Dolování dat a bezpečnost

PV056

May 11, 2006

Obsah přednášky

Citlivé informace

Možnost a nutnost zpracování velkých objemů dat vede k problémům s citlivostí některých informací.
Citlivost informací je v zásadě dvojího druhu:

- daná ze zákona - viz. Úřad pro ochranu osobních údajů
  `http://www.uoou.cz`
  Zákon č. 101/2000 Sb., o ochraně osobních údajů
  a další normy . . .
- daná osobním pohledem na věc - některé informace z jistých důvodů prostě nechceme zveřejňovat

Citlivé informace

Možnost a nutnost zpracování velkých objemů dat vede k problémům s citlivostí některých informací.
Citlivost informací je v zásadě dvojího druhu:

- daná ze zákona - viz. Úřad pro ochranu osobních údajů
  `http://www.uoou.cz`
  Zákon č. 101/2000 Sb., o ochraně osobních údajů
  a další normy . . .
- daná osobním pohledem na věc - některé informace z jistých důvodů prostě nechceme zveřejňovat

Jednoduše řečeno: Data, která vedou k identifikaci konkrétní osoby, podléhají ochraně zákonem. . .

Různé pohledy na soukromí při dolování dat

Čistě naivním pohledem můžeme rozlišit různé typy situací, ve kterých je třeba chránit soukromí při dolování dat:

- uživatel - citlivost zadávaných dotazů, vystopovatelnost

Různé pohledy na soukromí při dolování dat

Čistě naivním pohledem můžeme rozlišit různé typy situací, ve kterých je třeba chránit soukromí při dolování dat:

- uživatel - citlivost zadávaných dotazů, vystopovatelnost
- data - citlivé informace, možnost zneužití

Různé pohledy na soukromí při dolování dat

Čistě naivním pohledem můžeme rozlišit různé typy situací, ve kterých je třeba chránit soukromí při dolování dat:

- uživatel - citlivost zadávaných dotazů, vystopovatelnost
- data - citlivé informace, možnost zneužití
- vlastník databáze - neoprávněné šíření, výsledky dolování použitelné proti vlastníkovi

Dolování se zachováním soukromí

PPDM - *Privacy Preserving Data Mining* je oblast, zabývající se dolováním s přihlédnutím k bezpečnostní stránce zpracovávaných dat.
PPDM encompasses the dual goal of meeting privacy requirements and providing valid data mining results - (Oliveira, Zaine)

Dolování se zachováním soukromí

PPDM - *Privacy Preserving Data Mining* je oblast, zabývající se dolováním s přihlédnutím k bezpečnostní stránce zpracovávaných dat. *PPDM encompasses the dual goal of meeting privacy requirements and providing valid data mining results - (Oliveira, Zaine)*

Základní přístupy:

- modifikace dat před procesem dolování
  - odstranění citlivých dat
  - změna hodnot dat

Dolování se zachováním soukromí

PPDM - *Privacy Preserving Data Mining* je oblast, zabývající se dolováním s přihlédnutím k bezpečnostní stránce zpracovávaných dat. *PPDM encompasses the dual goal of meeting privacy requirements and providing valid data mining results - (Oliveira, Zaine)*

Základní přístupy:

- modifikace dat před procesem dolování
  - odstranění citlivých dat
  - změna hodnot dat
- použití nezměněných dat - kryptografické přístupy

Dolování se zachováním soukromí

PPDM - *Privacy Preserving Data Mining* je oblast, zabývající se dolováním s přihlédnutím k bezpečnostní stránce zpracovávaných dat. *PPDM encompasses the dual goal of meeting privacy requirements and providing valid data mining results - (Oliveira, Zaine)*

Základní přístupy:

- modifikace dat před procesem dolování
  - odstranění citlivých dat
  - změna hodnot dat
- použití nezměněných dat - kryptografické přístupy
- modifikace výsledků

Modifikace dat

Perturbation, Obfuscation, Distortion

- zašumění (perturbation)
- blokace (blocking)
- agregace (agregation)
- záměna (swapping)
- vzorkování (sampling)

Modifikace výsledků

Podobné metody jako při modifikaci vlastních dat.
Problém citlivosti řešíme až po získání výsledků.
Výsledky obsahující citlivé informace můžeme odstranit, zamlžit (přidáním šumu), ...

Lze také kombinovat s technikami modifikace dat:

- nejdříve provedeme dolování na nezměněných datech
- analyzujeme citlivá data, která jsme získali - zjistíme, která původní data je třeba změnit, abychom tato data nezískali
- provedeme patřičné změny na původních datech

Dolování na původních datech

Data zůstávají nezměněná $\Rightarrow$ soukromí je potřeba zachovat jiným způsobem.
Zaměříme se proto na vlastní proces dolování.

Dolování na původních datech

Data zůstávají nezměněná $\Rightarrow$ soukromí je potřeba zachovat jiným způsobem.
Zaměříme se proto na vlastní proces dolování.

Hlavní myšlenka - citlivá data se vůbec do nepovolaných rukou nedostanou...

Dolování na původních datech

Data zůstávají nezměněná $\Rightarrow$ soukromí je potřeba zachovat jiným způsobem.
Zaměříme se proto na vlastní proces dolování.

Hlavní myšlenka - citlivá data se vůbec do nepovolaných rukou nedostanou...

Často je třeba vyřešit problém v distribuovaném prostředí - více uživatelů chce dolovat nad společnými daty, ale přitom nikdo nechce svoje data zpřístupnit ostatním uživatelům...
Vytváření protokolů, které to umožní - využívání kryptografických postupů

SMC

SMC - “Secure Multiparty Computation”
Protokol mezi dvěma a více stranami. Počítáme nějakou funkci nad vstupními daty.
Účelem je, aby po skončení výpočtu každá strana vlastnila pouze svůj vstup a výslednou hodnotu.
Prakticky jakýkoliv protokol, který běhá v kruhu, lze předělat na SMC...
Typické příklady SMC protokolů:

- součet
- množinové operace (velikost průniku, ...)
- skalární produkt

Příklad SMC - bezpečný součet

Problém - spočítat součet čísel, která vlastní skupina uživatelů a přitom utajit hodnoty jednotlivých čísel.
algoritmus:

- seřadíme uživatele (1..n)
- první vygeneruje náhodné číslo $R(0 \leq R \leq n)$ a vypočítá $(u_1 + R)$ mod n
- pošle výsledek druhému v pořadí
- každý uživatel i poté přičte svoje číslo u_i a pošle výsledek mod n dalšímu
- když první dostane zpět výsledek, odečte náhodné číslo R a zbylým uživatelům pošle součet u

Pokud nikdo nebude s nikým spolupracovat, bude každý uživatel na konci znát pouze součet u a svou hodnotu u_i.

Další dělení PPDM technik

PPDM techniky lze dělit také podle několika dalších kritérií:

- Distribuce dat
  - centralizované
  - distribuované horizontálně
  - distribuované vertikálně

Další dělení PPDM technik

PPDM techniky lze dělit také podle několika dalších kritérií:

- Distribuce dat
  - centralizované
  - distribuované horizontálně
  - distribuované vertikálně
- Použitý algoritmus
  - rozhodovací stromy
  - asociační pravidla/časté vzory
  - shlukování

Příklady základních algoritmů

Oblivious Transfer Protocol (OTP)
protokol mezi dvěmi stranami, kde jedna strana (server) vlastní tajný bit b a druhá strana (uživatel) po proběhnutí tohoto protokolu s pravděpodobností 1/2 tento bit zjistí. Hlavní je, že server neví, jestli uživatel informaci získal či ne...

Existují i složitější verze - 1-out-of-N OT, distribuovaný OT, ...

Příklady základních algoritmů

Private Information Retrieval (PIR)
Protokol umožňuje uživateli získat informaci z databáze a přitom jeho dotaz uchovat v tajnosti...
Příklad: databáze s lékařskými daty, patenty apod. Uživatel nemá zájem zveřejnit dotaz někomu jinému. Pokud zveřejníme náš dotaz do databáze patentů, který hledá vše související s nějakým tématem, odhalujeme náš zájem něco v tomto směru patentovat...

Příklady základních algoritmů

Private Itemset Support Counting (PISC)
Zde chce uživatel zjistit support nějaké množiny a je to umožněno se stejnými zárukami jako v předcházejících příkladech...

Bezpečné dolování asociačních pravidel

- na vertikálně rozdělených datech
  každý itemset rozdělen - atributy uchovávány na různých místech
  cílem je spočítat support
  založeno na skalárním součinu
- na horizontálně rozdělených datech
  globální support je součtem lokálních supportů
  DIODA (Jan Frieser)

DIODA

založeno na FDM modelu (*fast ditributed mining of association rules*)
pouze semi-honest přístup, dvě použití SMC protokolu
bezpečné sjednocení (založeno na komutativním šifrování) pro vztvoření množin k-itemsetů s dostatečným globálním supportem
bezpečný součet pro nalezení velikosti globálních supportů

Dvě strany:

- server
  - generuje páry klíčů pro komutativní šifrování
  - synchronizuje distribuované výpočty
  - neshromažďuje žádná data, neprovádí samotné dolování
- klient
  - zpracovává svou část horizontálně rozdělených dat

DIODA

FDM model má čtyři hlavní kroky:

- generování množin kandidátních itemsetů (vytvoření lokálních k-itemsets z průniku lokálních a globálních (k-1)-itemsetů
- lokální prořezávání množin (jen ty lokálně podporované jsou použity)
- výměna kandidátních itemsetů (sjednocení lokálních množin)
- vytvoření globálně frekventovaných itemsetů (výpočet globálního supportu)

implementační detaily DIODY

- Apriori algoritmus importován z Weky
- komunikace mezi klienty a serverem pomocí GNU OmniSockets
- komutativní šifrování pomocí upravené verze RSA (použití bezpečných prvočísel)

Porovnávání PPDM algoritmů

Je několik možných hledisek, podle kterých můžeme PPDM algoritmy porovnávat:

- výkon (časová a výpočetní složitost)
- užitečnost dat (data utility) - po použití PPDM algoritmu
- úroveň nejistoty (level of uncertainty) s jakou mohou být ukryté informace přece jen vydolovány
- odolnost vůči jiným DM technikám
- ...

Dolování v oblasti IT bezpečnosti

Metody dolování v datech jsou užitečné i v oblasti IT bezpečnosti. Jejich přínos je znát zejména při analýze nějakého chování.

- Intrusion Detection Systems (IDSs)
- Reputation systems (internet auctions)
- Forensics (criminal behaviour)
- ...

IDS

Intrusion: “Any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource.” Existují samozřejmě

techniky pro prevenci napadení - autentizace uživatelů, kontrola programátorských chyb, . . .
Často však nejsou samy o sobě dostačující. Vždy najdeme nějakou chybu . . .
IDS jsou nástroje pro automatickou detekci útoků v systému.

Mají tři základní části:

- zdroje - ty se mají ochránit
- modely - charakterizují normální chování systému
- techniky - porovnávání aktuálních aktivit systému s vytvořenými modely

IDS

Dvě základní techniky, které jsou v IDSs používány (často v kombinaci):

- detekce známých vzorů útoků a slabých míst systému
- detekce anomálií v chování - budeme danou anomálii považovat za útok?

Použití dolování v IDSs:

- klasifikace normálního a abnormálního chování systému
- analýza vztahů mezi jednotlivými událostmi
- analýza sekvencí - modelujeme časté vzory, vyskytující se v systému

Nutné používání učení bez učitele a adaptivního učení (aktualizace IDS)

IDS

Dobrým příkladem využití IDS je analýza síťového provozu

detekce a zabránění DoS útokům (SYN Flood, Ping of Death, Teardrop,...).
Většina těchto útoků má svoje typické chování, které lze detekovat a občas taky útoku zabránit...

Computer forensic

Computer forensic science is the science of acquiring, preserving, retrieving and presenting data that has been processed electronicaly and stored on computer media. (Noblett, at al., FBI)

Analýza dat a hledání důkazů použitelných pro soudy apod.

- hledání typických vzorů pro spronevěry, tunelování...
- analýza e-mailové korespondence podezřelých
- ...

Reputační systémy

Analýza chování uživatelů. Uživatelé zde nemají účty, mohou vystupovat pod pseudonymy. Existuje pouze informace o jejich předchozím jednání. . . Tyto systémy mají tři typy uzlů:

- requesters (uživatelé)
- servers (nabízí nějaké služby)
- recommenders (poskytují informace o chování uživatelů)

Reputační systémy

Server se musí rozhodnout, jestli přijme nebo odmítne požadavek uživatele.

- vyhledá vzory vedoucí k nějakému špatnému výsledku - tyto vzory jsou vytvořeny pomocí technik dolování z lokální evidence a dat od recommenders
- porovná výsledky se špatnými vzory a rozhodne se

The R project for statistical computing

http://www.r-project.org

The R project for statistical computing

`http://www.r-project.org`

- GNU projekt (GPL), založeno na jazyce S (Bell laboratories, John Chambers)
- jazyk a prostředí pro statistické výpočty a tvorbu grafiky (grafy apod.)
- jednoduchá práce s různými datovými strukturami

The R project for statistical computing

`http://www.r-project.org`

- GNU projekt (GPL), založeno na jazyce S (Bell laboratories, John Chambers)
- jazyk a prostředí pro statistické výpočty a tvorbu grafiky (grafy apod.)
- jednoduchá práce s různými datovými strukturami
- nabízí mnoho statistických metod - lineární i nelineární modelování, klasické statistické testy, analýza časových řad, klasifikace, shlukování,...)
- grafické techniky - dobré grafy vhodné pro publikační účely, ...
- balíky pro velké množství různých úloh
- možnost včlenění programů napsaných v C, C++ a Fortranu
- verze pro UNIX, Win i MacOS

R v LVZ

spuštění `/pub/packages/R/bin/R`

```
norm ¡- function(t){
for (i in 1:length(t)){
        ma ¡- max(t[[i]])
                mi ¡- min(t[[i]])
                po ¡- (ma - mi)
                for (j in 1:length(t[[i]])){
                t[[i]][j] ¡- (t[[i]][j]-mi)/po
                        }
                }
        return(t)
        }
```

Internetové zdroje

- Kun Liu, University of Maryland, Baltimore County `http://www.csee.umbc.edu/~kunliu1/research/privacy_review.html`
- Helger Lipmaa, University of Tartu (Estonia) `http://www.cs.ut.ee/~lipmaa/crypto/link/data_mining/`

Hlavní představitelé výzkumu...

- Chris Clifton - Purdue University, Indiana (USA)
  `http://www.cs.purdue.edu/people/clifton` PPDM
- Stanley R. M. Oliveira - University of Alberta, Edmonton (Canada)
  `http://www.cs.ualberta.ca/oliveira/` PPDM
- Hillol Kargupta - University of Maryland, Baltimore County (USA)
  `http://www.csee.umbc.edu/~hillol/Kargupta/` PPDM
- Helger Lipmaa - Helsinki University of Technology (FIN)
  `http://www.cs.ut.ee/~lipmaa/` cryptography
- Benny Pinkas - University of Haifa (Israel), HP Labs
  `http://www.pinkas.net` cryptography