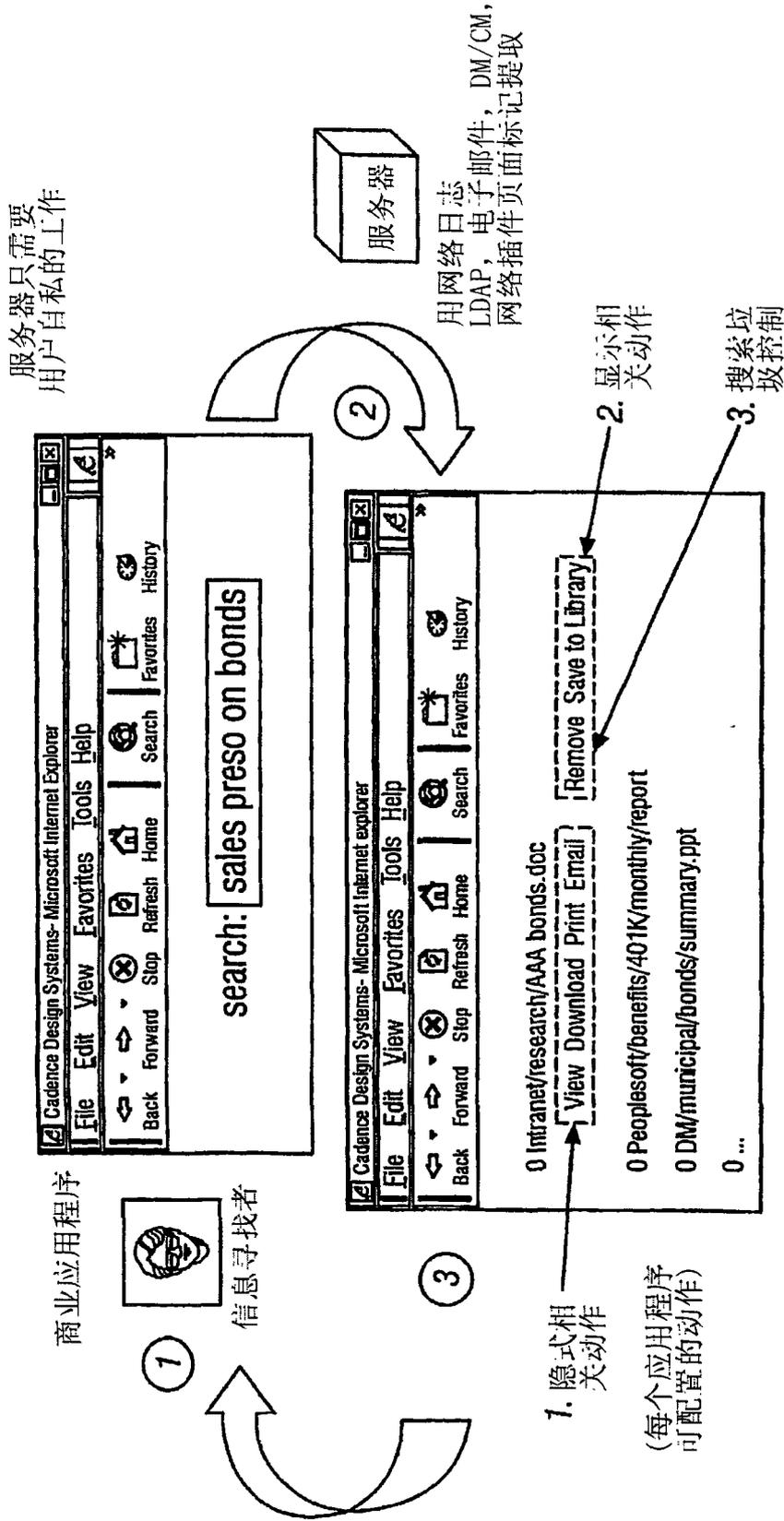


图4

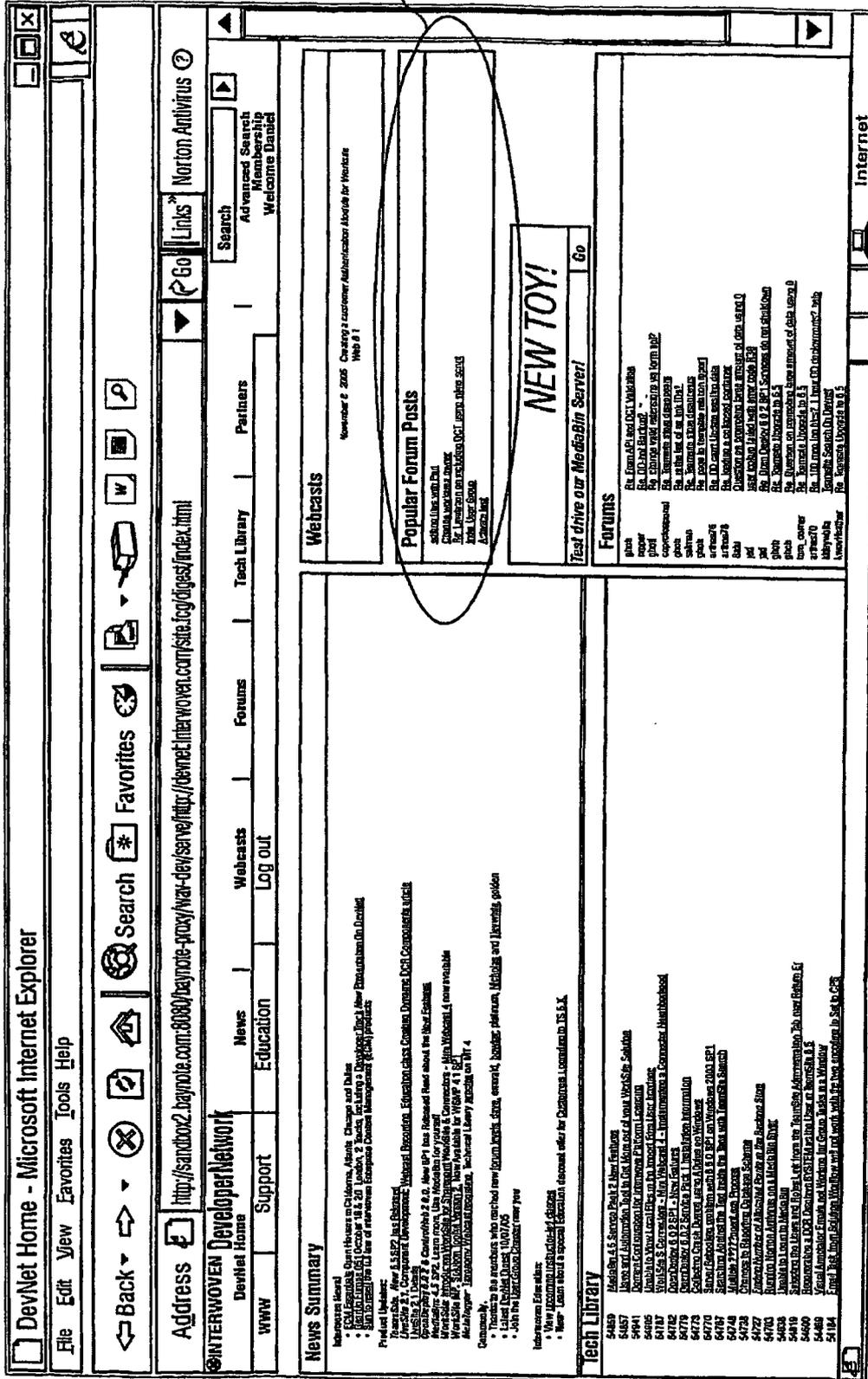


图5



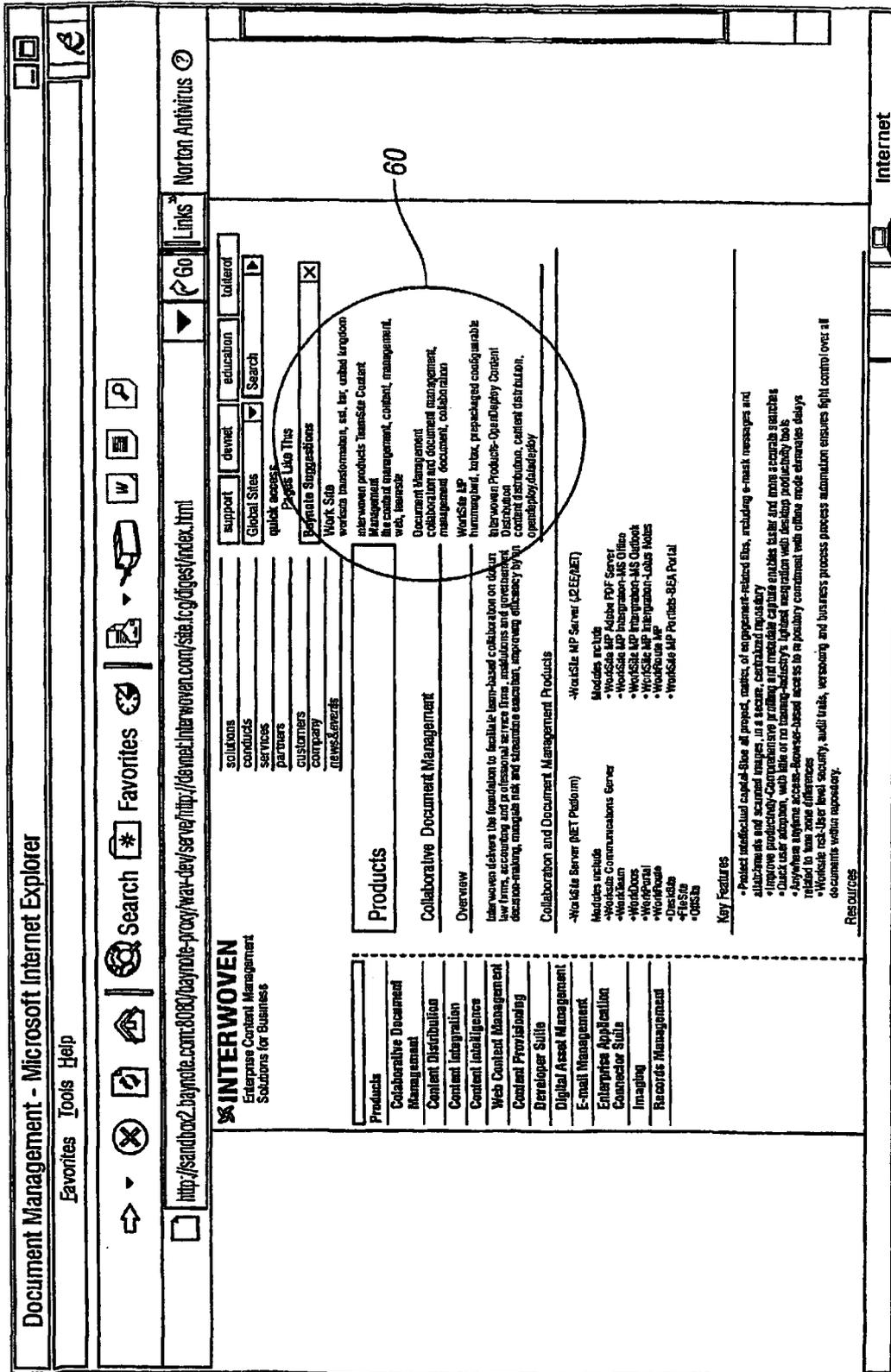


图7

域专业技术网络 (DEN)	内容重叠	映射重叠
个人收集	100%	20%
对等者收集	60%	40%
专家收集	50%	30%
团体收集	30%	10%

其它应用程序使用服务器的网络效果

我的库/行为日志

- 嵌入其它应用程序(不是新UI或应用程序)
- 通过搜索&发现产生
- 一般对用户可见
- 改善质量; 桥知识库
- 搜索“垃圾”控制
- 支持动态个人导航
- 近似哈希值, 装载率, 保密政策
- 浏览器和桌面插件
- 内容更新和缓存

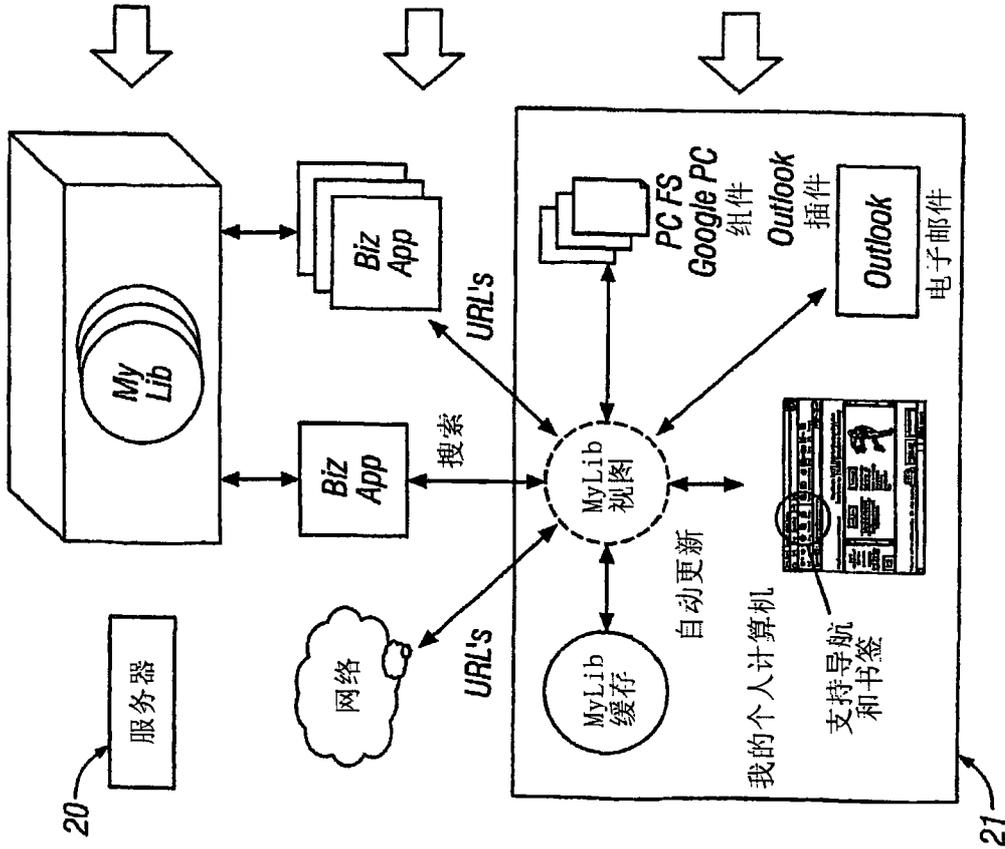


图8

WCI Extranet Portal My Library • Logout • Help

My Library for Dave Advanced **Sep. 15, 2004**

---

My Collections
Peer Collections
Expert Collections

---

**Content I Can Modify**

Dave PC > My documents > Wireless Products

- Airwave Product Executive Preso
- G10 phone design spec v5.0
- Remote DVD design spec v2.1

Airwave Email Server Archive > Dave Email > Critical

- Design Alerts

Airwave WCM Server > Dave WA > My Published

- G5 Installation

Airwave DM server > Consumer Products

- Remove DVD system layout v11

My Top 10  
My Most Recently Used  
My Highly Ranked

**Content I Can View**

Airwave > Products > G10 Phone

- MRDs
- Engineering Schedule
- Old Design Document v7.2

WCI > SupportI

- Customer Faq
- Training Books
- Developer Network

Cadence.com > Products > Custom IC

- System-on-Chip Solution
- Wireless Design Platform
- Analog-Digital Mixed Signal Tools

IEEE.org > Working Groups

- 802.11 standards

My Top 10  
My Most Recently Used  
My Highly Ranked

图9

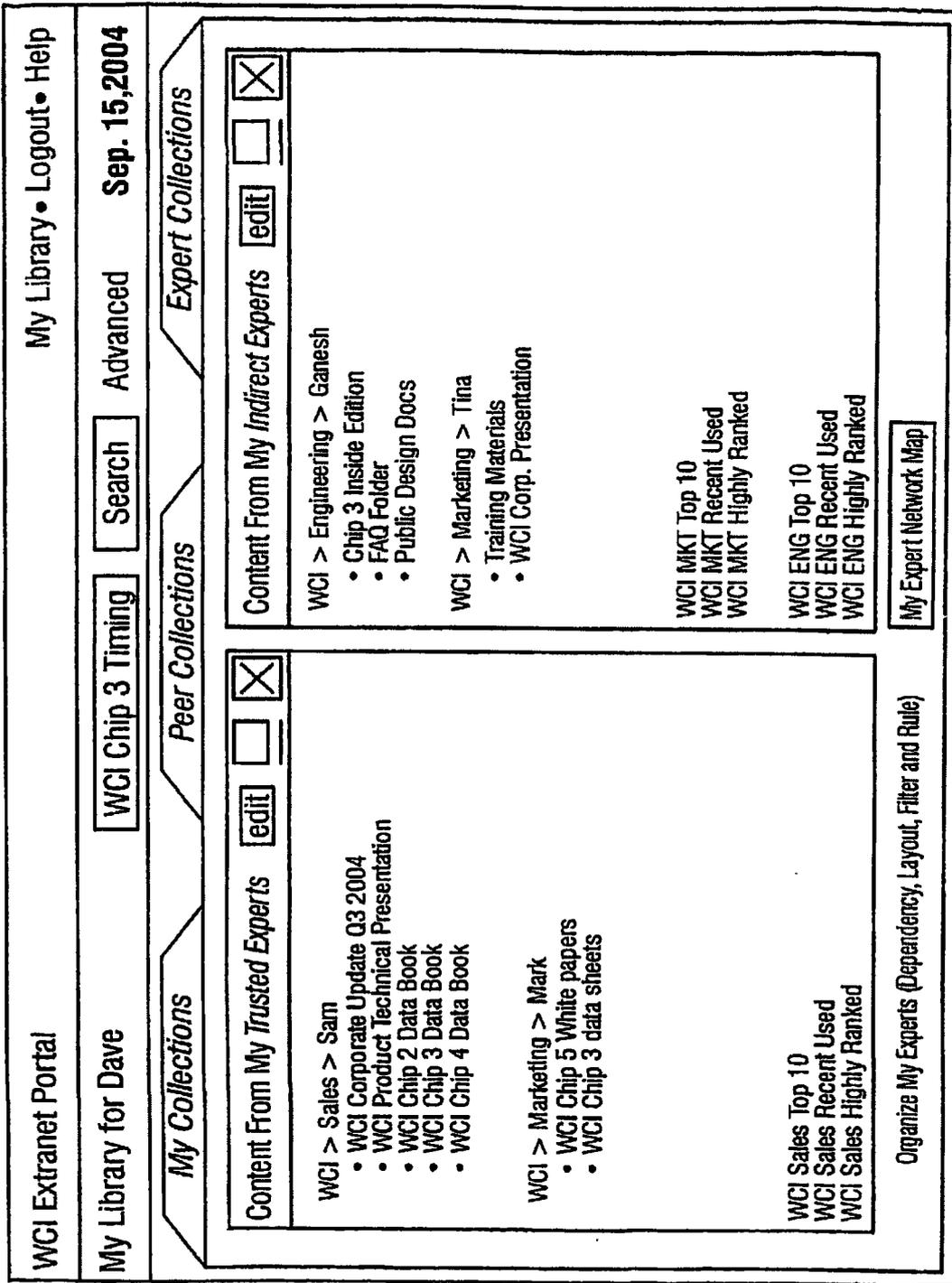


图10

WCI Extranet Portal My Library • Logout • Help

My Library for Dave WCI Chip 3 Timing  Advanced **Sep. 15, 2004**

Mr. Collections Dear Collections Expert Collections

### Dave's Peer and Expert Network

Name	Dept	Peer	Expert	Indirect Expert Depth	Ranking	Content Threshold	R/W
Doug	Design	X		1	2	50	R/W
Mark	MKT		X	1	3	50	R/W
Sam	Sales		X	1	1	100	R
Steve	Design	X		1	1	20	R/W

Organize My Experts (Dependency, Layout, Filter and Rule)

图 11

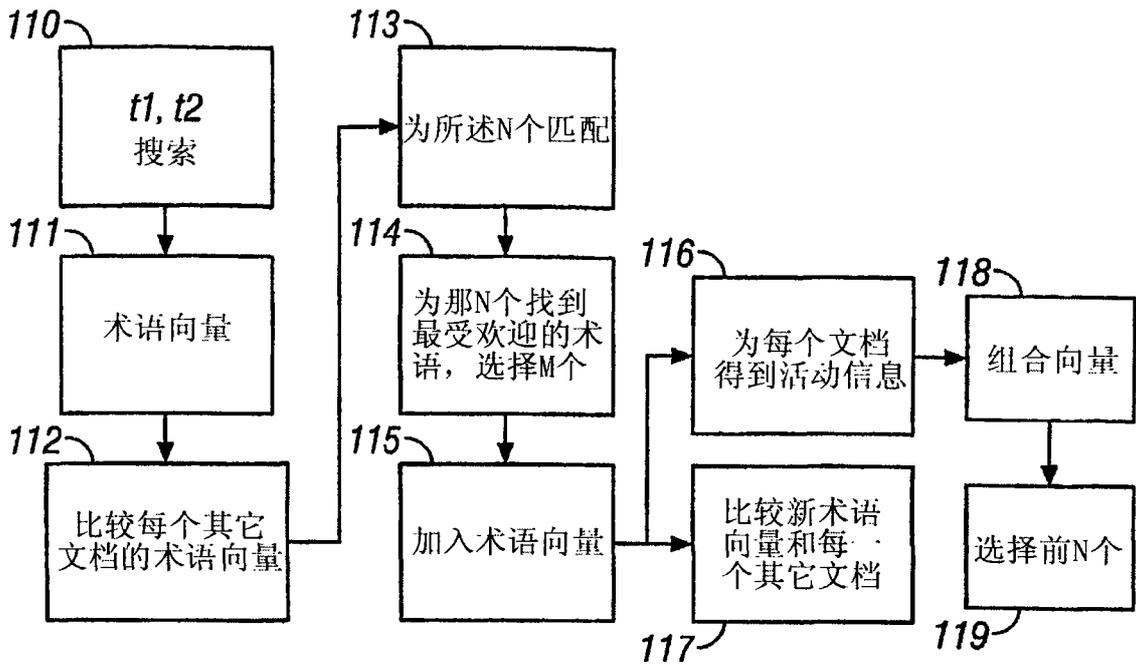


图 12

扩大搜索

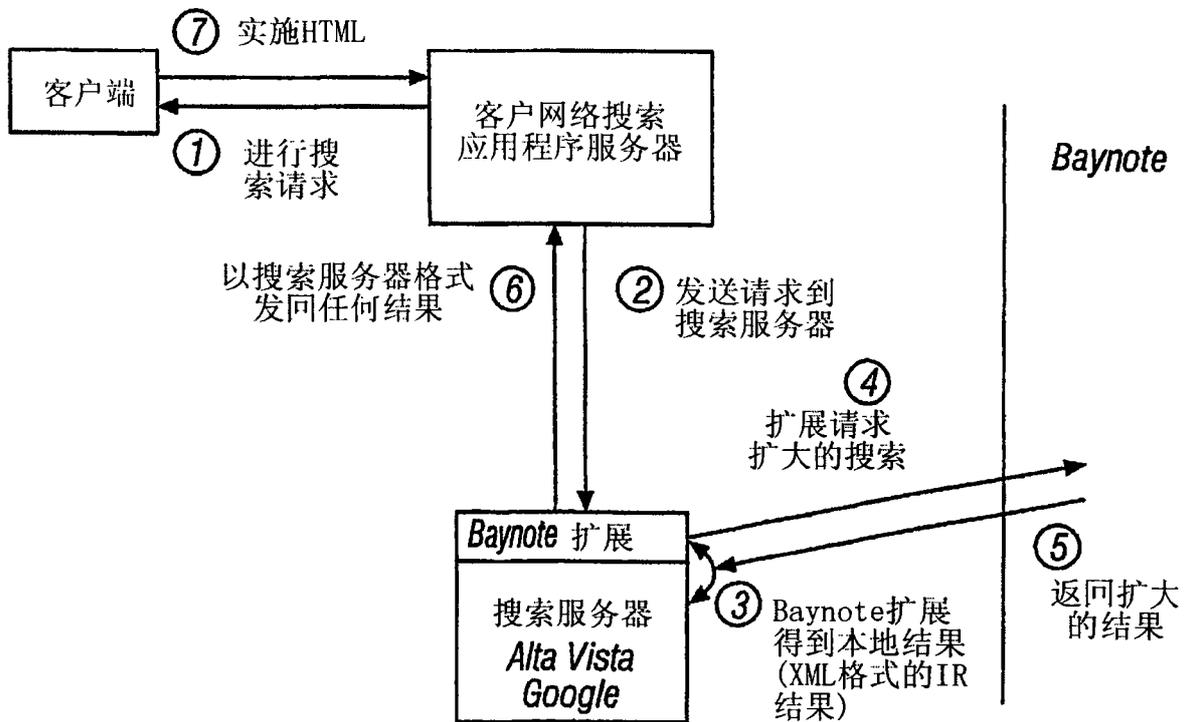


图 13