Protegendo a privacidade ao divulgar o anonimato das informações e sua aplicação por meio de generalização e supressão

Pierangela Samarati
Laboratório de Informática
SRI Internacional
Menlo Park CA EUA
samaraticslsricom

Latanya Sweeney
Laboratório de Informática
Instituto de Tecnologia de Massachusetts
Cambridge MA EUA
sweeneyaimitedu

Resumo

A sociedade globalmente conectada de hoje coloca grande demanda na disseminação e compartilhamento de dados pessoais específicos Situações em que informações estatísticas agregadas já foram a norma de relatórios agora dependem fortemente da transferência de transações microscopicamente detalhadas e informações de encontro Isso acontece em um momento em que cada vez mais informações históricas informações públicas também estão disponíveis eletronicamente Quando esses dados são vinculados, eles fornecem uma sombra eletrônica de uma pessoa ou organização que é tão identificável e pessoal quanto uma impressão digital, mesmo quando as fontes das informações não contêm identificadores explícitos, como nome e número de telefone A fim de proteger o anonimato dos indivíduos a quem os dados divulgados se referem, os detentores de dados geralmente removem ou criptografam identificadores explícitos, como nomes, endereços e números de telefone.

Neste artigo, abordamos o problema de liberar dados pessoais e, ao mesmo tempo, salvaguardar o anonimato dos indivíduos a quem os dados se referem. A abordagem é baseada na definição de k anonimato. o conteúdo mapeia a informação de forma ambígua para pelo menos k entidades Ilustramos como a canonímia pode ser fornecida usando técnicas de generalização e supressão

Ilustramos possíveis políticas de preferência para escolher entre diferentes generalizações mínimas Finalmente, apresentamos um algoritmo e resultados experimentais quando uma implementação do algoritmo foi usada para produzir lançamentos de informações médicas reais Também relatamos a qualidade dos dados divulgados medindo a precisão e completude dos resultados para diferentes valores de k

O trabalho de Pierangela Samarati foi financiado em parte pelo DARPARome Laboratory sob a bolsa FC e pela National Science Foundation sob a bolsa ECS pela Medical Informatics Training Grant LM pela National Library of Medicine

Universita di Milano

Introdução

Na era da Internet e do poder da computação barata, a sociedade desenvolveu um apetite insaciável por informações, todos os tipos de informações para muitos usos novos e empolgantes. A maioria das ações da vida diária são gravadas em algum computador em algum lugar.
Muitas pessoas podem não se importar que a mercearia local acompanhe quais itens compram, mas as informações compartilhadas podem ser bastante sensíveis ou prejudiciais para indivíduos e organizações A divulgação inadequada de informações médicas informações financeiras ou questões de segurança nacional pode ter consequências alarmantes e muitos abusos foram cometidos citado O objetivo é liberar informações livremente, mas fazê-lo de forma que a identidade de qualquer indivíduo contido nos dados não possa ser reconhecida. Dessa forma, as informações podem ser compartilhadas livremente e usadas para muitos novos propósitos

Surpreendentemente, permanece uma crença incorreta comum de que, se os dados parecem anônimos, eles são anônimos. Os detentores de dados, incluindo agências governamentais, geralmente removem todos os identificadores explícitos, como nome, endereço e número de telefone dos dados, para que outras informações nos dados possam ser compartilhadas incorretamente, acreditando que as identidades dos indivíduos não podem ser determinadas.

As informações divulgadas geralmente contêm outros dados, como sexo, data de nascimento e CEP, que combinados podem ser vinculados a informações disponíveis publicamente para reidentificar indivíduos A maioria dos municípios vende registros populacionais que incluem as identidades dos indivíduos junto com exemplos básicos de demonstração incluem dados do censo local listas de eleitores diretórios de cidades e informações de agências de veículos motorizados avaliadores de impostos e agências imobiliárias Por exemplo, uma versão eletrônica da lista de eleitores de uma cidade foi comprada para vinte dólares e usado para mostrar a facilidade de reidentificação de registros médicos Além dos nomes e endereços, a lista de eleitores incluía as datas de nascimento e gêneros dos eleitores. Destes, as datas de nascimento e sexo em relação à data nascimento e sexo em relação à data de nascimento e código postal de dígito e eram identificáveis apenas com o código

Esses resultados revelam como combinações de identificação única de atributos demográficos básicos, como CEP, data de nascimento, etnia, gênero e estado civil, podem ser

Para ilustrar este problema, a Figura exemplifica uma tabela de dados médicos liberados sem identificação pela supressão de nomes e números de CPF, SSNs, de modo a não divulgar as identidades dos indivíduos a quem os dados se referem.
Estado Civilg também pode aparecer em alguma tabela externa juntamente com a identidade individual e assim permitir que ela seja rastreada Conforme ilustrado na Figura ZIP Data de Nascimento e Sexo podem ser vinculados ao
Lista de eleitores para revelar o nome, endereço e cidade Da mesma forma, a etnia e o estado civil podem ser vinculados a outros registros populacionais disponíveis publicamente. Na tabela de dados médicos da Figura, há apenas uma mulher nascida e morando na área. a lista real de eleitores mais do que os eleitores podem ser identificados exclusivamente usando apenas estes atributos Esta Carlson

Main Street Cambridge e, portanto, revela que ela relatou falta de ar Observe que as informações médicas não são consideradas publicamente associadas aos indivíduos e a proteção desejada é liberar as informações médicas de forma que as identidades dos indivíduos não possam ser determinadas. características divulgadas para Sue J Carlson levam a determinar quais dados médicos entre os divulgados são dela Enquanto este exemplo demonstrou uma correspondência exata em alguns casos, as informações divulgadas podem ser vinculadas a um conjunto restritivo de indivíduos aos quais as informações divulgadas podem se referir

Várias técnicas de proteção foram desenvolvidas com relação a bancos de dados estatísticos, como embaralhar e trocar valores e adicionar ruído aos dados de forma a manter uma propriedade estatística geral do resultado

No entanto, muitos novos usos de dados, incluindo análise de custo de mineração de dados e pesquisa retrospectiva, geralmente precisam de informações precisas dentro da própria tupla.

SSN Nome	Etnia	Data de Nascimento	Sexo	feminino feminino		fecho eclair	Estado civil	Problema divorciado
		asiático		masculino			hipertensão divorciado	obesidade
		asiático					dor no peito casado com	obesidade
		asiático					hipertensão casado com	falta de ar
		asiático					falta de ar casado com	obesidade dor
		preto					com obesidade viúva	solteira falta de ar
		preto						
		preto						
		preto						
		branco						
		branco						
		branco		masculino	masculino		feminino feminino masculino	masculino feminino

casado com
casado com
casado com
não pode solteiro

Nome	Endereço	Cidade	CEP	DOB	Sexo		Festa	
Sue J Carlson	Main St	Cambridge				fêmea	democrata	

Figura Reidentificando dados anônimos por meio de links para dados externos

controlam mantendo a integridade dos valores dentro de cada tupla nomeadamente Datay nos Estados Unidos e MuArgus na Europa No entanto não foram fornecidas bases formais ou abstracções para as técnicas utilizadas por ambos Outras aproximações feitas pelos sistemas podem sofrer de inconvenientes tais como a generalização dos dados mais do que é necessário como

ou não fornecer proteção adequada, como

Neste artigo, fornecemos uma base formal para o problema do anonimato contra vinculação e para a aplicação de generalização e supressão para sua solução. dados com relação à inferência por ligação Mostramos como o anonimato pode ser garantido em lançamentos de informações generalizando e/ou suprimindo parte dos dados a serem divulgados Neste quadro, introduzimos os conceitos de tabela generalizada e de generalização mínima Intuitivamente uma generalização é mínima se os dados são não generalizado mais do que o necessário para fornecer kanonymity Além disso, a definição de generalização preferida permite ao usuário selecionar entre possíveis generalizações mínimas aquelas que satisfazem condições particulares, como favorecer certos atributos no processo de generalização Apresentamos um algoritmo para calcular uma generalização mínima preferida de um dado tabela n Finalmente, discutimos alguns resultados experimentais derivados da aplicação de nossa abordagem a um banco de dados médico contendo informações sobre pacientes

O problema que consideramos difere do controle de acesso tradicional e do banco de dados estatístico

problemas Os sistemas de controle de acesso abordam o problema de controlar o acesso específico aos dados com relação às regras que determinam se um dado pode ou não ser liberado. decisão pode ser tomada, mas sim o fato de que os dados se referem a uma entidade específica As técnicas de banco de dados estatísticos abordam o problema de produzir dados tabulares que representam um resumo das informações a serem consultadas A proteção é aplicada em tal estrutura, garantindo que não seja possível para os usuários inferirem dados individuais originais a partir do resumo produzido Em nossa abordagem, em vez disso, permitimos a liberação de dados pessoais específicos generalizados nos quais os usuários podem produzir resumos de acordo com suas necessidades. usuários Essa flexibilidade e disponibilidade tem como desvantagem do ponto de vista do usuário final nível de granularidade grosseira dos dados

Este novo tipo de desclassificação e liberação de informações parece ser cada vez mais exigido nas aplicações emergentes de hoje

O restante deste artigo está organizado da seguinte forma. Na Seção, apresentamos as suposições básicas e

denições Na Seção discutimos a generalização para fornecer anonimato e na Seção continuamos a discussão para incluir a supressão Na Seção as políticas básicas de preferência para escolher entre diferentes generalizações mínimas são ilustradas Na Seção discutimos uma implementação algorítmica de nossa abordagem
A seção relata alguns resultados experimentais A seção conclui o artigo

Suposições e negações preliminares

Consideramos a tabela de detentores de dados como uma organização individual de tabela privada e assim por diante Da tabela privada como uma liberação anônima da onde cada tupla se refere a uma entidade diferente, o titular dos dados constrói uma tabela que é Pro- uma questão de simplicidade iremos subsequentemente referir-nos à privacidade e reidentificação de indivíduos em casos igualmente aplicáveis a outras entidades Assumimos que todos os identificadores explícitos, por exemplo, nomes SSNs e endereços são encriptados ou suprimidos e, portanto, ignoramo-los no restante deste paper Tomando emprestada a terminologia de nós chamamos a combinação de atributos ligação a possível associação com registros externos Os quais quasiidentiers devem portanto ser protegidos Eles são definidos da seguinte forma

Denition Quasiidentier Seja TA An uma tabela Um quasiidentier de T é um conjunto de at tributos fAi Aj g fA Ang cuja liberação deve ser controlada por led

Dada uma tabela TA An um subconjunto de atributos fAi Aj g fA Ang e uma tupla t T tAi Aj denota a sequência dos valores de Ai Aj em t T Ai Aj denota a projeção mantendo tuplas duplicadas de atributos Ai Aj em T Também denota o conjunto de quase-identificadores associados a T e jT j denota cardinalidade que é o número de tuplas em T

Nosso objetivo é permitir a divulgação das informações da tabela garantindo o anonimato dos indivíduos A restrição de anonimato exige que as informações divulgadas se relacionem indistintamente com pelo menos um determinado número k de indivíduos, onde k é normalmente definido pelo detentor dos dados, conforme declarado pelo seguinte requisito

Definição Requisito de kanonimato Cada liberação de dados deve ser tal que cada combinação de valores de quase-identificadores possa ser combinada indistintamente com pelo menos k indivíduos

A adesão ao requisito de anonimato requer saber quantos indivíduos cada tupla liberada corresponde Isso pode ser feito vinculando explicitamente os dados liberados com os dados disponíveis externamente Esta é obviamente uma tarefa impossível para o detentor dos dados Embora possamos supor que o detentor dos dados saiba quais atributos podem aparecem em tabelas externas e, portanto, o que constitui os quase-identificadores os valores específicos dos dados no conhecimento externo não podem ser assumidos A chave para satisfazer o requisito de kanonymity, portanto, é traduzir o requisito em termos dos próprios dados liberados Para fazer isso, exigimos a seguinte suposição segurar

Suposições que devem ser liberadas e que estão disponíveis externamente em combinação, ou seja, aparecendo juntas em uma tabela externa ou em possíveis junções entre tabelas externas para um destinatário de dados, são definidas em um quase-identificador associado a Todos os atributos na tabela

Embora esta não seja uma suposição trivial a sua aplicação é possível O titular dos dados estima quais os atributos que podem ser usados para ligar com conhecimento externo isto obviamente forma a base para um quasi-identificador Embora a expectativa desse conhecimento seja um tanto razoável para dados disponíveis publicamente, reconhecemos que existem muitas fontes de informações semipúblicas e privadas, como registros de farmácias longitudinais

Uma relação universal combinando tabelas externas pode ser imaginada

estuda registros financeiros responde a pesquisas listas ocupacionais e listas de membros para considerar a priori todas as possibilidades de ligação Suponha que a escolha de atributos para um quase-identificador esteja incorreta, ou seja, o detentor dos dados julga incorretamente quais atributos são sensíveis para vinculação Nesse caso, os dados liberados podem ser menos anônimos do que o necessário e, como resultado, os indivíduos podem ser identificados mais facilmente Sweeney examina esse risco e mostra que não pode ser perfeitamente resolvido pelo titular dos dados uma vez que o titular dos dados nem sempre pode saber o que cada destinatário dos dados sabe propõe soluções para esse problema, mas nos concentramos no quase-identificadores foram reconhecidos

Introduzimos a definição de canonimato para uma tabela da seguinte forma

A definição seja os quase-identificadores associados a T diz-se que satisfaz a canonímia i para cada quase-identificador QI T cada sequência de valores em T QI aparece pelo menos com k ocorrências em T QI

sob suposição e sob a hipótese de que a tabela armazenada privadamente contém no máximo uma tupla para cada identidade a ser protegida, ou seja, a quem um quase-identificador se refere, a canonímia de uma tabela liberada representa uma condição suciente para a satisfação do requisito de canonimato Em outras palavras, uma tabela que satisfaz a Definição para um determinado k satisfaz o requisito de kanonimato para tal k Considere um quase-identificador QI se a definição for satisfeita cada tupla em QI tem pelo menos a aparência de Dado que a tabela privada contém no máximo uma tupla para cada indivíduo, o conjunto de que correspondem a esses valores Além disso, como todos os atributos disponíveis em combinação externa estão incluídos no QI, nenhum atributo adicional pode ser associado ao QI para reduzir a cardinalidade de tal conjunto Observe também que qualquer subconjunto dos atributos no QI se referirá a kk indivíduos Para ilustrar, considere a situação exemplificada na Figura, mas suponha que os dados divulgados continham duas ocorrências de th e sequência viúva branca Então existirão pelo menos dois indivíduos correspondentes a tais ocorrências na lista de eleitores ou na tabela que combina a lista de eleitores com todas as outras tabelas externas e não será possível para o destinatário dos dados determinar qual dos dois pertence a qual dos dois indivíduos Desde que a canonimato foi fornecido na liberação cada prontuário poderia pertencer indistintamente a pelo menos dois indivíduos

Dada a suposição e denições acima e dada uma tabela privada que satisfaça o o problema de produzir uma versão canonimato para ser lançado nos concentramos em

Generalizando dados

Nossa primeira abordagem para fornecer kanonymity é baseada na definição e uso de relações de generalização entre domínios e entre valores que os atributos podem assumir

Relações de generalização

Em um sistema de banco de dados relacional clássico, os domínios são usados para descrever o conjunto de valores que os atributos assumem Por exemplo, pode haver um domínio de código postal um domínio de número e um domínio de string Nós estendemos esta noção de domínio para facilitar a descrição de como generalizar os valores de um atributo Na base de dados original onde cada valor é tão específico quanto possível cada atributo está no domínio básico
Por exemplo está no ZIP do domínio do ZIP básico Para obter o kanonimato podemos tornar o CEP menos informativo Fazemos isso dizendo que existe um domínio mais geral e menos específico que pode ser usado para descrever os CEPs Z em que o último dígito tem foi substituído por um Há também um mapeamento de Z a Z, como no Esse mapeamento entre domínios é feito por meio de uma generalização as seguintes condições Cada domínio Dois que é necessário para o generalizado tipo representa uma ordem parcial, satisfaz

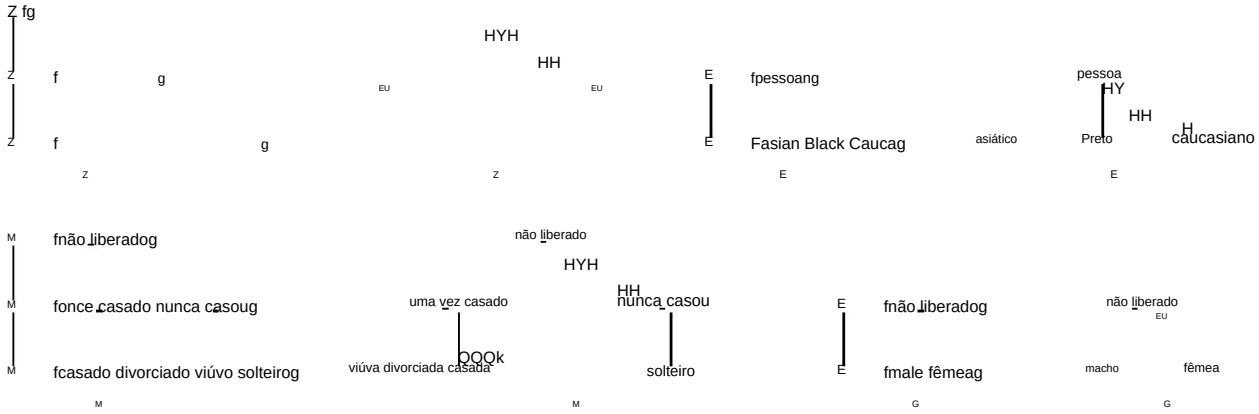

Figura Exemplos de hierarquias de generalização de domínio e valor

elementos maximais de são singleton A denição desta generalização implica a existência de cada domínio D de uma hierarquia que chamamos de hierarquia de generalização de domínio D Uma vez que valores generalizados podem ser usados no lugar de valores mais específicos, é importante que todos os domínios em uma hierarquia sejam compatíveis A compatibilidade pode ser assegurada usando o mesmo formulário de representação de armazenamento para todos os domínios em uma hierarquia de generalização Uma ordem parcial de relação de generalização de valor também é definida que associa a cada valor vi em um domínio Di um valor único no domínio Tal relação implica a existência para cada domínio D de uma hierarquia de generalização de valores

D

Exemplo A Figura il ilustra um exemplo de domínio e hierarquias de generalização de valor para domínio Z representando códigos postais da área de Cambridge MA E representando etnias M representando estado civil e G representando gênero

No restante deste artigo iremos nos referir freqüentemente a um domínio ou hierarquia de generalização de valor em termos do grafo que representa todas e somente as relações de generalização direta entre os elementos nele, isto é, as relações de generalização implícitas não aparecem como arcos no gráfico Nós usaremos o termo hierarquia de forma intercambiável para denotar um conjunto parcialmente ordenado ou o grafo que representa o conjunto e todas as relações de generalização direta entre seus elementos Iremos nos referir explicitamente ao conjunto ordenado ou ao grafo quando não estiver claro no contexto

Além disso, como estaremos lidando com conjuntos de atributos, é útil visualizar o relacionamento de generalização e as hierarquias em termos de tuplas compostas de elementos de ou de seus valores Dada uma tupla DT hD n definimos a hierarquia de generalização de domínio de DT como

Dni tal que Di

eu

A D Dn assumindo que o produto cartesiano é ordenado pela imposição de coordenadas ordem DT sábia define uma rede cujo elemento mínimo é DT A hierarquia de generalização de uma tupla de domínio DT define as diferentes maneiras pelas quais DT pode ser generalizado Em particular, cada caminho de DT para o único elemento máximo de no grafo que descreve um possível caminho alternativo que podem ser seguidos no processo de generalização Nos referimos ao conjunto de nós em cada Um desses caminhos juntamente com as relações de generalização como uma

A Figura DT ilustra as hierarquias de hierarquia de generalização de domínio de E e Z conforme ilustrado na Figura

A motivação por trás da condição é garantir que todos os valores em cada domínio possam ser eventualmente generalizados para um único valor

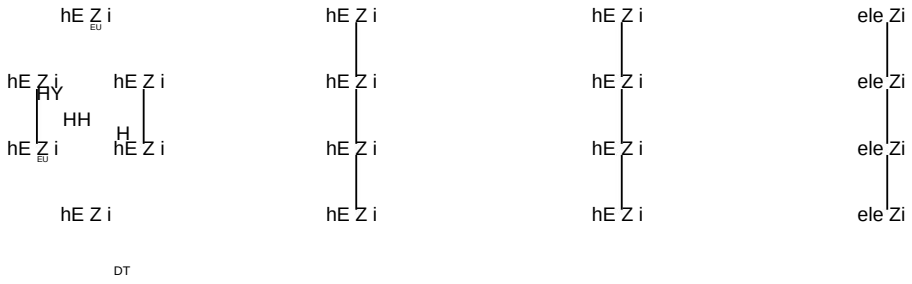

Figura Hierarquia de generalização de domínio DT e estratégias para DT

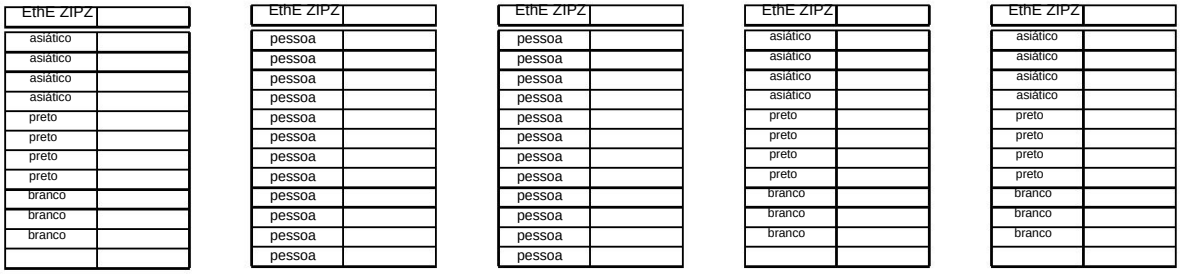

Figura Exemplos de tabelas generalizadas para

Tabela generalizada e generalização mínima

Dada uma tabela privada a nossa primeira abordagem para fornecer o anonimato consiste em generalizar os valores armazenados na tabela Intuitivamente os valores dos atributos armazenados na tabela privada podem ser substituídos após o lançamento com valores generalizados Uma vez que vários valores podem ser mapeados para um único valor generalizado a generalização pode diminuir o número de tuplas distintas, possivelmente aumentando o tamanho dos clusters contendo tuplas com os mesmos valores Realizamos a generalização no nível do atributo Generalizar um atributo significa substituir seus valores por valores correspondentes de um domínio mais geral A generalização no nível do atributo garante que todos os valores de um o atributo pertence ao mesmo domínio No entanto, como resultado do processo de generalização, o domínio de um atributo pode mudar No seguinte domAi T denota o domínio do atributo

Ai na tabela T Di domAi denota o domínio associado ao atributo Ai na tabela privada

negação Tabela Generalizada Let TiA mesmo An e Tj A An são duas tabelas denidas no conjunto de atributos Tj é dito ser uma generalização de Ti escrito Ti Tj i

jTi j jTj j

z n domAz Ti domAz Tj

É possível definir um mapeamento bijetivo entre Ti e Tj que associa cada tupla ti e tj tal que ti Az tj Az

negação afirma que uma tabela Tj é uma generalização de uma tabela Ti definida nos mesmos atributos Ti e Tj têm o mesmo número de tuplas o domínio de cada atributo em Tj é igual ou a

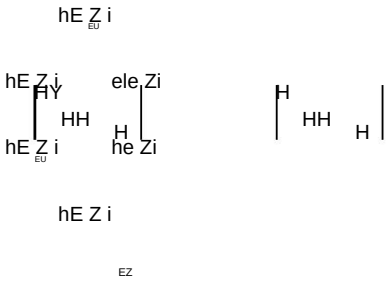

Figura Hierarquia e rede correspondente em vetores de distância

generalização do domínio do atributo em Ti e cada tupla ti em Ti tem uma tupla correspondente tj em Tj e vice-versa tal que o valor para cada atributo em tj é igual ou uma generalização do valor do atributo correspondente em ti

Exemplo ilustrado na Figura e as hierarquias de generalização de domínio e valor para E e Z ilustradas na Figura As quatro tabelas Considere a tabela restantes na Figura são todas tabelas generalizadas possíveis para, mas a mais alta generaliza cada tupla para hpessoa i Para clareza do exemplo cada tabela informa o domínio para cada atributo na tabela Com relação ao kanonymity kanonymity para kk e satisfaz

satisfaz o anonimato para k Satisfaz o anonimato para

satisfaz o anonimato para k

Dada uma tabela, existem diferentes generalizações possíveis Nem todas as generalizações, entretanto, podem ser consideradas igualmente satisfatórias Por exemplo, a generalização trivial trazendo cada atributo para o nível mais alto possível de generalização, reduzindo assim todas as tuplas em T para a mesma lista de valores fornece kanonimato ao preço de um generalização forte dos dados Tal generalização extrema não é necessária se existir uma tabela mais específica ie contendo valores mais específicos que satisfaça a canonímia Este conceitoécapturado pela definição de generalização kminimal

negação Vetor de distância Let TiA An e Tj A An são duas tabelas tais que Ti Tj

O vetor de distância de Tj de Ti é o vetor ij dn onde cada dz é o comprimento do caminho único entre D domAz Ti e domAz Tj na hierarquia de generalização de domínio d D

Exemplo Considere a tabela e suas tabelas generalizadas ilustradas na Figura Os vetores de distância e entre suas diferentes generalizações são os vetores que aparecem como um subscrito de cada tabela

Dados dois vetores de distância d dn e d d n eu morro di por tudo que eu n e Uma hierarquia de generalização para uma tupla de domínio pode ser vista como uma rede hierárquica nos vetores de distância correspondentes. Por exemplo, a Figura ilustra a rede que representa a relação entre os vetores de distância correspondentes à possível generalização de hE Zi

Podemos agora introduzir a denição de generalização kminimal

Denição generalização kminimal Sejam Ti e Tj duas tabelas tais que Ti Tj Tj é dito ser uma generalização kminimal de Ti i

Tj satisfaz o anonimato

Ethn	DOB	Sexo	fecho eclair	Status
asiático		feminino		
asiático		feminino		
asiático		masculino		
asiático		masculino		
preto		masculino		
preto		feminino		
preto		feminino		
preto		masculino		
branco		masculino		
branco		feminino		
branco				

divorciados divorciados casados casados casados casados solteiros solteiros viúvos

Ethn	DOB	Sexo		fecho eclair	Status
asiático		não rel			não rel
asiático		não rel			não rel
asiático		não rel			não rel
asiático		não rel			não rel
preto		não rel			não rel
preto		não rel			não rel
preto		não rel			não rel
preto		não rel			não rel
branco		não rel			não rel
branco		não rel			não rel
branco		não rel	não rel		não rel não rel

Ethn	DOB	Sexo		fecho eclair	Status
pessoa		feminino			estive
pessoa		feminino			estive
pessoa		masculino			estive
pessoa		masculino			estive
pessoa		masculino			estive
pessoa		feminino			
pessoa		feminino			
pessoa		masculino			
pessoa		masculino			nunca
pessoa		feminino			nunca
pessoa					estive

Figura Um exemplo de tabela e suas generalizações mínimas

Tz Ti Tz Tz satisfaz kanonymity e DViz DVij

Intuitivamente uma generalização Tj é mínima i não existe outra generalização Tz satisfazendo k anonimato que é dominado por Tj na hierarquia de generalização de domínio de hD Dni ou equivalentemente na rede correspondente de vetores de distância Se este fosse o caso Tj seria uma generalização para Tz Observe também que uma tabela pode ser uma generalização mínima de si mesma se a tabela já tiver alcançado o anonimato

Exemplo para Considere a tabela e suas tabelas generalizadas ilustradas na Figura Assume QI Eth ZIP ser um quasiidentier É fácil ver que para k existem duas generalizações kminimal que são e A tabela que satisfaz os requisitos de anonimato não é mínima, pois é uma generalização de não pode ser mínima Assim também é generalização de ambos e Existem também apenas duas tabelas generalizadas kminimal para k que são e

Note que desde que o kanonymity requer a existência de koccurrences para cada seqüência de valores apenas para quasiidentiers para cada generalização mínima Tj DVij dz para todos os atributos Az que não pertencem a nenhum quasiidentier

Suprimindo dados

Na Seção, discutimos como, dada uma tabela privada, uma tabela generalizada pode ser produzida, liberando uma versão mais geral dos dados e satisfazendo uma restrição de kanonimato. A generalização tem a vantagem de preservar a veracidade dos dados em cada célula individual da tabela, embora de kanonymity que é supressão Suprimir significa remover dados da tabela para que eles não sejam liberados e como uma técnica de controle de divulgação não é novo Aplicamos supressão no nível da tupla, ou seja, uma tupla pode ser suprimida apenas em sua totalidade A supressão é usada para moderar o processo de generalização quando um número limitado de outliers é

Ethn DOB		Sexo	fecho eclair	Status
asiático		feminino		divorciados
asiático		feminino		divorciados
asiático		masculino		casados
asiático		masculino		casados
preto		masculino		casados
preto		feminino		casados
preto		feminino		casados
preto				
Branco		macho		solteiro
Branco		macho		solteiro

Ethn DOB Sexo			fecho eclair	Status
asiático		fêmea		divorciado
asiático		fêmea		divorciado
asiático		macho		casado
asiático		macho		casado
preto		macho		casado
preto		fêmea		casado
preto		fêmea		casado
preto		macho		casado
branco		macho		solteiro
branco				solteiro

Figura Um exemplo de tabela e sua generalização mínima

EthE ZIPZ	
asiático	
asiático	
asiático	
asiático	
preto	
preto	
preto	
branco	

EthE ZIPZ	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	

EthE ZIPZ	
asiático	
asiático	
asiático	
asiático	
preto	
preto	
preto	
branco	

EthE ZIPZ	
asiático	
asiático	
asiático	
asiático	
preto	
preto	
preto	
branco	

EthE ZIPZ	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	

Figura Exemplos de tabelas generalizadas para

tuplas com menos que k ocorrências forçariam uma grande quantidade de generalização Para esclarecer considere a tabela ilustrada na Figura cuja projeção no quase-identificador considerado é ilustrada na Figura e suponha que seja fornecida a canonimato com k Atributo Data de Nascimento tem uma data de domínio com as seguintes generalizações da data específica mmaa para o mês mmaa para o ano yy para um intervalo de ano, por exemplo Nascimento As duas possíveis generalizações mínimas são conforme ilustrado na Figura passo de generalização no Código Postal um passo de generalização no Estado Civil e mais um passo no Código Postal de Sexo e Estado Civil ou alternativamente na Etnia e Data de

Na prática, em ambos os casos, quase todos os atributos devem ser generalizados. Pode ser facilmente visto ao mesmo tempo que, se esta última tupla não estivesse presente, a canonímia poderia ter sido simplesmente alcançada por duas etapas de generalização no atributo Data de Nascimento, conforme ilustrado na Figura Suprimindo a tupla, neste caso, permitiria a aplicação de menos generalização

Ao ilustrar como a supressão interage com a generalização para fornecer kanonymity, começamos por re declarando a definição de tabela generalizada da seguinte forma

negação Tabela generalizada com supressão Let TiA An e Tj A An são duas tabelas definidas no mesmo conjunto de atributos Tj é dito ser uma generalização de Ti escrito Ti Tj i

jTj j jTi j

z n domAz Ti domAz Tj

É possível denir um mapeamento injetivo entre Ti e Tj que associa as tuplas ti Ti e tj Tj tais que ti Az tj Az

Repare que embora a generalização possa parecer mudar o formato dos dados a compatibilidade pode ser assegurada usando a mesma forma de representação Por exemplo o mês pode ser representado sempre como um dia específico Esteérealmente o truque que usamos na nossa aplicação de generalização

EthE	ZIPZ
asiático	
asiático	
asiático	
asiático	
preto	
preto	
preto	
branco	

EthE	ZIPZ
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	

EthE	ZIPZ
asiático	
asiático	
asiático	
asiático	
Preto	
Preto	

EthE	ZIPZ
asiático	
asiático	
asiático	
asiático	
preto	
preto	
preto	

EthE	ZIPZ
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	
pessoa	

Figura Exemplos de tabelas generalizadas para

A denição acima difere da denição uma vez que permite que as tuplas que aparecem em Ti não tenham nenhuma tupla generalizada correspondente em Tj Intuitivamente, as tuplas em Ti que não têm nenhum correspondente em Tj são tuplas que foram suprimidas

A denição permite qualquer quantidade de supressão em uma tabela generalizada Obviamente não estamos interessados em tabelas que suprimam mais tuplas do que o necessário para alcançar a canonímia em um dado nível de generalização
Isso é capturado pela seguinte denição

negação Supressão mínima necessária Seja Ti uma tabela e Tj uma generalização de Ti que satisfaça a canonímia Tj é dito para impor a supressão mínima necessária i jTzj e Tz satisfaz a canonimato Tz tal que Ti Tz iz
ij jTj j

Considere a tabela Exemplo e suas generalizações ilustradas na Figura As tuplas escritas em negrito e marcadas com linhas duplas em cada tabela são as tuplas que devem ser suprimidas para atingir k anonimato de Supressão de um subconjunto delas não atingiria o anonimato necessário Supressão de qualquer superconjunto seria desnecessário não satisfazendo a supressão mínima exigida

Permitir que as tuplas sejam suprimidas normalmente resulta em mais tabelas por nível de generalização É trivial provar, no entanto, que para cada vetor de distância possível a tabela generalizada que satisfaz uma restrição de kanonimato aplicando supressão mínima é única Esta tabela é obtida aplicando primeiro a generalização descrita por o vetor de distância e, em seguida, removendo todos l e apenas as tuplas que aparecem com menos de k

ocorrências

No restante deste artigo, assumimos a condição declarada na Definição a ser satisfeita, ou seja, todas as generalizações que consideramos impõem a supressão mínima necessária. pretenda a generalização única para aquele vetor de distância que satisfaça a restrição de kanonimato aplicando a supressão mínima necessária Para ilustrar, considere a tabela na Figura em relação ao kanonimato com k nos referiríamos às suas generalizações conforme ilustrado na Figura

Observe que, para fins de clareza, deixamos uma linha vazia para corresponder a cada tupla removida

A generalização e a supressão são duas abordagens diferentes para obter de uma dada tabela uma tabela que satisfaça o canonimatoÉ trivial notar que as duas abordagens produzem os melhores resultados quando aplicadas em conjunto é insatisfatório ver Figura

A supressão sozinha no outro lado exigiria a supressão de todas as tuplas na tabela A aplicação conjunta das duas técnicas permite, em vez disso, a liberação de uma tabela como a da Figura A questão é, portanto, se é melhor generalizar ao custo de menos precisão em os dados ou suprimir ao custo da integridade A partir de observações de aplicativos da vida real e requisitos especificados declarando o número máximo de tuplas suprimidas que é considerado aceitável Dentro desse limite aceitável, a supressão é considerada preferível à generalização, em outras palavras, é melhor

assumimos o seguinte Consideramos um limite de supressão aceitável Como

suprimir mais tuplas do que impor mais generalização A razão para istoéque a supressão afecta as tuplas individuais enquanto a generalização modifica todos os valores associados a um atributo afectando assim todas as tuplas na tabela Tabelas que reforçam a supressão para além são considerados inaceitáveis

Dadas essas suposições, podemos agora reafirmar a definição de generalização kminimal levando em consideração a supressão

negação generalização kminimal com supressão Sejam Ti e Tj duas tabelas tais que seja o limite
Ti Tj e especificado de supressão aceitável Tj é considerado um kminimal
generalização de uma tabela Ti i

Tj satisfaz o anonimato jTi

j jTj j

Tz Ti Tz Tz satisfaz as condições e e DViz DVij

Intuitivamente a generalização Tj é kminimal i satisfaz a canonímia não impõe mais supressão do que o permitido e não existe outra generalização que satisfaça estas condições com um vetor de distância menor que o de Tj nem existe outra tabela com o mesmo nível de generalização satisfazendo essas condições com menos supressão

Exemplo ilustrado na Figura a supressão é que a kanonimato com k é necessária As generalizações possíveis, Com Figura mas a mais alta, colapsando cada tupla para hpessoa i são ilustradas na Figura Dependendo do limite de supressão aceitável, as seguintes generalizações são consideradas mínimas

ou suprimir mais tupla do que é permitido não é
mínimo por causa de
e suprime mais tuplas do que é permitido GT não é mínimo não é
por causa de e mínimo por causa de GT e
e e não é mínimo por causa do GT GT e não são
mínimo por causa de

Preferências

Fica claro na Seção que pode haver mais de uma generalização mínima para um determinado limite de supressão de tabela e restrição de kanonimato. é aplicado No entanto, podem existir várias soluções que satisfaçam esta condição Qual das soluções é preferida depende de medidas subjetivas e preferências do destinatário dos dados Por exemplo, dependendo do uso dos dados liberados, pode ser preferível generalizar alguns atributos de outros Nós esboçamos aqui algumas políticas de preferência simples que podem ser aplicadas na escolha de uma generalização mínima preferida Para fazer isso, primeiro introduzimos duas medidas de distância definidas entre tabelas distância absoluta e distância relativa Let TiA

An ser uma tabela e Tj A An ser uma de suas generalizações com vetor de distância é a soma das distâncias para cada atributo Formalmente di A distância relativa de Tj de Ti escala distância $d_{eu\,j}$ d_{ni} $d_{eu\,j}$ $d_{eu\,j}$ A distância absoluta de Tj de Ti é escrita relativa para cada atributo onde a distância relativa de cada atributo

$eu\,j$

é obtido dividindo a distância sobre a altura total da hierarquia Formalmente onde hz é a $eu\,j$ n z $\frac{dz}{hz}$ altura da hierarquia de generalização de domínio de domAz Ti

Dadas essas medidas de distância, podemos delinear as seguintes políticas básicas de preferência

Distância absoluta mínima prefere as generalizações que tem uma distância absoluta menor que está com um número total menor de passos de generalização independentemente das hierarquias em que foram tomadas

A distância relativa mínima prefere as generalizações que têm uma distância relativa menor, ou seja, que minimiza o número total de etapas relativas que são consideradas em relação à altura da hierarquia na qual são tomadas

A distribuição máxima prefere as generalizações que contêm o maior número de tuplas distintas

A supressão mínima prefere as generalizações que suprimem menos que contém o maior número de tuplas

Exemplo Considere o exemplo Suponha Generalizações mínimas são e Sob as políticas de distância absoluta preferidas Sob distância relativa mínima distribuição máxima mínima e supressão mínima, as duas generalizações são igualmente preferíveis Suponha que Generalizações mínimas são e Sob a política de distância absoluta mínima, as duas generalizações são igualmente preferíveis Sob a política de supressão mínima é preferida Sob a distância relativa mínima, as duas políticas máxima são preferidas

A lista acima obviamente não está completa e ainda existem políticas de preferência adicionais que podem ser aplicadas a melhor a ser usada obviamente depende do uso específico para os dados liberados O exame de um conjunto exaustivo de possíveis políticas está fora do escopo deste documento A escolha de uma política de preferência específica é feita pelo solicitante no momento do acesso Diferentes políticas de preferência podem ser aplicadas a diferentes quase-identificadores nos mesmos dados liberados

Calculando uma generalização preferida

Definimos o conceito de generalização kminimal preferida correspondente a uma determinada tabela privada Aqui ilustramos uma abordagem para calcular tal generalização Antes de discutir o algoritmo fazemos algumas observações esclarecendo o problema de encontrar uma generalização mínima e sua complexidade Usamos o termo outlier para nos referir a uma tupla com menos de k ocorrências onde k é a restrição de anonimato requeridos

Em primeiro lugar, dado que a propriedade de kanonimato é necessária apenas para atributos em quase-identificadores, consideramos a generalização de cada quasi-identificador específico dentro da tabela independentemente, Consideramos quase-identificador QI A tabela generalizada é obtida aplicando a generalização composta de cada quasiidentier QI

A exatidão da combinação das generalizações produzidas independentemente para cada quase-identificador é assegurada pelo fato de que a definição de uma tabela generalizada requer correspondência de valores em tuplas inteiras e pelo fato de que os quase-identificadores de uma tabela são disjuntos

Na seção, ilustramos os conceitos de uma hierarquia de generalização e estratégias para uma tupla de domínio Dado um quasiidentier QI A An a hierarquia de domínio correspondente em DT hD Dni retrata todas as generalizações possíveis e seus relacionamentos Cada estratégia de caminho define uma maneira diferente na qual a generalização pode ser aplicada Com relação a uma estratégia, poderíamos definir o conceito de generalização mínima local como a generalização que é mínima em relação ao conjunto de generalizações na estratégia intuitivamente o primeiro encontrado no caminho do elemento inferior DT para o elemento superior
Cada generalização kminimal é localmente mínima em relação a alguma estratégia, conforme declarado pelo seguinte teorema

Esta última restrição pode ser removida desde que a generalização de quase-identificadores não disjuntos seja executada serialmente

Teorema Seja TA Um QI seja a tabela a ser generalizada e seja DT hD Dni seja a tupla onde Dz domAz T zn seja uma tabela a ser generalizada Toda kminimal generalização de Ti é uma generalização mínima local para alguma estratégia de DT

Proofsketch Por contradição Suponha que Tj seja kminimal mas não seja localmente mínimo em relação a qualquer estratégia Então existe uma estratégia contendo Tj tal que existe outra generalização Tz dominada por Tj nesta estratégia que satisfaz a canonimato suprimindo não mais tuplas do que o permitido Daí Tz satisfaz condições e de Denição

Além disso, como Tz é dominado por Tj DViz DVij Portanto, Tj não pode ser mínimo, o que contradiz a suposição

Como as estratégias não são disjuntas, o inverso não é necessariamente verdadeiro, ou seja, é uma generalização mínima local em relação a uma estratégia pode não corresponder a uma generalização kminimal

Do teorema seguir cada estratégia de generalização da tupla de domínio para o elemento maximal da hierarquia revelaria então todas as generalizações locais mínimas das quais as generalizações kminimais podem ser selecionadas e uma eventual generalização preferida escolhida A consideração de preferências implica que não podemos parar o pesquisa na primeira generalização encontrada que é conhecida por ser kminimal

No entanto, este processo é muito caro devido ao grande número de estratégias que devem ser seguidas Pode-se provar que o número de estratégias diferentes para uma tupla de domínio DT hD onde cada hi é Dni é $\frac{hhn\ hhn}{}$ o comprimento do caminho de Di para o domínio superior em di

Na implementação de nossa abordagem realizamos um algoritmo que calcula uma generalização preferida sem precisar seguir todas as estratégias e computar as generalizações O algoritmo faz uso do conceito de vetor distância entre tuplas Seja T uma tabela e xy T duas tuplas tal que x hv v ni e y hv é um valor no domínio Di O vetor de distância entre x e y é o vetor e v para seu ancestral comum mais próximo na hierarquia de generalização de v onde cada vd vi

xy dn onde di é o comprimento dos caminhos de v di Por exemplo, com referência ao ilustrado na Figura, a distância entre hasiani e hblacki é Intuitivamente a distância entre duas tuplas x e y na tabela Ti é o vetor distância entre Ti e a tabela Tj com Ti Tj onde os domínios do atributo em Tj são os domínios mais específicos para os quais x e y generalizam para a mesma tupla t

O teorema a seguir estabelece a relação entre vetores de distância entre tuplas em uma tabela e uma generalização mínima para a tabela

Teorema Seja TiA Um QI e Tj são duas tabelas tais que Ti Tj Se Tj é kminimal então ij Vxy para algumas tuplas xy em Ti tais que x ou y tem um número de ocorrências menor que k

Esboço de prova Por contradição Suponha que exista uma generalização kminimal Tj tal que eu j não satisfaz a condição acima Seja dn Considere uma estratégia contendo uma generalização com aquele vetor de distância haverá mais de uma dessas estratégias e qual delas é considerada não importante Considere as diferentes etapas de generalização executadas de acordo com a estratégia de baixo para cima chegando à generalização correspondente a Tj Desde que nenhum outlier está na distância exata d dn de qualquer tupla nenhum outlier é mesclado com qualquer tupla na última etapa da generalização considerada Então a generalização diretamente abaixo de Tj na estratégia satisfaz a mesma restrição de kanonimato que Ti com a mesma quantidade de supressão Também por denição de estratégia

iz eu j Então por Denição Tj não pode ser mínimo, o que contradiz a suposição

De acordo com o Teorema o vector distância de uma generalização mínima cai dentro do conjunto dos vectores entre os outliers e outras tuplas na tabela Esta propriedadeéexplorada pelo algoritmo de generalização para reduzir o número de generalizações a considerar

O algoritmo funciona da seguinte forma Seja QI a projeção do quase-identificador QI Primeiro todas as tuplas distintas em QI são determinadas junto com o número de suas ocorrências Depois a distância

os vetores entre cada outlier e cada tupla na tabela são calculados Em seguida, um DAG com como nós todos os vetores de distância encontrados é construído Há um arco de cada vetor para todo o menor vetor que o domina no conjunto Intuitivamente o DAG corresponde a um resumo do estratégias a serem consideradas nem todas as estratégias podem ser representadas e nem todas as generalizações de uma estratégia podem estar presentes Cada caminho no DAG é seguido de baixo para cima até que uma generalização local mínima seja encontrada O algoritmo determina se uma generalização é localmente mínima simplesmente por controlando como as ocorrências das tuplas se combinariam com base na tabela de distância construída no início sem realmente realizar a generalização Quando uma generalização local é encontrada, outro caminho é seguido Como os caminhos podem não ser disjuntos, o algoritmo mantém o controle das generalizações que foram feitas considerado de modo a parar em um caminho quando se depara com outro caminho no qual um mínimo local já b encontrado Uma vez que todos os caminhos possíveis foram examinados a avaliação dos vetores de distância permite a determinação das generalizações entre aquelas encontradas que são kminimais Entre elas uma generalização preferida a ser calculada é então determinada com base nos vetores de distância e de como as ocorrências de tuplas combinariam

As características que reduzem o custo de computação são, portanto, que o cálculo dos vetores de distância entre as tuplas reduz muito o número de generalizações a serem consideradas as generalizações não são realmente computadas, mas previstas observando como as ocorrências das tuplas se combinariam o fato de o algoritmo acompanhar as generalizações avaliadas permite que ele pare avaliação em um caminho sempre que cruzar um caminho já avaliado

A exatidão do algoritmo descende diretamente dos Teoremas e

A condição necessária e suciente para que uma tabela T satisfaça a canonímia é que a cardinalidade da tabela seja pelo menos k e só neste caso é aplicado o algoritmo Isto é afirmado pelo seguinte teorema

Teorema Seja T um número de jT j seja o limiar de supressão aceitável ek seja um k natural então tabela Se jT j k não vazio existe pelo menos uma generalização kminimal para T Se jT jk não há generalizações mínimas para T

Esboço de prova Suponha que jT jk Considere a generalização generalizando cada tupla para o domínio mais alto possível Como os elementos máximos de são singleton Portanto, valores de atributo jT j ocorrência do mesmo valor tupla Como jT jk satisfaz a canonímia Suponha jT jk nenhuma generalização pode satisfazer a canonímia a que pode ser apenas suprimindo todas as tuplas em T

Aplicação da abordagem alguns resultados experimentais

Construímos um programa de computador que produz tabelas aderindo a generalizações mínimas, dados limites específicos de supressão O programa foi escrito em C usando ODBC para fazer interface com um servidor SQL que por sua vez acessava um banco de dados médico qualidade dos dados divulgados A maioria dos estados tem mandatos legislativos para coletar dados médicos de hospitais, então reduzimos o banco de dados médico original em uma única tabela consistente com o formato e os atributos primários que a Associação Nacional de Organizações de Dados de Saúde recomenda que as agências estaduais coletem Cada tupla representa um paciente e cada paciente é único Os dados continham registros médicos para pacientes A figura lista os atributos usados, a tabela é considerada sem identidade porque não contém informações de identificação explícitas, como nome ou endereço Conforme discutido anteriormente, CEP, data de nascimento e sexo podem ser vinculados a registros populacionais que estão disponíveis publicamente para reidentificar os pacientes Portanto, o quasiidentier QI fZIP data de nascimento gênero etniag foi considerado Cada tupla dentro do QI foi considerada única

A tabela superior na Figura é uma amostra dos dados originais e a tabela inferior ilustra uma generalização mínima dessa tabela dado um limite de supressão O campo ZIP foi generalizado para o

Atributo	valores distintos	frequência mínima	frequência máxima	frequência mediana	comentários
fecho eclair					
Ano de Nascimento					intervalo de anos
Gênero					
Etnia					

Tabela Distribuição dos valores na tabela considerada no experimento

primeiros dígitos e data de nascimento para o ano A tupla com o código postal incomum de foi suprimida O destinatário dos dados é informado dos níveis de generalizações e quantas tuplas foram suprimidas Nota O valor padrão para o mês é janeiro e para o dia é o st quando as datas são generalizadas
Isso é feito por considerações práticas que preservam o tipo de dados originalmente atribuído ao atributo, veja
Seção

A tabela detalha a distribuição básica de valores dentro dos atributos CEPs foram armazenados nos níveis Datas de nascimento ano e forma de dígito completo com uma hierarquia de generalização substituindo os dígitos mais à direita por foram generalizados primeiro para o mês e depois o ano a hierarquia ainda foi considerada para gênero e etnia. O produto do número de possíveis

O programa construiu um clique onde cada nó era uma tupla e as arestas eram ponderadas por vetores de distância entre tuplas adjacentes. Lendo esses vetores do clique, o programa gerou um conjunto de generalizações a serem consideradas. generalizações lidas a partir do descarte do clique e ou
Para nossos testes usamos valores de k to be ou tuplas um limite máximo de supressão de

A figura mostra a relação entre supressão e generalização dentro do programa em uma aplicação prática e realista. Medimos a perda de qualidade dos dados devido à supressão como a razão entre o número de tuplas suprimidas dividido pelo número total de tuplas nos dados originais. medida inversa de completude para determinar quanto dos dados permanece calculado como um menos a perda devido à supressão A generalização também reduz a qualidade dos dados, pois os valores generalizados são menos precisos Medimos a perda devido à generalização como a razão do nível de generalização dividido pela altura total da hierarquia de generalização Nós denominamos precisão como a quantidade de especificidade remanescente nos dados calculados como um menos a perda devido à generalização

Nos Gráficos A e B da Figura, comparamos a perda de qualidade dos dados à medida que o requisito de kanonimato aumenta. valores encontrados nestes atributos Dada a distribuição de homens e mulheres nos dados, o próprio atributo sexo pode atingir esses valores de k, então não vemos nenhuma perda devido à generalização ou supressão

Por outro lado existiam datas de nascimento distintas Claramente a data de nascimento e o código postal são os valores mais discriminativos pelo que não é de estranhar que tenham de ser generalizados mais do que outros atributos
As linhas nessas curvas indicam que os valores estão um pouco agrupados

Os gráficos C e D da Figura relatam medições de completude e precisão para as generalizações mínimas encontradas Basicamente, as generalizações que satisfazem valores menores de k aparecem mais à direita no gráfico C e as generalizações que atingem valores maiores de k são mais à esquerda. quanto maior o valor de k mais generalização pode ser necessária resultando, é claro, em uma perda de precisão Também não é surpreendente que a completude permaneça acima porque nosso limite de supressão durante esses testes foi

Embora não seja mostrado nos gráficos, pode-se entender facilmente que aumentar o limite de supressão geralmente melhora a precisão, pois mais valores podem ser suprimidos para atingir k Claramente a generalização é cara para a qualidade dos dados, pois é executada em todo o atributo, cada tupla é afetada Por outro lado, permanece semanticamente mais útil ter um valor

Figura Exemplo de prática de lançamento atual e equivalente minimamente generalizado

Figura Resultados experimentais baseados em registros médicos

presente, mesmo que seja menos preciso do que não ter nenhum valor, pois é o resultado da supressão

A partir desses experimentos fica claro que as técnicas de generalização e supressão podem ser usadas em aplicações práticas É claro que a proteção contra vinculação envolve uma perda de qualidade de dados nos atributos que compõem o quase-identificador embora tenhamos mostrado que a perda não é severa Essas técnicas são claramente mais eficaz quando os atributos primários exigidos pelo destinatário não são os mesmos que o quase-identificador que pode ser usado para vincular. a fim de desenvolver ferramentas de diagnóstico realizar pesquisas retrospectivas e avaliar os custos hospitalares

Conclusões

Apresentamos uma abordagem para divulgar informações específicas da entidade de modo que a tabela liberada não possa ser vinculada de forma confiável a tabelas externas O requisito de anonimato é alcançado generalizando e possivelmente suprimindo as informações após a liberação Demos a noção de generalização mínima capturando a propriedade de que a informação não é generalizada mais do que o necessário para atingir o requisito de anonimato Discutimos possíveis políticas de preferência para escolher entre diferentes generalizações mínimas e um algoritmo para calcular uma generalização mínima preferida Finalmente ilustramos os resultados de algumas experiências da aplicação da nossa abordagem ao lançamento de uma base de dados médica contendo informação relativa aos pacientes

Este trabalho representa apenas um primeiro passo para a definição de uma estrutura completa para controle de divulgação de informações Muitos problemas ainda estão em aberto Do ponto de vista da modelagem, a definição de quase-identificadores e de um tamanho apropriado de k deve ser abordada A qualidade dos dados generalizados é melhor quando os atributos mais importantes para o destinatário não pertencem a nenhum quase-identificador Para arquivos de uso público isso pode ser aceitável, mas a determinação da qualidade e utilidade em outras configurações deve ser mais pesquisada Do ponto de vista técnico, trabalhos futuros devem incluir a investigação de um algoritmo eficiente para aplicar as técnicas propostas e a consideração de consultas específicas de vários lançamentos ao longo do tempo e de atualização de dados que possam permitir ataques de inferência

Agradecimentos

Agradecemos a Steve Dawson, da SRI, pelas discussões e apoio a Rema Padman, da CMU, pelas discussões sobre métricas e ao Dr. Lee Mann, da Inova Health Systems Lexical Technology Inc, e ao Dr. Fred Chu, por disponibilizar dados médicos para validar nossas abordagens. Também agradecemos a Sylvia Barrett e Henry Leitner da Universidade de Harvard pelo apoio

Referências

NR Adam e JC Wortman Métodos de controle de segurança para bancos de dados estatísticos Um estudo comparativo ACM Computing Surveys

Ross Anderson Um modelo de política de segurança para sistemas de informação clínica In Proc of the IEEE Páginas do Simpósio sobre Segurança e Privacidade Oakland CA maio

Silvana Castano Maria Grazia Fugini Giancarlo Martella e Pierangela Samarati Segurança de Banco de Dados Addison Wesley

Metodologia de supressão de PC Chu Cell A importância de suprimir totais marginais IEEE Trans em sistemas de dados de conhecimento Julho agosto

Metodologia de supressão de LH Cox na análise de divulgação estatística em ASA Proceedings of Social Páginas da seção de estatísticas

Tore Dalenius Encontrando uma agulha no palheiro ou identificando um registro de censo anônimo Journal of Estatísticas Oficiais

BA Davey e HA Priestley Introdução a Lattices and Order Cambridge University Press

Dorothy E Denning Criptografia e Segurança de Dados AddisonWesley

Dan Guseld Um pouco de conhecimento ajuda muito Detecção mais rápida de dados comprometidos em tabelas D In Proc do IEEE Symposium on Security and Privacy pages Oakland CA maio

J Hale e S Shenoi Análise de inferência catalítica Detectando ameaças de inferência devido à descoberta de conhecimento Em Proc do Páginas do Simpósio IEEE sobre Segurança e Privacidade Oakland CA maio

A Hundepool e L Willenborg e Argus Software para controle de divulgação estatística em Terceiro Seminário Internacional sobre Confidencialidade Estatística Bled

Ram Kumar Garantindo a segurança de dados em dados tabulares inter-relacionados In Proc of the IEEE Symposium on Páginas de segurança e privacidade Oakland CA maio

Teresa Lunt Agregação e inferência Fatos e falácias In Proc do IEEE Symposium on Páginas de segurança e privacidade Oakland CA maio

Associação Nacional de Organizações de Dados de Saúde Falls Church Um Guia para o Ambulatório Estadual Atividades de coleta de dados de assistência Outubro

P Samarati e L Sweeney Generalizando dados para fornecer anonimato ao divulgar informações Em Procedimento do ACM SIGACTSIGMODSIGART Simpósio sobre Princípios de Sistemas de Banco de Dados PODS Seattle EUA junho

Latanya Sweeney Controle de divulgação computacional para microdados médicos In Record Linkage Workshop Bureau do Censo Washington DC

Latanya Sweeney Garantia de anonimato ao compartilhar dados médicos o sistema Datay In Proc Jornal da Associação Americana de Informática Médica Washington DC Hanley Belfus Inc

Latanya Sweeney Unindo tecnologia e política para manter a confidencialidade Journal of Law Ética na Medicina

Questões de privacidade de informações Rein Turn para o s In Proc do IEEE Symposium on Security and páginas de privacidade Oakland CA maio

Jerey D Ullman Princípios de Bancos de Dados e ConhecimentoSistemas Básicos volume I Ciência da Computação Imprensa

L Willenborg e T De Waal Controle de divulgação estatística na prática New York SpringerVerlag

L Willenborg e T De Waal Controle de Divulgação Estatística na Prática SpringerVerlag

Beverly Woodward A confidencialidade do registro do paciente baseado em computador The New England Journal of Medicina