基于量子计算的用户识别算法

朱皖宁[1],刘志昊[2]
(1. 金陵科技学院软件工程学院,江苏南京 210000;2. 东南大学计算机科学与工程学院,江苏南京 210096)

摘 要: 本文提出了基于量子算法的快速用户识别算法. 当代社会进入互联网时代后,大量的信息充斥在网络上,许多有价值的信息被隐藏在 Weblog 中,大数据分析的一项任务就是通过对 Weblog 的分析得到用户行为模式等重要的信息,在这之前必须要做的是对用户进行识别. 以往对用户识别算法的研究较为侧重在准确度方面,识别的速度尚不能令人满意. 本文基于 Grover 搜索算法提出了扩展记录模式和非扩展记录模式的两种快速 IP 地址搜索算法,将搜索的查询复杂度进行了二次加速.

关键词: 用户识别; 量子计算; 大数据; Grover 搜索算法; 无结构数据库搜索
中图分类号: TP387; TN911.73 **文献标识码:** A **文章编号:** 0372-2112 (2018)01-0024-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.01.004

User Identifying Algorithm Based on Quantum Computing

ZHU Wan-ning[1], LIU Zhi-hao[2]
(1. *Institute of Software Engineering, Jinling Institute of Technology, Nanjing, Jiangsu* 210000, *China*;
2. *Institute of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu* 210096, *China*)

Abstract: This paper presents an IP address finding algorithm based on improved Grover algorithm. At present, Internet is full of massive information. The weblogs contain lots of valuable information that must be analyzed for useful detection like behavior pattern of user. And the user identifying is the previous work. In the past researching of user identifying algorithms, most results focus on the accuracy of identifying user instead of the performance. This paper shows two IP address quick searching algorithms, namely record expansion searching algorithm and record non-expansion searching algorithm based on Grover searching algorithm. The query complexity of the record non-expansion searching algorithm gets quadratic acceleration.

Key words: user identify; quantum computing; big data; Grover searching algorithm; searching on unstructured database

1 引言

当今世界已经进入一个信息爆炸的时代,而互联网处于信息交互的一个核心的位置. 多种不同的数据和信息在互联网上交叉并存储在社交网络上,从而促使多种不同的研究领域都开始关注社交网络,例如社会学、经济学、计算机科学等[1~3]. 社交网络上的用户通过访问网页来获得需要的资源,研究社交网络的关键是在于分析这些社交网络是如何被使用的[4]. 分析出来的数据可以用来改进社交网络本身,让用户更加方便的浏览所需要的数据;也可以用于分析用户的喜好,从而进行广告定点投送;还可以用于分析用户的行为,对此用户参与的交易进行预测[5]. 分析出这些结果的主要方法之一就是对这些网站的 Weblog 进行大数据分析.

每当一个用户请求一个页面或者是这个页面中的一些资源,例如视频,声音等,在网站的 Weblog 中就会新添一条记录. 这些信息里蕴含了用户最喜欢的页面(即访问次数最多)、普通页面的访问序列甚至还暗示了这个用户本身的一些特点[6,7]. 这种信息分析手段可以称为基于 Weblog 信息的网页使用数据挖掘(Web Usage Mining, WUM). 使用 WUM 需要先抽取单个用户与

收稿日期:2016-05-20;修回日期:2016-10-26;责任编辑:孙瑶
基金项目:金陵科技学院高层次人才科研启动基金(No. jit-b-201624);国家自然科学基金(No. 61502101);江苏省自然科学基金(No. BK20140651);南京信息工程大学 PAPD 和 CICAEET 资助

网页交互的序列,这个序列所形成的文件中需要至少包含以下字段:用户的 IP 地址、时间戳、请求的资源、操作结果的代码、进入此网页前的前一个网页地址和使用的浏览器. 使用和分析这个序列,可以得到此用户访问网页的模式[8~10].

为了便于数据分析,单个用户与网页交互的序列还需要进行进一步的划分,这个过程称为会话划分(Sessionization). 而会话划分的前提是已经做好了用户识别工作(User Identifying). 在以往对此问题的研究中,学者们已经提出了很多算法,黄健青等人提出了使用二分查找算法进行用户识别工作,将效率提高了 5 倍[11];侯枫提出了构造序列集识别用户的算法[12];纪良浩等人提出了基于协作过滤的 Web 日志数据预处理过程结构图和一种可行的数据预处理方法,提高了用户识别的准确度[13];邹根等人提出了基于支持向量机的用户识别算法,从 Cookies 中提取用户的以往记录,提高了用户识别的准确度[14];Stenmark 等人提出了利用一种特殊的搜索引擎来区分不同行为用户的族[15];Santra 等人则提出了使用贝叶斯算法来识别兴趣用户,大大提高了识别效率,其效率与网站深度成反比[16]. 但是以上的研究成果仍然有一定缺陷,文献[12-14]较为关注用户识别的准确度,而对于海量的 Weblog 来说,其查询效率必须要同时被关注;文献[11]对查询效率进行了一定的改进,但是常数系数的效率提高仍然不能满足查询的需要;文献[15]主要研究的是用户的行为以便于以后对于行为模式进行发现和验证,但是无法查询某一特定的兴趣用户;文献[16]局限于某一特定网站的 Weblog,但是在当今的大数据分析中,往往是使用多个网站的 Weblog 进行联合分析.

对于用户识别算法来说,不仅仅是进行会话划分之前的预处理过程,还需要能够快速查找特定的兴趣用户记录. 本文提出了一种基于量子计算机制的用户识别算法,先使用量子搜索算法在海量的 Weblog 中快速查找用户的 IP 地址,在查询出的小集合中再运用以往的文献算法进行二次分类,不仅保证了用户识别的准确率,同时大大提高了查询的效率.

2　预备知识

量子计算源于上个世纪 80 年代,是一门新兴的交叉学科,利用量子力学来解决计算问题. 在上个世纪 90 年代末时,多个高效率量子算法的提出,让量子计算被人们重视起来,其中就有在无结构数据库中进行快速搜索的 Grover 搜索算法[17]. 在随后的十几年中,量子计算飞速发展,尤其是今年墨子量子卫星的上天标志着量子计算与量子通信的研究进入到了一个新的领域. 尽管当前的量子计算机硬件技术尚不足以支撑大规模的数据运算,但是科技革命的规律表明了科学革命可以领先于技术革命. 因此研究如何用量子算法来解决当前的经典问题有极其重要的价值和意义. 本节将会简单介绍量子计算的基本原理以及 Grover 搜索算法.

量子系统是由量子比特表示,一个量子比特由希尔伯特空间中的一个向量表示,最简单的量子比特表示如下:

$$\begin{cases} |0\rangle = [1 \quad 0]^{\mathrm{T}} \\ |1\rangle = [0 \quad 1]^{\mathrm{T}} \end{cases} \tag{1}$$

与经典比特不同的是,量子系统可以同时以不同的概率处于不同的状态上,这种状态称为量子叠加态,例如系统状态 $|\psi\rangle$ 以 α^2 的概率为 $|0\rangle$,以 β^2 的概率为 $|1\rangle$,那么可表示为:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{2}$$

其中,$\alpha^2 + \beta^2 = 1$,即满足归一化. 对 $|\psi\rangle$ 做操作会同时对 $|0\rangle$ 和 $|1\rangle$ 进行操作,这就是量子算法的高度并行特性. 1997 年,Grover 提出了利用量子的并行性原理进行无结构数据库上搜索的算法[17],将时间复杂度降低到了 $O(\sqrt{N})$. Grover 搜索算法的核心是迭代如图 1 所示的 U 算子.

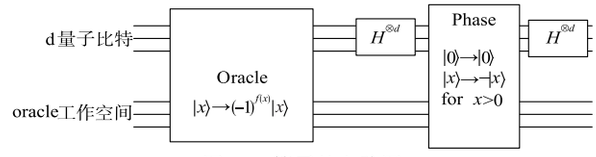

图1　U算子的电路图

输入是由 d 个量子比特叠加成的均匀叠加态:

$$|s\rangle = \frac{1}{\sqrt{D}} \sum_{i=0}^{D-1} |i\rangle \tag{3}$$

其中 $D = 2^d$,为搜索空间. 如图 1 所示的 U 算子可以分成两个部分:Oracle 算子和均值反演算子. 其中 Oracle 算子令所求目标分量相位翻转:$\text{Oracle} = I - 2|\varphi\rangle\langle\varphi|$;作用在搜索空间为 D 的均值反演算子通过简单计算可以得到如式(4)表示:

$$\begin{aligned} G_D &= H^{\otimes d}(2|0\rangle\langle 0| - I)^{\otimes d} H^{\otimes d} \\ &= 2H^{\otimes d}(|0\rangle\langle 0|)^{\otimes d} H^{\otimes d} - H^{\otimes d} I^{\otimes d} H^{\otimes d} \\ &= 2|s\rangle\langle s| - I \end{aligned} \tag{4}$$

设 M 为搜索目标的数量,那么迭代 $O\left(\sqrt{\frac{D}{M}}\right)$ 次后对系统状态进行测量,就能以超过 $\frac{1}{2}$ 的概率得到解.

在之前的工作中,对 Grover 搜索算法进行了优化,解决了算法在 M 未知情况下无法确定迭代次数的问题,为本文解决用户识别问题提供了算法基础[18].

3 基于 Grover 搜索算法的用户识别算法

考虑以下场景:网络警察发现某个匿名用户在某网站上发布了匿名的信息,而此信息违法. 警察只能查到发布者的 IP 地址,希望通过扫描此 IP 地址的所有会话序列以发现其他的犯罪事实. 在此场景中,并不需要对所有的用户进行会话划分,只需要针对某个特定的 IP 地址用户进行会话划分即可. 在做会话划分之前,必须先将所有包含此 IP 地址的 Weblog 记录抽取出来. 现在一个网站的 Weblog 记录条目数多达几十万条,如果要对多个网站的记录进行联合搜索,其数量更是可能达到几百万条. Weblog 中的记录按照请求时间进行排序,对于 IP 地址或者 Agent 字段,既没有建立索引,也没有经过排序,即这个搜索问题是一个无结构数据库上的搜索问题,其效率非常低下. 在经典的求解方法中,只能按照顺序一条一条的进行比对,假设总记录条目的数量为时,搜索到所有包含某一特定 IP 地址的记录,需要的时间复杂度为 $O(N)$. 下面本文将会介绍如何使用 Grover 搜索算法降低搜索特定 IP 地址所需的时间复杂度.

3.1 扩展记录模式的 IP 地址搜索算法

假设共有 N' 条记录,设 n 满足 $2^{n-1}\leqslant N'\leqslant 2^n$,$N=2^n$. 将搜索空间扩展到 N,扩展的记录 IP 地址设为本机地址 127.0.0.1,设 $\mathrm{IP}_{\mathrm{Local}}=00\ 111111\ \underbrace{00\cdots01}_{23}$. 由于访问网站的 IP 地址中不会有此 IP 地址,因此不会影响到搜索结果. 设序号空间 H_{SN} 为 n 维希尔伯特空间,IP 地址空间 H_{IP} 为 32 维希尔伯特空间,因此量子搜索空间 $H_S=H_{SN}\otimes H_{IP}$,为 $32n$ 维希尔伯特空间. 设空间 H_{SN} 中的一个向量为 $|x_i\rangle$,空间 H_{IP} 中的一个向量为 $|y_j\rangle$,因此量子搜索空间 H_S 中的一个向量为 $|x_i\rangle\otimes|y_j\rangle$. 设装载算子 S_L 作用如下:

$$|i'\rangle\otimes|y_j\rangle^{\otimes 32}\xrightarrow{S_L}|i'\rangle\otimes|IP\rangle \tag{5}$$

式(5)中 $|0\rangle^{\otimes 32}$ 为对 $|0\rangle$ 自身张量 32 次的结果,IP_i 为第 i 条记录的 IP 地址. 由于搜索目标数量未知,因此需要使用文献[18]算法来进行搜索;并且搜索算法是在 IP 地址空间中进行,所以需要对文献[18]中的算法结构进行一定改动才可以使用,改动过的量子搜索算法如图 2 所示.

如图 2 所示的 U′算子电路,Oracle 算子作用在 IP 地址空间和 Oracle 工作空间上,第一个 Phase 判断相位和的门电路作用在序号空间和判断终止位上,第二个 Phase 进行均值反演的门电路作用在序号空间和 Oracle 工作空间上.

扩展记录模式的 IP 地址搜索算法如算法 1.

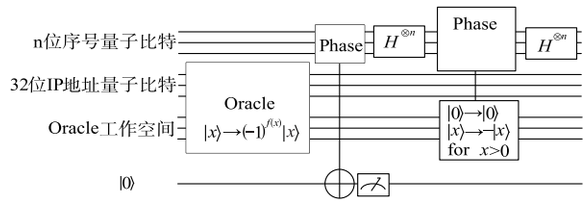

图2 改进后的U′算子电路图

算法 1 扩展记录模式的 IP 地址搜索算法

设修正算子 $S_{adj}=\sum_{i=0}^{N-1}|i\rangle\langle i|\otimes\sum_{j=0}^{2^{32}-1}|j\rangle\langle j|$,标记 Flag = True.

当 Flag 为真时循环以下过程:

1. 制备初态 $|\phi_0'\rangle=|0\rangle^{\otimes n}\otimes|0\rangle^{\otimes 32}$.
2. 对初态 $|\phi_0'\rangle$ 作用算子 $H^{\otimes n}\otimes I^{\otimes 32}$,得到量子态 $|\phi_0''\rangle$.
3. 对 $|\phi_0''\rangle$ 作用装载算子 S_L 和修正算子 S_{adj},得到量子态 $|\phi_0\rangle$.
4. 对 $|\phi_0\rangle$ 作用 U′算子最多次得到终态 $|\phi_{\mathrm{final}}\rangle$.
5. 测量 $|\phi_{\mathrm{final}}\rangle$,得到塌缩的状态 $|k\rangle\otimes\langle p|$,若 p 为所搜索地址,则记录 k 并将修正算子中 $|k\rangle\langle k|\otimes\sum_{j=0}^{2^{32}-1}|j\rangle\langle j|$ 分量改为

$$\begin{aligned}&|k\rangle\langle k|\otimes|IP_{\mathrm{Local}}\rangle\langle p|+|p\rangle\langle k|\\&\otimes|p\rangle\langle IP_{\mathrm{Local}}|+|k\rangle\langle k|\\&\otimes\sum_{j:0\rightarrow 2^{32}-1\text{且}j\neq p,IP_{\mathrm{Local}}}|j\rangle\langle j|\end{aligned}\tag{6}$$

6. 若 p 不是所搜索地址,则令 Flag = False.

循环结束.

算法分析:

定理 1 修正算子 S_{adj} 在每次修改后都为酉算子.

证明 在没有修改时:

$$\begin{aligned}S_{adj}\circ S_{adj}^{\dagger}&=\sum_{i=0}^{N-1}|i\rangle\langle i|\sum_{i=0}^{N-1}|i\rangle\langle i|\otimes\sum_{j=0}^{2^{32}-1}|j\rangle|\langle j|\\&\quad\cdot\sum_{j=0}^{2^{32}-1}|j\rangle\langle j|\\&=\sum_{i=0}^{N-1}|i\rangle\langle i||i\rangle\langle i|\otimes\sum_{j=0}^{2^{32}-1}|j\rangle\langle j||j\rangle\langle j|\\&=\sum_{i=0}^{N-1}|i\rangle\langle i|\otimes\sum_{j=0}^{2^{32}-1}|j\rangle\langle j|=I^{\otimes 32n}\end{aligned}$$

因此在没有修改时 S_{adj} 为酉算子. 在每次修改后,都是对某个序号 $|k\rangle$ 的 IP 地址 $|p\rangle$ 和本机地址 $|IP_{\mathrm{Local}}\rangle$ 做一个置换操作. 对于每个序号来说,其置换操作都是相同的,且作用在序号空间的算子没有发生改变,所以只需要证明式(6)所示的作用在 IP 地址空间算子是酉算子即可证明修正算子为酉算子:

$$\begin{aligned}&\left(|IP_{\mathrm{Local}}\rangle\langle p|+|p\rangle\langle IP_{\mathrm{Local}}|+\sum_{j:0\rightarrow 2^{32}-1\text{且}j\neq p,IP_{\mathrm{Local}}}|j\rangle\langle j|\right)\\&\cdot\left(|IP_{\mathrm{Local}}\rangle\langle p|+|p\rangle\langle IP_{\mathrm{Local}}|+\sum_{j:0\rightarrow 2^{32}-1\text{且}j\neq p,IP_{\mathrm{Local}}}|j\rangle\langle j|\right)^{\dagger}\end{aligned}$$

$$
\begin{aligned}
&=\left(|IP_{\text{Local}}\rangle\langle p|+|p\rangle\langle IP_{\text{Local}}|+\sum_{j:0\to 2^{32}-1\text{且}j\neq p,IP_{\text{Local}}}|j\rangle\langle j|\right)\\
&\cdot\left(|IP_{\text{Local}}\rangle\langle p|+|p\rangle\langle IP_{\text{Local}}|+\sum_{j:0\to 2^{32}-1\text{且}j\neq p,IP_{\text{Local}}}|j\rangle\langle j|\right)\\
&=|IP_{\text{Local}}\rangle\langle p||p\rangle\langle IP_{\text{Local}}|+|p\rangle\langle IP_{\text{Local}}||p\rangle\langle p|\\
&+\sum_{j:0\to 2^{32}-1\text{且}j\neq p,IP_{\text{Local}}}|j\rangle\langle j|\sum_{j:0\to 2^{32}-1\text{且}j\neq p,IP_{\text{Local}}}|j\rangle\langle j|\\
&=|IP_{\text{Local}}\rangle\langle IP_{\text{Local}}|+|p\rangle\langle p|+\sum_{j:0\to 2^{32}-1\text{且}j\neq p,IP_{\text{Local}}}|j\rangle\langle j|\\
&=\sum_{j=0}^{2^{32}-1}|j\rangle\langle j|=I^{\otimes 32}
\end{aligned}
$$

证毕

根据定理 1 可知,算法 1 为一个量子算法,因为每一步操作都是酉操作. 算法步骤 1 ~ 3 制备了可以作用 U′算子的初态,即序号状态为均匀叠加态,IP 地址状态为序号对应的 IP 地址. 在作用 U′算子之前,通过作用装载算子 S_L 使初态变成一个纠缠态,因此对于 IP 地址状态进行相位取反操作(Oracle 操作)就等于对序号状态做了相位取反操作,从而使得后面的均值反演操作可以正常进行.

在第一次循环时,修正算子实际为恒等算子,不做任何操作. 因此直接开始做改进的 Grover 搜索算法,假设搜索目标有 M 个,则根据文献[18,19]中的分析结果可知当迭代 $\left\lceil\frac{\pi}{4}\sqrt{\frac{N}{M}}\right\rceil$ 次时,搜索到目标的概率为近似于$\frac{N-M}{N}$,由于 $M\ll N$,所以最多迭代$\sqrt{N}$次 U′算子后,搜索到目标的概率近似于 1.

当每次循环结束后通过改变修正算子,将 IP 地址置换为本机地址. 在下次循环时,通过作用修正算子,将此序号对应的 IP 地址换成本机地址. 这样在进行 Grover 搜索时,将不会再搜索到已经搜索过的序号. 由于共有 M 个搜索目标,因此最多循环 M 次后算法结束. 算法查询复杂度为迭代 $\left\lceil\frac{\pi}{4}M\sqrt{\frac{N}{M}}\right\rceil=\left\lceil\frac{\pi}{4}\sqrt{MN}\right\rceil$,即 $O(\sqrt{MN})$.

3.2　非扩展记录模式的 IP 地址搜索算法

由于扩展记录的数量会导致额外的开销,因此本节将会讨论在不进行记录扩展的情况下如何进行 Grover 搜索.

虽然不再扩展记录,但是为了能够表示 N'条记录,仍然需要 n 个量子比特,即做 Grover 搜索算法的空间没有发生变化,为 $H_S=H_{SN}\otimes H_{IP}$,因此均值反演算子也没有变化,为 $2|s\rangle\langle s|-I$. 将初态制备成$\frac{1}{\sqrt{N'}}\sum_{i=0}^{N'-1}|i\rangle$,若直接运行 Grover 搜索算法将不会降低查询复杂度.

定理 2　在非扩展记录模式下直接使用 Grover 搜索算法,查询复杂度仍然为 $O\left(\sqrt{\frac{N}{M}}\right)$.

证明　设 H_{SN}的维度为 n,$N=2^n$,H_{SN}上的一个均匀叠加态$|s\rangle=\frac{1}{\sqrt{N}}\sum_{i=0}^{N-1}|i\rangle$,$H_{SN}$上的均值反演算子 G_N 表示如下:

$$G_N=2|s\rangle\langle s|-I^{\otimes n}=\frac{2}{N}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}|i\rangle\langle j|-I^{\otimes n}\tag{7}$$

现在考察进行一次 Grover 迭代后的结果. 设一个由 N'个分量组成的一般叠加态$|\varphi'\rangle\equiv\sum_{k=0}^{N'-1}\alpha_k|k\rangle$,其中假设 α_i 满足归一化. 当 Oracle 算子作用后即可获得新状态$|\varphi''\rangle=\sum_{k=0}^{N'-1}(-1)^{f(k)}\alpha_k|k\rangle$. 为了研究目标分量相位的变化,可以将目标分量(即令 $f(k)=1$)表示为 $\sum_{k'=0}^{N'-1}-\alpha_{k'}|k'\rangle$,一般分量表示为$\sum_{k''=0}^{N'-1}\alpha_{k''}|k''\rangle$. 现在将 Grover 均值反演算子作用在新状态$|\varphi''\rangle$上:

$$
\begin{aligned}
&(2|s\rangle\langle s|-I^{\otimes n})\sum_{k=0}^{N'-1}(-1)^{f(k)}\alpha_k|k\rangle\\
&=\frac{2}{N}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}|i\rangle\langle j|\sum_{k=0}^{N'-1}(-1)^{f(k)}\alpha_k|k\rangle\\
&-\sum_{k=0}^{N'-1}(-1)^{f(k)}\alpha_k|k\rangle\\
&=\frac{2\sum_{k=0}^{N'-1}(-1)^{f(k)}\alpha_k\langle k|k\rangle}{N}\sum_{i=0}^{N-1}|i\rangle\\
&-\left(\sum_{k'=0}^{N'-1}-\alpha_{k'}|k'\rangle+\sum_{k''=0}^{N'-1}\alpha_{k''}|k''\rangle\right)
\end{aligned}
\tag{8}
$$

如式(8)所示,$\sum_{k=0}^{N'-1}(-1)^{f(k)}\alpha_k\langle k|k\rangle$为叠加态所有分量的相位和,但是均值是按照 N 个分量来进行计算,这里设$\langle\alpha\rangle_N=\frac{\sum_{k=0}^{N'-1}(-1)^{f(k)}\alpha_k\langle k|k\rangle}{N}$,那么式(8)可变换为如下形式:

$$
\begin{aligned}
&\left(\sum_{k'=0}^{N'-1}2\langle\alpha\rangle_N|k'\rangle+\sum_{k''=0}^{N'-1}2\langle\alpha\rangle_N|k''\rangle+\sum_{k'''=0}^{N'-1}2\langle\alpha\rangle_N|k'''\rangle\right)\\
&-\left(\sum_{k'=0}^{N'-1}-\alpha_{k'}|k'\rangle+\sum_{k''=0}^{N'-1}\alpha_{k''}|k''\rangle\right)\\
&=\left(\sum_{k'=0}^{N'-1}2\langle\alpha\rangle_N+\alpha_{k'}\right)|k'\rangle+\left(\sum_{k''=0}^{N'-1}2\langle\alpha\rangle_N-\alpha_{k''}\right)|k''\rangle\\
&+\sum_{k'''=N'}^{N-1}2\langle\alpha\rangle_N|k'''\rangle
\end{aligned}
\tag{9}
$$

式(9)中$\sum_{k'''=N'}^{N-1}2\langle\alpha\rangle_N|k'''\rangle$为非搜索空间中的分量,容易发现虽然在搜索目标前的相位也进行了幅度扩大,但

是其增幅最大为 $2\langle\alpha\rangle_N$,在文献[18]已证明,第一次迭代时 $\langle\alpha\rangle_N$ 的值最大,为 $\frac{1}{\sqrt{N}}-O\left(\frac{M}{N\sqrt{N}}\right)$,其中 M 为搜索目标的数量. 而目标态的相位增加幅度是每个目标分量增加幅度归一化的结果:

$$\begin{aligned}&\sqrt{\frac{4M}{(\sqrt{N})^2}-O\left(\left(\frac{M}{N\sqrt{N}}\right)^2\right)}\\&\approx 2\sqrt{\frac{M}{N}}-O\left(\frac{M}{N\sqrt{N}}\right)\\&=O\left(\sqrt{\frac{M}{N}}\right)\end{aligned}$$

因此需要 $O\left(\sqrt{\frac{N}{M}}\right)$次迭代,才能以比较大的概率测量到搜索目标.

证毕

通过定理 2 可知,在非扩展记录模式下直接使用 Grover 搜索算法,不仅不会提高搜索效率,还会变换出非搜索空间中的状态. 为了能够提高效率,构造新的均值反演算子如下:

$$G_{db}=\frac{2}{N'}\sum_{i=0}^{N'-1}\sum_{j=0}^{N'-1}|i\rangle\langle j|-I^{\otimes n}+2\sum_{k=N'}^{N-1}|k\rangle\langle k| \tag{10}$$

令非扩展记录模式下进行搜索的 Grover 迭代算子 $U_{db}=G_{db}\circ(I-2|\delta\rangle\langle\delta|)$,其中 $|\delta\rangle$ 为搜索的目标.

算法 2 非扩展记录模式的 IP 地址搜索算法

设修正算子 $S_{adj}=\sum_{i=0}^{N-1}|i\rangle\langle i|\otimes\sum_{j=0}^{2^{32}-1}|j\rangle\langle j|$,标记 Flag = True.

当 Flag 为真时循环以下过程:

1. 制备初态 $|\phi_0'\rangle=|0\rangle^{\otimes n}\otimes|0\rangle^{\otimes 32}$.
2. 对初态 $|\phi_0'\rangle$ 作用算子 $H^{\otimes n}\otimes I^{\otimes 32}$,得到量子态 $|\phi_0''\rangle$.
3. 对 $|\phi_0''\rangle$ 作用装载算子 S_L 和修正算子 S_{adj},得到量子态 $|\phi_0\rangle$.
4. 对 $|\phi_0\rangle$ 作用 U_{db} 算子最多 $\sqrt{\frac{N'}{M}}$ 次得到终态 $|\phi_{\text{final}}\rangle$.
5. 测量 $|\phi_{\text{final}}\rangle$,得到塌缩的状态 $|k\rangle\otimes\langle p|$,若 p 为所搜索地址,则记录 k 并将修正算子中 $|k\rangle\langle k|\otimes\sum_{j=0}^{2^{32}-1}|j\rangle\langle j|$ 分量改为
$|k\rangle\langle k|\otimes|IP_{\text{Local}}\rangle\langle p|+|k\rangle\langle k|$
$\otimes|p\rangle\langle IP_{\text{Local}}|+|k\rangle\langle k|$
$\otimes\sum_{j:0\to 2^{32}-1\text{且}j\neq p,IP_{\text{Local}}}|j\rangle\langle j|$.
6. 若 p 不是所搜索地址,则令 Flag = False.

循环结束.

算法分析:

容易发现算法 2 与算法 1 只有步骤 4 不同,因此只需要对步骤 4 进行分析即可. 首先需要证明新的 Grover 迭代算子 U_{db} 为酉算子,然后分析算法复杂度.

定理 3 在非扩展记录模式下迭代 $O\left(\sqrt{\frac{N'}{M}}\right)$次 U_{db} 算子后,能以比较大的概率测量到搜索目标.

证明 先证明 G_{db} 为酉算子.

注意到 $I^{\otimes n}$ 可以分成一个 N' 行 N' 列的恒等矩阵 $I_{N'}$ 加上 $\sum_{k=N'}^{N-1}|k\rangle\langle k|$,那么 G_{db} 可以简化为:

$$\begin{aligned}G_{db}&=\frac{2}{N'}\sum_{i=0}^{N'-1}\sum_{j=0}^{N'-1}|i\rangle\langle j|-I_{N'}-\sum_{k=N'}^{N-1}|k\rangle\langle k|+2\sum_{k=N'}^{N-1}|k\rangle\langle k|\\&=\frac{2}{N'}\sum_{i=0}^{N'-1}\sum_{j=0}^{N'-1}|i\rangle\langle j|-I_{N'}+\sum_{k=N'}^{N-1}|k\rangle\langle k|\end{aligned}$$

G_{db} 可以用分块矩阵表示,其形式如下:

$$\begin{bmatrix}\frac{2-N'}{N'} & \frac{2}{N'} & \frac{2}{N'} & & & & \\ \frac{2}{N'} & \frac{2-N'}{N'} & \frac{2}{N'} & \cdots & 0\ 0\ 0 & 0\ 0\ 0 & \\ \frac{2}{N'} & \frac{2}{N'} & \frac{2-N'}{N'} & & & & \\ \vdots & & & \ddots & & \ddots & \\ 0\ 0\ 0 & & & & 1\ 0\ 0 & & \\ 0\ 0\ 0 & & & \cdots & 0\ 1\ 0 & & \\ 0\ 0\ 0 & & & & 0\ 0\ 1 & & \end{bmatrix}$$

矩阵的左上角 N' 维方阵为 $\boldsymbol{G}_{N'}$,右下角 $N-N'$ 维方阵为 $\boldsymbol{I}_{N-N'}$,右上角和左下角的矩阵为 $\boldsymbol{0}$ 矩阵. 由于 $\boldsymbol{G}_{N'}$ 和 $\boldsymbol{I}_{N-N'}$ 均为酉矩阵,所以整个矩阵为酉矩阵. 又由于 $(I-2|\delta\rangle\langle\delta|)$ 为酉算子,因此 U_{db} 为酉算子.

假设初态为 $|\varphi'\rangle\equiv\sum_{p=0}^{N'-1}\alpha_p|p\rangle$,类似的,一般分量和目标分量都按照定理 2 证明中方法定义. 进行一次 U_{db} 迭代后:

$$\begin{aligned}&U_{db}|\varphi'\rangle\\&=\left(\frac{2}{N'}\sum_{i=0}^{N'-1}\sum_{j=0}^{N'-1}|i\rangle\langle j|-I_{N'}+\sum_{k=N'}^{N-1}|k\rangle\langle k|\right)\sum_{p=0}^{N'-1}(-1)^{f(p)}\alpha_p|p\rangle\\&=\frac{2\sum_{p=0}^{N'-1}(-1)^{f(p)}\alpha_p\langle p|p\rangle}{N'}\sum_{i=0}^{N'-1}|i\rangle-\sum_{p=0}^{N'-1}(-1)^{f(p)}\alpha_p|p\rangle\end{aligned}$$

设 $\langle\alpha\rangle_{N'}\equiv\frac{\sum_{p=0}^{N'-1}(-1)^{f(p)}\alpha_p\langle p|p\rangle}{N'}$,那么以上公式可简化为:

$$\sum_{p'=0}^{N'-1}(2\langle\alpha\rangle_{N'}+\alpha_{p'})|p'\rangle+\sum_{p''=0}^{N'-1}(2\langle\alpha\rangle_{N'}-\alpha_{p''})|p''\rangle \tag{11}$$

根据定理 2 证明中相同的计算方法可得,$\langle\alpha\rangle_{N'}=O\left(\frac{1}{\sqrt{N'}}\right)$,由归一化可知目标态在进行一次 U_{db} 迭代后幅度增大 $O\left(\sqrt{\frac{M}{N'}}\right)$,因此最多需要迭代 $O\left(\sqrt{\frac{N'}{M}}\right)$次 U_{db}

后能够以比较大的概测量到目标.

证毕

将新的 Grover 迭代使用到 IP 地址搜索中,得到非扩展记录模式的 IP 地址搜索算法. 由于其他部分都没有发生改变,只是通过对 Grover 迭代算子进行改进得到了更好的查询复杂度,因此其他部分的分析都不变. 根据定理 3 可知,最终的时间复杂度为 $O(MN')$.

使用算法 2,就可以快速的得到所需 IP 地址记录的序列,在此基础上,使用已有的经典算法,通过判断是否使用相同浏览器和操作系统进行进一步的用户识别.

4　小结

当今的网络社交媒体网站中蕴含了大量的信息,其中就包括用户行为模式这一重要信息,如果可以挖掘出这些知识,就可以实现定点广告投放,喜好推荐等增值服务. 进行网络大数据挖掘的前提是已经对网站 Weblog 中的数据进行了数据预处理,包括数据清洗、用户识别和会话划分等. 其中用户识别又是其中的基础步骤,能够迅速、准确的进行用户识别,是实现大数据分析的关键.

本文基于量子计算,提出了对兴趣用户进行快速识别的算法. 由于 Weblog 没有加入任何索引,在其中查找某条特定记录属于无结构数据库上搜索的问题. 量子算法中的 Grover 搜索算法可以高效的对无结构数据库进行搜索,本文基于这一点,提出了扩展记录模式的 IP 地址搜索算法,将 N'扩展到 N 再进行 Grover 搜索,并在搜索出一条记录后,通过作用修正算子将记录在量子态中消去,最终以 $O(\sqrt{MN})$的时间复杂度得到所需的 IP 地址记录序列;本文对扩展记录模式的 IP 地址搜索算法进行了改进,得到了非扩展记录模式的 IP 地址搜索算法,将时间复杂度降低到了 $O(\sqrt{MN'})$.

参考文献

[1] LIN S, JHENG Y, YU C. Combining ranking concept and social network analysis to detect collusive groups in online auctions [J]. Expert Systems with Applications, 2012, 39(10): 9079 – 9086.

[2] DEVI B N, DEVI Y R, RANI B P, RAO R R. Design and implementation of web usage mining intelligent system in the field of e-commerce [J]. Procedia Engineering, 2012, 30: 20 – 27.

[3] FONG A C M, ZHOU B, HUI S, TANG T, HONG G. Generation of personalized ontology based on consumer emotion and behavior analysis [J]. IEEE Transactions on Affective Computing, 2012, 3(2): 152 – 164.

[4] COOLEY Robert, MOBASHER Bamshad, SRIVASTAVA Jaideep. Data preparation for mining World Wide Web browsing patterns [J]. Knowledge and Information Systems, 1999, 1(1): 5 – 32.

[5] CHEN Min, MAO Shiwen, LIU Yunhao. Big data: A survey [J]. Mobile Networks and Applications, 2014, 19(2): 171 – 209.

[6] CHOI J, LEE G. New techniques for data preprocessing based on usage logs for efficient web user profiling at client side [A]. Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence And Intelligent Agent Technology [C]. Washington DC, USA: IEEE Press, 2009. 3. 54 – 57.

[7] NASRAOUI O, SOLIMAN M, SAKA E, BADIA A, GERMAIN R. A web usage mining framework for mining evolving user profiles in dynamic web sites [J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(2): 202 – 215.

[8] MAYER J R, MITCHELL J C. Third-party web tracking: Policy and technology [A]. Proceedings of the 2012 IEEE Symposium on Security and Privacy [C]. San Francisco, USA: IEEE Press, 2012. 413 – 427.

[9] INBARANI H H, THANGAVEL K, PETHALAKSHMI A. Rough set based feature selection for web usage mining [A]. Proceedings of the 2007 International Conference on Computational Intelligence and Multimedia Applications [C]. Sivakasi, TamilNadu, India, 2007, 1. 33 – 38.

[10] KOSALA R, BLOCKEEL H. Web mining research: A survey [J]. SIGKDD Explorations, 2000, 2: 1 – 15.

[11] 黄健青,黄浩. Web 日志分析中数据预处理的设计与实现[J]. 河南科技大学学报(自然科学版),2009,30(5):45 – 48.
HUANG Jianqing, HUANG Hao. Design and implementation of data preprocessing in web log analysis [J]. Journal of Henan University of Science & Technology (Natural Science), 2009, 30(5): 45 – 48. (in Chinese)

[12] 侯枫. Web 日志数据预处理[J]. 河南工程学院学报(自然科学版),2008,20(3):54 – 57.
HOU Feng. Web log data preparation [J]. Journal of Henan Institute of Engineering (Natural Science), 2008, 20(3): 54 – 57. (in Chinese)

[13] 纪良浩,王国胤,杨勇. 基于协作过滤的 Web 日志数据预处理研究[J]. 重庆邮电学院学报(自然科学版),2006,18(5):646 – 649.
JI Lianghao, WANG Guoying, YANG Yong. Research of web log data preprocessing based on collaborative filtering [J]. Journal of ChongQing University of Posts and Telecommunications (Natural Science), 2006, 18(5): 646 – 649. (in Chinese)

[14] 邹根,闻立杰. 基于支持向量机的 Web 日志用户标志修

正算法[J]. 计算机集成制造系统,2011,17(8):1851-1855.
ZHOU Gen, WEN Lijie. User identifier correction algorithm in web logs based on support vector machine [J]. Computer Integrated Manufacturing Systems, 2011, 17(8):1851-1855. (in Chinese)

[15] DICK Stenmark. Identifying clusters of user behavior in intranet search engine log files [J]. Journal of the American Society for Information Science and Technology, 2008,59(14):2232-2243.

[16] SANTRA A K, JAYASUDHA S. Classification of web log data to identify interested users using naïve bayesian classification [J]. International Journal of Computer Science Issues, 2012, 9:381-384.

[17] GROVER L K. Quantum mechanics helps in searching for a needle in a haystack [J]. Physical Review Ltters, 1997, 79(2):325-328.

[18] ZHU Wanning, CHEN Hanwu, LIU Zhihao, XUE Xilin. Grover algorithm for multi-objective searching with iteration auto-controlling [A]. Proceedings of the Fifth International Conference on Swarm Intelligence [C]. Hefei, China: ICSI Press, 2014, 8794. 357-364.

[19] BOYER M, BRASSARD G, HΦYER P, et al. Tight bounds on quantum searching [J]. Fortschritte Der Physik (Progress of Physiks), 1998, 46(4-5):493-505.

作者简介

朱皖宁(通讯作者) 男,1983 年 1 月出生,安徽淮南人. 2015 年毕业于东南大学计算机科学与工程学院,博士,讲师,主要研究领域为量子可逆逻辑综合与量子计算.
E-mail: granny025@163.com

刘志昊 男,1982 年 10 月出生,湖南邵阳人,博士,讲师. 2013 年毕业于东南大学计算机科学与工程学院. 主要研究领域为量子信息安全与量子计算.
E-mail: liuzhtopic@163.com