# Relatório do Trabalho de Recuperação da Informação

Alex Pungartnik Handel
Departamento de Ciência da Computação
Universidade Federal do Rio de Janeiro

## 1.Introdução

O artigo escolhido foi "Anonymizing Query Logs by Differential Privacy", escrito por Sicong Zhang, Hui Yang e Lisa Singh do Departamento de Computação da Universidade de Georgetown e publicado na conferência SIGIR em 2016. O artigo em si trata do problema de anonimização de registros de consultas (Query Logs) e propõe uma solução usando o conceito de privacidade diferencial, bem como testes. Nesse relatório proponho a explicar o problema de anonimização de registros de consultas, o que é e como é implementada a privacidade diferencial, o conceito de Utilidade de um registro de consultas e como ele se relaciona com o conceito de Privacidade de um registro de consultas e também descrever uma implementação de anonimização de registros de consultas como proposto no artigo.

## 2. Anonimização de registros de consultas

### 2.1 registros de consultas

Um registro de consultas é um dataset composto por um conjunto de Queries em algum serviço qualquer de recuperação de informação, geralmente de serviços de busca na web. O registro de consultas contém não apenas a query mas também seus resultados e outras informações como identificação do usuário, tempo até clicar em um resultado, queries que seguem a original e qual resultado foi clicado. Um registro de consultas usado muito em pesquisas é o registro de consultas da American Online (AOL) que foi disponibilizado pela empresa em 2006. Um exemplo de uma entrada nesse registro de consultas pode ser visto na tabela 1.

### 2.2 O problema de anonimização

Um registro de consultas é muito útil para pesquisas e pode ser usado como conjunto de testes para vários sistemas de RI, e registro de consultas disponibilizados por empresas de serviços na web são valiosos por conterem grande volume de informações coletadas de usuários reais. Porém existe um grande problema de segurança envolvido em sua divulgação: mesmo que nomes sejam omitidos, agentes maliciosos ainda podem identificar certos pelas suas buscas combinadas com outras informações. Assim, é necessária a implementação de algoritmos e processos para garantir a privacidade e anonimidade dos usuários envolvidos. Na divulgação do registro de consultas da AOL para pesquisa, tal registro de consultas não havia sido propriamente anonimizado, o que levou a um escândalo onde múltiplos usuários tiveram sua privacidade comprometida.[2]

Muitos métodos já foram propostos para anonimizar registros de consultas. Um deles utiliza de técnicas de clusterização e k-anonimidade, onde é assumido que cada query foi disparada por k diferentes usuários, porém estes métodos ainda eram vulneráveis a adversários que possuíam certas informações externas.

| AnonID | Query | QueryTime | ItemRank | ClickUrl |
|---|---|---|---|---|
| 1268 | ozark horse blankets | 2006-03-01 17:39:28 | 8 | http://www.blanketsnmore.com |

Tabela 1: formato de uma entrada no registro de consultas da AOL

## 3. Privacidade Diferencial

O método de anonimização proposto no artigo escolhido faz uso de privacidade diferencial.

Pela definição, uma função aleatória K provém ($\boldsymbol{\varepsilon}$, $\boldsymbol{\delta}$)-privacidade diferencial se para dois datasets $D_1$ e $D_2$ que diferem em no máximo um elemento e todo S $\subseteq$ imagem(K) se verifica a inequação $P[K(D1) \in S] \leq e^{\varepsilon} * P[K(D2) \in S] +$ , onde $\delta$ e ε são parâmetros, e quanto tanto menor forem $\delta$ e ε melhor a garantia de privacidade. Caso $\delta = 0$, diz-se que a função provém ε-privacidade diferencial.[3] O que essa equação nos diz é que a probabilidade de um evento em $D_1$ acontecer após passar pela função K é menor ou igual a probabilidade do mesmo evento acontecer em $D_2$ após passar por K vezes um fator $e^{\varepsilon}$. Na prática, isso significa que devido a uma aleatoriedade introduzida por K, dois datasets que diferem de apenas 1 elemento tem aproximadamente a mesma probabilidade de gerar um evento. Assim, a falta ou não deste elemento não influencia nos resultados destes datasets e é plausível negar que o elemento esteve no dataset real antes de ser anonimizado com K e isso gera uma garantia de privacidade. Esse mesmo conceito pode ser estendido de um indivíduo para um grupo com k pessoas multiplicando se a probabilidade no lado direito por $e^{k\varepsilon}$ ao invés de por $e^{\varepsilon}$.
Um exemplo muito simples de um mecanismo de aleatoriedade que provem certo grau de privacidade diferencial seria dado um dataset composto apenas por usuários e suas respostas "Sim" ou "Não" a uma pergunta, o mecanismo para cada entrada joga uma moeda. Caso a moeda caia em cara, o verdadeiro resultado da pergunta é mostrado. Caso coroa, outra moeda é jogada e o resultado é mostrado como "Sim" se esse segundo resultado for cara e "Não" caso for coroa. Isso significa que uma resposta "Sim" no dataset após passar pelo mecanismo ainda pode não ser verdadeiro com uma probabilidade de ¼ e isso significa que mesmo que a pessoa que deu essa resposta seja identificada ela pode negar que o resultado no dataset publicado corresponde à verdade.

## 4.Utilidade de um registro de consultas

### 4.1 Conceito de Utilidade

No campo de pesquisa relacionado a privacidade, a utilidade de um registro de consultas após ser anonimizado é fundamental. A função utilidade varia com o dataset e procura medir o quanto ele é usável para os propósitos desejados. A Utilidade de um processo de anonimização deve ser medida aplicando uma função U tanto ao dataset original Q quanto ao dataset anonimizado Q'.

### 4.2 Utilidade e privacidade

É importante notar que a Utilidade de um dataset público tem uma relação inversamente proporcional ao grau de privacidade deste dataset. Um dataset real provém a maior utilidade e resultados de pesquisas com ele serão os mais fiéis à realidade possíveis, porém com estes datasets é muito fácil de identificar os usuários, havendo assim uma falta de privacidade. Após passar por um mecanismo de anonimização como privacidade diferencial, o dataset teve uma certa aleatoriedade introduzida, o que pode fazer com que resultados de pesquisas sobre o dataset não correspondam completamente aos mesmos resultados sobre o dataset real.
O desafio da área de pesquisa de privacidade é buscar algoritmos que fornecem alto grau de privacidade sem sacrificar muita utilidade e como veremos mais à frente a implementação proposta de privacidade diferencial nos dá bons resultados.

## 5. Formulação do problema

Nesta seção será apresentada a formulação do problema de anonimização de registros de consultas.
**registro de consultas Q**: Nosso registro de consultas Q será um documento textual que contém registros de queries feitas por usuarios. O registro de consultas que usaremos é o da AOL de 2006, que contém informações sobre o nome do usuário, o que ele digitou na query em texto, a data e hora da query, o rank do link que ele clicou e qual link foi clicado.
**Busca na web usando registros de consultas:** Dada uma consulta q e um registro de consultas Q, a tarefa da busca é retornar uma lista ranqueada de documentos ou URLs D que é relevante para a consulta q tirada de um conjunto de documentos ou URLs C pré-indexado. Informações podem ser usadas como cliques de usuários, reformulações da

consulta, tempo passado examinando documentos retornados e documentos clicados em consultas similares pelo mesmo usuário

**A tarefa de anonimizar o registro de consultas:** Dado um registro de consultas de entrada Q, a tarefa é produzir uma versão do registro de consultas na qual os dados identificáveis são removidos e o resto dos dados são propriamente anonimizados de modo a reduzir ao máximo a chance de re-identificação dos usuários. A saída dessa tarefa é um registro de consultas anonimizado Q' com um grau de privacidade garantido.

**Função de privacidade A:** Um registro de consultas anonimizado Q' é gerado aplicando-se uma função A ao registro de consultas original Q, tendo assim Q'=A(Q), Geralmente A têm parâmetros para indicar o nível de privacidade desejado, e no exemplo da privacidade diferencial, o parâmetro em A é ε e assim temos que Q'=Q(ε,A)

**Função de utilidade U:** A função de utilidade U é necessária para avaliação do resultados. Ela precisa ser específica para o domínio e é aplicada tanto em Q quanto Q' para comparação.

**Objetivo:** O objetivo final é ter que $|U(Q) - U(Q')| < \sigma$ onde σ é mantido pequeno. Ao mesmo tempo, o algoritmo de anonimização deve garantir que que o nível de privacidade ε é pequeno o suficiente para manter alta garantia de privacidade.

## 6. Formato do registro anonimizado

Essa seção explica o formato final do registro de consultas após passar pelo algoritmo de anonimização com privacidade diferencial como proposto no artigo [1].

Para melhorar o nível de privacidade dos registros de consultas, algum registro de busca pode ser removido ou modificado. Tais mudanças fazem sentido apenas quando os registros que sobram nos dão informação suficiente para serem úteis.

Primeiramente, são removidos dados altamente identificáveis como IDs e nomes. É preciso que sejam mantidas as consultas, em formato de texto puro. Consultas com frequência muito baixa são removidas pois não contribuem muito e tem alto risco para privacidade do registro. Os links que foram clicados são mantidos pois também são informações chave em um registro de consultas. Porém para mantermos privacidade diferencial é preciso que eles estejam em um formato estatístico e para isso eles são agregados e mostrados como contagens. Finalmente, serão mantidos transições de consultas em um registro. O formato proposto pode ser visto na Figura 2. Cada registro de consultas anonimizado Q' consiste de três partes. A primeira parte contém as buscas e consultas publicadas em Q em formato de texto puro junto com sua frequência. A segunda parte contém dados de quais links foram clicados em quais consultas. A terceira parte de Q' contem as informações de transição de consultas onde cada linha mostra um par de consultas adjacentes e a frequência dessa transição.

Figura 2: formato proposto para registro anonimizado

Part 1: Query Counts

| *Query* | *Counts* |
|---|---|
| weather | 13826 |
| weather channel | 1175 |
| aol weather | 284 |
| las vegas weather | 126 |
| ... | |

Part 2: Click-Through Counts

| *Query* | *Clicked-through URL* | *Counts* |
|---|---|---|
| weather | http://www.weather.com | 4190 |
| weather | http://weather.yahoo.com | 1035 |
| aol weather | http://weather.aol.com | 30 |
| aol weather | http://aolsvc.weather.aol.com | 16 |
| ... | | |

Part 3: Query Transitions Counts

| *Query* | *Next Query* | *Counts* |
|---|---|---|
| weather | aol weather | 44 |
| weather | weather channel | 25 |
| weather | las vegas weather | 9 |
| aol weather | aol yellow pages | 5 |
| ... | | |

## 7. Anonimizando o registro de consultas

Nos importamos tanto com os níveis de privacidade quanto com a utilidade do registro anonimizado.

A implementação proposta pelo artigo [1] busca alcançar ε-privacidade usando-se de um conjunto de consultas para estender o registro original e fazendo uso de uma distribuição de laplace como fator de aleatoriedade.

O processo é composto de 5 passos:

**1. Remoção de dados sensíveis:** Consultas com baixa frequência (menos de 5) são removidas e consultas contendo termos únicos são removidas, isso é feito para filtrar dados muito específicos que possam levar a identificação.

**2. Limitar atividade dos usuários:** Mantemos apenas as primeiras f consultas e os primeiros c URLs clicados por cada usuário em Q. Quanto mais consultas temos de um único usuário, maior a chance dele ser identificado

**3. Expandir o conjunto de consultas:** Não é incomum para donos de registros de consultas incluírem dados de fora dos registros em sua publicação. Esses dados geralmente vem de amostras de outros registros de busca. Na implementação do artigo, esse processo é simulado.

**4. Introduzir função de aleatoriedade:** Usaremos aqui uma função de Ruído de Laplace. Definiremos a contagem de consultas como a contagem original mais o ruído de laplace correspondente. Consultas apenas serão publicadas caso sua contagem com ruído esteja acima de um limite arbitrário K

**5. Gerar estatísticas:** As contagens de consultas e contagem de links clicados são publicadas, com seus valores já alterados pelo ruído de laplace

**6. Gerar transições de consultas:** informações de consultas sequenciais são computadas e publicadas. Consultas sequências são publicadas apenas se ambas seriam publicadas após aplicação do ruído.

## 8. Avaliação e resultados

**8.1 definição da função de avaliação:** O resultado de um processo de anonimização de um registro de consultas deve ser avaliado em termos de sua Utilidade. No contexto de recuperação da informação, a função utilidade de um registro de consultas pode ser definida em dois passos: o primeiro é utilizar o registro para alguma recuperação de documentos usando métodos de recuperação como o método probabilístico junto com algum sistema de feedback usando consultas e o segundo passo é avaliar essa recuperação. Para avaliar a recuperação podemos usar métodos como o Mean Average Precision. Então, temos que nossa função de utilidade U é definida como U(Q)=E(R(Q)), onde Q é um registro de consultas, R é uma função de recuperação de documentos e E é uma função de avaliação da função R. O objetivo de um algoritmo de anonimização será minimizar |U(Q)-U(Q')| onde Q é o registro original e Q' é o registro anonimizado enquanto ainda mantém alto grau de privacidade.

**8.2 Resultados:** Usando o MAP como método de avaliação de recuperação e um método de recuperação por feedback implícito de consultas, os artigo conseguiu bons resultados, tendo o registro de consultas anonimizado alcançado um MAP apenas 0,0014 menor que o do registro original, diferença essa que é aceitável. Como esperado, aumento nos parâmetros de privacidade como o nível do ruído de laplace aplicado aumentam essa diferença.

## 9.2 Experimentação

De acordo com a proposta, foi planejada uma implementação própria porém esta não pode ser concluída a tempo. Mudanças planejadas eram: Mudança no método de recuperação, mudança na avaliação, mecanismo de privacidade diferencial que não fosse ruído de Laplace, e o uso de um outro dataset inteiramente.

O uso de outro dataset não teria sido possível pois registros de consultas ainda são escassos, especialmente não-anonimizados. Assim só foi encontrado o da AOL 2006 que é o usado pelo artigo.

## 8. Conclusão

O problema de anonimização de registros de consultas é um obstáculo a pesquisas na área de recuperação da informação pois dificulta a publicação de datasets que poderiam ser muito úteis. Assim, a solução proposta usando privacidade diferencial é um ótimo caminho para melhorar o acesso a dados. O conceito de privacidade diferencial em si também é extremamente útil e pode ser usado em vários datasets com funções de diferentes formas e complexidades.

Infelizmente uma implementação própria do algoritmo de privacidade diferencial no registro de consultas da AOL usando algoritmos de avaliação e busca aprendidos na sala foi planejado mas não pode ser completado por questão

de tempo. Além disso, a dificuldade de encontrar outros registros de consultas não-anonimizados preveniu a possibilidade de testes em outro dataset.

**Referencias**


1. Zhang, S. Yang, H. Singh, L. “Anonymizing Query Logs by Differential Privacy”. SIGIR 2016
2. “A Face is Exposed for AOL Searcher No. 4417749”, https://www.nytimes.com/2006/08/09/technology/09aol.html
3. Dwork, C. “Differential Privacy”. 2006