CNRIA 2010

Datamining et vie privée

Fodé Camara* — Yahya Slimani** — Samba Ndiaye*

* Département mathématiques-informatique, Faculté des Sciences et Techniques
Université Cheikh Anta Diop de Dakar
SENEGAL
{fode.camara, samba.ndiaye}@ucad.edu.sn

** Département d'informatique, Faculté des Sciences
Université Tunis
TUNISIE
yahya.slimani@fst.rnu.tn

RÉSUMÉ. Récemment le problème de la protection de la vie privée est devenu important en datamining, surtout quand les données sont partitionnées sur plusieurs sites. Dans le cas d'un partitionnement vertical, plusieurs problèmes de datamining peuvent simplement être réduits au calcul sécurisé du produit scalaire. Parmi ces problèmes nous pouvons citer l'extraction de règles d'association sur des données partitionnées verticalement. L'efficacité du calcul sécurisé du produit scalaire peut se mesurer par le nombre de messages nécessaire pour assurer la protection de la vie privée. Pour les différentes solutions qui ont été proposées pour préserver la vie privée dans l'extraction de règles d'association à partir de données partitionnées verticalement, ce nombre de messages est souvent excessif. Dans ce papier, nous proposons un nouveau protocole de calcul sécurisé du produit scalaire qui réduit considérablement ce coût de communication.

ABSTRACT. Recently, privacy issues have becomes important in data mining, especially when data is partitioned over several parties. For the vertically partitioned case, many data mining problems can essentially be reduced to securely computing the scalar product. Among these problems, we can mention association rule mining over vertically partitioned data. Efficiency of a secure scalar product can be measured by the overhead of communication needed to ensure this security. Several solutions have been proposed for privacy preserving association rule mining in vertically partitioned data. But the main drawback of these solutions is the excessive overhead communication needed for ensuring data privacy. In this paper we propose a new secure scalar product with the aim to reduce the overhead communication.

MOTS-CLÉS : datamining, règles d'association, vie privée, cryptographie.

KEYWORDS : data mining, associate rules, privacy, cryptography

1. Introduction

Les nouvelles technologies de l'information permettent à la fois un stockage de larges volumes de données et une contribution à leur croissance exponentielle. Ceci a pour conséquence de créer une disproportion entre les volumes de données et les moyens matériels et humains pour les traiter. Face à ces problèmes, les technologies du datamining permettent de traiter ces ensembles de données en vue d'extraire de la connaissance, qui sera déterminante pour la prise de décisions efficaces. Cependant l'application de ces technologies sur des données à caractère personnel permet de définir des profils d'individus, de prédire leurs comportements et d'agir en conséquence. Or, une telle connaissance se heurte aux droits des personnes qui veulent préserver les données relatives à leur vie privée. Récemment, de nouvelles lois ont été mises en place pour devenir de nouvelles contraintes sur la confidentialité des données, afin de préserver la vie privée des personnes. Parmi ces lois, nous pouvons citer :

– la loi HIPAA (*Health Insurance Portability and Accountability Act*)

– la Directive n° 95/46 du Parlement européen et du conseil du 24 octobre 1995 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données

– etc.

Bien entendu le problème de confidentialité peut être adapté à de nombreux domaines, comme par exemple l'analyse de transactions financières, l'analyse de comportements sur des sites de e-commerce, etc. Comme scénario, nous pouvons citer l'exemple de deux compagnies de bioinformatique. Chaque compagnie possède une base de données gigantesque constituée de mesures collectées à partir d'expériences effectuées dans leurs laboratoires. Les deux sont prêtes à coopérer pour réaliser une tâche d'apprentissage d'intérêt commun, mais aucune d'elles ne souhaitent communiquer sa base de données en clair. Comment peuvent-elles atteindre ce but sans divulguer aucune information sensible ?

PPDM (*Privacy Preserving Data Mining*) est un domaine émergeant qui étudie comment les algorithmes de datamining affecte la protection de la vie privée et essaye de trouver et d'analyser de nouveaux algorithmes qui vont respecter cette contrainte de vie privée.

Dans ce papier, nous nous intéressons particulièrement au problème de la préservation de vie privée dans l'extraction de règles d'association à partir de données distribuées verticalement. Pour un partitionnement vertical, l'extraction de règles d'association peut se réduire au calcul du produit scalaire. Cette primitive est utilisée de façon répétée dans ce processus, son calcul de manière efficace et sécurisée est donc important. L'efficacité de ce calcul peut se mesurer par le nombre de messages nécessaire pour assurer la protection de la vie privée. Pour les différentes solutions qui ont été proposées, ce nombre de messages

est souvent excessif. Dans ce papier, nous proposons un nouveau protocole de calcul sécurisé du produit scalaire qui réduit considérablement ce coût de communication.

Le reste de ce papier est organisé comme suit. La section 2 fournit l'état de l'art du datamining et vie privée. Dans la section 3, nous présentons l'approche proposée. L'évaluation du coût de calcul et de communication de notre protocole est présentée dans la section 4. Dans la section 5 nous analysons la sécurité de notre proposition. La section 6 évalue ce travail en le comparant à des travaux relatifs. Enfin, la section 7 résume l'ensemble de nos travaux et donne quelques perspectives de leur prolongement.

2. Datamining et vie privée

Le problème de la protection de vie privée devient de plus en plus important ces dernières années à cause des nombreux besoins de partage de données, de la contrainte de vie privée et du besoin sans cesse croissant d'extraire de la connaissance à partir de données. Deux problèmes sont étudiés dans ce domaine : le premier est la protection des données privées ; le second est la protection de la connaissance sensible contenue dans les données, ce problème est plus connu sous le terme KHD (*Knowledge Hiding in Database*). Nous pouvons classer les techniques de KHD en deux groupes : les approches basées sur la modification des données et les approches basées sur la reconstruction. L'idée de base des approches de modification des données est de modifier directement la base de données originale, nous pouvons pousser encore la classification en distinguant deux familles d'algorithmes : ceux utilisant des techniques basées sur la distorsion et ceux utilisant des techniques basées sur le blocage des données. La distorsion ou la perturbation est le changement d'une valeur d'attribut par une nouvelle valeur [1] (par exemple le changement de la valeur 1 en 0), tandis que le blocage est le remplacement d'une valeur d'attribut existante par ? [2]. Toutefois, les approches basées sur la modification des données ne contrôlent pas directement les effets de bord (i.e. une règle non-fréquente peut devenir fréquente après le processus de préservation de vie privée) et nécessitent de nombreuses opérations d'E/S, surtout quand la base de données initiale contient plusieurs transactions. Une autre approche consiste à utiliser des méthodes basées sur la reconstruction des données [3]. L'idée de base dans ces approches est d'extraire d'abord la connaissance K à partir de la base de données originale. Ensuite la nouvelle base de données D' est alors reconstruite à partir de K. L'idée de base de la reconstruction des données est inspirée d'un problème récent appelé *Inverse Frequent Set Mining* [3]. Contrairement aux techniques de l'autre approche, les techniques de reconstruction contrôlent directement les effets de bord. La principale proposition pour résoudre le problème de la protection des données privées est le calcul distribué sécuritaire multi-parties ou SMC (*Secure Multi-Party Computation*), dont la problématique est de permettre le calcul d'une fonction quelconque sur un ensemble de données réparties entre plusieurs entités. Chaque entité posséde une partie des données et le calcul doit être réalisé de manière à ce qu'aucune des parties ne

puisse déduire, de quelque manière que ce soit, les données des autres entités à partir des résultats du calcul et de ses propres données. Les études dans ce domaine ont été initiées par Yao dans [4, 5] pour une approche bipartie. Elles ont ensuite été généralisées à un nombre quelconque de parties dans [6]. Cette approche a été introduite dans la communauté datamining pour la première fois par Lindell et Pinkas dans [7], avec une méthode permettant à deux parties de construire un arbre de décision sans qu'aucune des parties ne puisse apprendre quelque chose sur l'autre. Depuis, des techniques ont été développées pour l'extraction de règles d'association dans le cas des bases de données distribuées verticalement [8, 9] et horizontalement [10], la recherche de motifs séquentiels dans un contexte collaboratif [11], etc.

Le protocole proposé dans ce papier est une technique de calcul distribué sécuritaire multi-parties.

3. Approche proposée

3.1. Définition du problème

Le produit scalaire est un composant technique puissant. Plusieurs problèmes de datamining peuvent essentielement être réduits au calcul du produit scalaire. Pour donner une idée sur la manière dont cette primitive est utilisée, considérons le problème de l'extraction des règles d'association dans une base de données distribuée verticalement. Le problème de l'extraction des règles d'association peut être formulé comme suit : Soit $I = \{i_1, ..., i_n\}$ un ensemble d'items. Soit D un ensemble de transactions où chaque transaction est composée d'un identifiant unique *TID* et contient un ensemble d'items tels que $T \subset I$. Une transaction T contient X, un ensemble d'items dans I, si $X \subseteq T$. Une règle d'association est une implication de la forme $X \Rightarrow Y$, où $X \subseteq I$, $Y \subset I$, et $X \cap Y = \emptyset$. La règle $X \Rightarrow Y$ existe dans la base de données D avec une confiance c, si $c\%$ des transactions qui contiennent X contiennent aussi Y. La règle $X \Rightarrow Y$ a un support s, si $s\%$ des transactions de D contiennent $X \cup Y$. Dans ce cadre, nous considérons une représentation binaire de la base de transactions. L'absence ou la présence d'un item est représentée par une valeur prise dans $\{0, 1\}$. Les transactions sont représentées par des chaines de 0 et de 1, quand la base de transactions peut être représentée par une matrice de $\{0, 1\}$. L'algorithme d'extraction de règles d'association dans une base de données distribuées verticalement est basé sur l'algorithme Apriori de Agrawal et Srikant [12]. Le problème principal de cet algorithme est de trouver tous les itemsets fréquents dans la base de données. Pour déterminer si un itemset donné est fréquent ou non, nous comptons le nombre de transactions, dans D, quand la valeur de tous les attributs dans l'itemset est égale à 1. Ce problème peut être transformé en un simple problème de mathématique utilisant les définitions suivantes : Soit $l + m$ le nombre d'attributs de la base de données,

le site A a l attributs, $\{A_1, ..., A_l\}$ et le site B a m attributs, $\{B_1, ...B_m\}$. Les transactions sont des séquences de $l + m$ 1 ou 0. Soit $minsupp$ le support minimal, et n le nombre total de transactions dans la base de données D. Soient $\overrightarrow{X}$ et $\overrightarrow{Y}$ les colonnes dans D, i.e. $x_i = 1$ si la ligne i a la valeur 1 pour l'item ou l'attribut X. Le produit scalaire des deux vecteurs $\overrightarrow{X}$ et $\overrightarrow{Y}$ de cardinalité n est défini comme suit : $\overrightarrow{X} \bullet \overrightarrow{Y} = \sum_{i=1}^{n} x_i \times y_i$. Déterminer si le 2-itemset $< XY >$ est fréquent peut être réduit à tester si $\overrightarrow{X} \bullet \overrightarrow{Y} \geq minsupp$. La généralisation de cette procédure à un w-itemset est facile. Supposons que le site A a p attributs, $a_1, ..., a_p$ et B a q attributs, $b_1, ..., b_q$. Nous voulons calculer la fréquence du w-itemset $< a_1...a_p, b_1...b_q >$, où $w = p + q$. Chaque item dans $\overrightarrow{X}$ (respectivement dans $\overrightarrow{Y}$) est composé du produit des éléments individuels correspondants, i.e., $x_i = \prod_{j=1}^{p} a_j$ (respectivement $y_i = \prod_{j=1}^{q} b_j$).

A ce stade, nous pouvons formuler notre problème comme suit : Supposons que nous avons deux parties, par exemple $Alice$ et Bob tel que chacune d'elles posséde un vecteur binaire de cardinalité n, i.e. $\overrightarrow{X} = (x_1, ..., x_n)$ et $\overrightarrow{Y} = (y_1, ..., y_n)$. Le problème est de calculer de manière sécurisée le produit scalaire de ces deux vecteurs, i.e. $\overrightarrow{X} \bullet \overrightarrow{Y} = \sum_{j=1}^{n} x_i \times y_i$.

3.2. Outils cryptographiques

Pour définir notre protocole de calcul sécurisé du produit scalaire, nous avons utilisé un cryptosystème à clé publique homomorphe additif et sémantiquement sécurisé. Afin d'obtenir la sécurité souhaitée dans la transmission des données, nous avons choisi un cryptage asymétrique. Dans ce type de cryptage, une clé est utilisée pour crypter et une autre pour décrypter. La clé secrète est conservée par l'usager, en toute sécurité. Le cryptosystème que nous avons choisi est homomorphe. Ce choix est motivé par le fait que, connaissant $Enc(x)$ et $Enc(y)$, nous pouvons calculer $Enc(x \perp y)$ sans déchiffrer x, y pour une certaine opération $\perp$. De plus, le cryptosystème homomorphe est additif ; en d'autres termes, une partie peut effectuer une opération d'addition sur les messages en clair en faisant simplement des calculs sur les messages chiffrés, sans disposer de la clé secrète. Enfin, le cryptosystème choisi est sémantiquement sécurisé. Cette propriété est très importante dans notre contexte binaire, car un adversaire peut toujours crypter 0 et 1 en utilisant la clé publique et comparer alors le résultat des chiffrés correspondants avec le message reçu pour déduire la vraie valeur du bit. Un des plus populaires cryptosystèmes qui regroupe les quatres propriétés citées ci-haut est le cryptosystème de Paillier [13].

3.3. Algorithme

Dans notre approche proposée, Alice génère d'abord une paire de clés et calcule $Enc_{pk}(x_i, r)$ qu'elle envoie à Bob. La sécurité sémantique, équivalente à indisguabilité [13], garantit qu'aucune information n'est révélée à travers ce message. Alice envoie aussi la clé publique à Bob. Bob calcule $\prod_{i=1}^{n} p_i$, avec $p_i = Enc_{pk}(x_i)$ si $y_i = 1$, $p_i = 1$ dans

Algorithm 1 Private Scalar Product Protocol

Require: N=2 (Nombre de sites ; Alice et Bob),
Require: Le vecteur d'Alice : $\overrightarrow{X} = (x_1, ..., x_n)$,
Require: Le vecteur de Bob : $\overrightarrow{Y} = (y_1, ..., y_n)$
1: **for** Alice **do**
2: Générer une paire de clé (sk, pk) ;
3: Générer $(Enc_{pk}(x_1), ..., Enc_{pk}(x_n))$ en utilisant la clé pk ;
4: Envoi de $(Enc_{pk}(x_1), ..., Enc_{pk}(x_n))$ à Bob ;
5: **end for**
6: **for** Bob **do**
7: Calculer $\prod_{i=1}^{n} p_i$, où $p_i = Enc_{pk}(x_i)$ si $y_i = 1$ et $p_i = 1$ si $y_i = 0$
ensuite utiliser la propriété additive du chiffrement homomorphe pour calculer $\prod_{i=1}^{n} p_i = Enc_{pk}(\overrightarrow{X} \bullet \overrightarrow{Y})$;
8: Envoi de $Enc_{pk}(\overrightarrow{X} \bullet \overrightarrow{Y})$ à Alice ;
9: **end for**
10: **for** Alice **do**
11: Calculer $Dec_{sk}(Enc_{pk}(\overrightarrow{X} \bullet \overrightarrow{Y}))$;
12: Envoi du résultat à Bob ;
13: **end for**

le cas contraire. Il utilise ensuite la propriété additive de l'homomorphisme pour obtenir $Enc_{pk}(\overrightarrow{X} \bullet \overrightarrow{Y})$. Nous rappelons qu'un cryptage à clé publique est homomorphe additif quand $\text{Enc}_{pk}(x_1, r_1) \times Enc_{pk}(x_2, r_2) \times ... \times Enc_{pk}(x_n, r_n) = Enc_{pk}(x_1 + x_2 + ... + x_n, r_1 \times r_2 \times ... \times r_n)$, où + et $\times$ sont des opérations de groupe. Pour simplifier la notation, nous ne noterons plus explicitement la valeur aléatoire comme entrée dans la fonction de cryptage. Dans l'étape 8, Bob envoie donc $Enc_{pk}(\overrightarrow{X} \bullet \overrightarrow{Y})$ à Alice. Possédant la clé secrète, Alice calcule $Dec_{sk}(Enc_{pk}(\overrightarrow{X} \bullet \overrightarrow{Y}))$ pour retrouver le résultat final qu'elle envoie à Bob.

4. Analyse du coût de communication et de calcul

Du point de vue communication, notre protocole a besoin des messages suivants : (i) pour chaque entrée de vecteur, notre protocole nécessite un message ; (ii) un message est requis pour envoyer la clé publique ; (iii) Bob a besoin d'envoyer $Enc_{pk}(\overrightarrow{X} \bullet \overrightarrow{Y})$ à Alice ; (iv) enfin, Alice doit envoyer le résultat du produit scalaire à Bob. Ainsi, le nombre de messages est $n + 3$, où n est la dimension du vecteur. Dans ce cas, nous obtenons une complexité, en terme de coût de communication, de l'ordre de $O(n)$. Du point de vue calcul, Alice exécute, dans notre protocole, n opérations de cryptage et 1 opération de décryptage. Bob exécute moins de $n - 1$ additions. Ceci donne une complexité linéaire, de l'ordre de $O(n)$.

5. Analyse de la sécurité

La sécurité de notre protocole est basée sur les propriétés du cryptosystème utilisé : (i) le système de cryptage de Paillier [13] est probabiliste ; (ii) le système de cryptage de Paillier [13] est un chiffrement homomorphe.

Pour analyser la sécurité, examinons les informations propagées par chaque site participant au protocole :

– Vue d'Alice : Dans les étapes 2, 3 et 4, pour chaque $x_i, i \in \{1..n\}$, Alice choisit un nombre aléatoire r et calcule $Enc_{pk}(x_i, r)$, puis l'envoie à Bob.

– Vue de Bob : Bob reçoit $Enc_{pk}(x_i, r)$. Ce message est indistinguable au calcul [13], i.e. qu'il n'est pas identifiable par étude statistique, puisque le cryptage est sémantiquement sécurisé. Donc pour chaque entrée $x_i, i \in \{1..n\}$, un adversaire ne peut reconnaître le chiffré correspondant. La probabilité de retrouver le vecteur d'Alice en entier est alors très faible, sachant que la dimension de celui-ci est généralement grande (notons que la technologie du datamining s'applique sur de larges volumes de données). Dans l'étape 7, Bob calcule $Enc_{pk}(\overrightarrow{X} \bullet \overrightarrow{Y}) = \prod Enc_{pk}(x_i)$ si $y_i = 1$. La propriété homomorphe du cryptosystème garantit que ce calcul ne révèlera pas les valeurs de $x_i, i \in \{1..n\}$. Aussi, dans l'étape 8, Bob envoie $Enc_{pk}(\overrightarrow{X} \bullet \overrightarrow{Y})$ à Alice. La sécurité de cette étape, de même que les étapes suivantes, n'est pas importante car le résultat du produit scalaire n'est pas privé.

6. Travaux relatifs

La préservation de la vie privée ne peut se faire et être efficace qu'à un certain coût. Ce coût est souvent comparé au coût dans DNSP (*Distributed Non-private Scalar Product*) qui est défini comme le calcul du produit scalaire de deux vecteurs sans contrainte de vie privée (i.e. une des parties envoie la totalité de ses données à l'autre partie qui calcule le produit scalaire et envoie le résultat au premier). Le coût de communication dans DNSP est de l'ordre de $n + 1$ messages. DNSP nécessite n opérations de multiplication et $n - 1$ opérations d'addition. Dans le tableau 1, nous comparons notre protocole avec DNSP et deux autres [9, 14] qui sont très populaires. Les détails sur l'évaluation de leur coût de communication et de calcul peuvent être retrouvés dans [14](section 3). Donc ce tableau résume la comparaison de notre protocole avec les trois autres.

L'analyse du tableau 1 met en évidence la différence entre notre protocole et les autres, suivant trois métriques : la complexité de calcul, le coût de communication et le surcoût de communication. Si la complexité de calcul n'est pas très critique en datamining c'est parcequ'il existe des algorithmes et des architectures parallèles permettant d'obtenir des complexités de calcul acceptables. Par contre, le nombre de messages échangés entre les

	DNSP	[9]	[14]	Notre protocole
Coût de communication	$n+1$	$3n/2+2$	$2n+2$	$n+3$
Surcoût de communication	0	$n/2+1$	$n+1$	2
Complexité de calcul	$O(n)$	$O(n^2)$	$O(n)$	$O(n)$

Figure 1. *Tableau comparatif entre protocoles de sécurisation.*

différents sites peut créer un goulet d'étranglement, qui pourrait considérablement dégrader les performances d'un processus de datamining. C'est donc un critère non négligeable dans la proposition d'un protocole de cryptage.

7. Conclusion

Le calcul sécurisé du produit scalaire est un important problème dans le domaine du datamining intégrant la contrainte de vie privée. Récemment, plusieurs protocoles ont été proposés pour résoudre ce problème. L'évaluation de ces protocoles montre qu'ils génèrent un important surcoût de communication. Dans ce papier, nous avons proposé un nouveau protocole de calcul sécurisé du produit scalaire basé sur des méthodes standards de cryptographie. Cette proposition présente deux avantages : (i) elle est sécurisée ; (ii) elle réduit considérablement le coût de communication entre les sites participant au calcul. La version actuelle de notre protocole ne s'applique que dans un contexte binaire, nous envisageons de l'étendre pour gérer des données catégoriques, puis numériques. Comme autre perspective, nous prévoyons de mesurer l'efficacité de notre protocole sur des bases de données (denses et éparses).

8. Bibliographie

[1] OLIVEIRA S. R.M., ZAIANE O., SAYGIN Y. : « Secure Association Rule Sharing », Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sidney, Australia, pp. 74-85, Springer Verlag.

[2] SAYGIN, Y., VERYKIOS, V., CLIFTON, C, « Using Unknowns to prevent discovery of Association Rules », ACM SIGMOD Record, 30(4), 2001.

[3] GUO, Y.H., TONG, Y.H., TANG, SW., YANG, DQ., « Knowledge hiding in database », Journal of Software, November 2007, 18(11), pp. 2782-2799.

[4] YAO, A. C., « Protocols for secure computations ». Proc. of the 23rd annual IEEE Symposium on Foundations of Computer Science, IEEE, London, pp. 160-164, 1982.

[5] YAO, A. C., « How to generate and exchange secrets ». Proc. of the 27th Symposium on Foundations of Computer Science (FOCS), IEEE, Toronto, pp. 162-167, 1986.

[6] DU, W., ATALLAH, M. J. « Secure multi-party computation problems and their applications : a review and open problems ». New Security Paradigms Workshop, Cloudcroft, USA, pp. 11-20, 2001.

[7] LINDELL, Y., PINKAS, B. « Privacy preserving data mining ». In Advances in Cryptology - CRYPTO 2000, pages 36-54. Springer-Verlag, August 20-24 2000.

[8] DWORK, C. , NISSIM, K., « Privacy-preserving Data mining on Vertically Partitioned Databases ». In Proceedings of The 24rd Annual International Cryptology Conference (CRYPTO 2004), Santa Barbara, CA, August 2004.

[9] VAIDYA, J. S., CLIFTON, C., « Privacy preserving association rule mining in vertically partitioned data ». In The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA pages 639-644, July 23-26 2002.

[10] KANTARCIOGLU, M. , CLIFTON, C., « Privacy-Preserving distributed mining of association rules on horizontally partitioned data ». IEEE Trans. Knowl. Data Eng. 16, 9 (Sept.) 2004, pp. 1026-1037.

[11] ZHAN, J. Z., MATWIN, S., CHANG,L., « Privacy-preserving collaborative sequential pattern mining ».In Proceedings of Workshop on Link Analysis, Counter-terrorism and Privacy, Florida, April 24 2004.

[12] AGRAWAL, R., SRIKANT, R.. « Fast algorithm for mining association rules in large databases ». In Research Report RJ 9839, IBM Almaden Research Center, San Jose, CA, June 1994.

[13] PAILLIER,P., « Public-key cryptosystems based on composite degree residuosity classes ». In EUROCRYPT, Vol. 1592 of Lecture Notes in Computer Science, Springer Verlag, 1999, pp. 223-238.

[14] AMIRBEKYAN, A., ESTIVILL-CASTRO, V., « A New Efficient Privacy-Preserving Scalar Product Protocol. » In Proc. Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia.