Suis-je celui que je prétends être ?

Diyé Dia, Olivier Coupelon, Yannick Loiseau, Olivier Raynaud

Suis-je celui que je prétends être ?

Diyé Dia[1,2], Olivier Coupelon[1], Yannick Loiseau[2], Olivier Raynaud[2]

[1] ALMERYS, solution santé d'Orange Business Services, Clermont-Ferrand, France
diye.dia, olivier.coupelon@almerys.com

[2] LIMOS, Informatique, Modélisation et Optimisation des Systèmes, Clermont-Ferrand, France
yannick.loiseau@univ-bpclermont.fr
raynaud@isima.fr

Résumé : L'usurpation d'identité est une fraude génératrice de grande méfiance des internautes envers l'utilisation des services numériques en ligne. La mise en place d'un système d'authentification implicite, c'est-à-dire basée sur l'étude du comportement de l'internaute, est un moyen original pour lutter contre cette fraude, pour restaurer la confiance et ainsi favoriser l'usage des services en lignes. Dans notre étude, l'authentification implicite se présente comme un élément de sécurité complémentaire aux éléments de sécurité traditionnels. Le rôle de notre système d'authentification est de détecter le plus tôt possible qu'un internaute n'est pas celui qu'il prétend être et/ou de valider le plus longtemps possible son identité. Plus précisément, ce système automatique peut être appelé à la demande - mode ponctuel - pour permettre l'accès à une fonctionnalité plus critique par exemple ou en mode continu pour élever le niveau de sécurité global de la plateforme d'accès. Le principe théorique du système consiste à générer des signatures pour chaque utilisateur à partir de l'historique de son comportement et de comparer ces signature à la trace locale pour valider ou non son identité.

Mots-clés : Comportement utilisateur, identité, authentification implicite, sécurité, confiance

1 Introduction

L'entreprise Almerys, porteuse du projet, déploie un espace de vie numérique composé d'un ensemble de services. Parmi ces service nous trouvons par exemple un coffre fort numérique, un accès à des communautés virtuelles et à des services d'e-commerce [1]. Pour accéder à cet espace et ainsi utiliser l'ensemble des fonctionnalités de la plateforme, un internaute doit se connecter avec un moyen d'authentification classique (login/mot de passe ou carte à puce/code PIN). Seulement, ce déploiement se fait dans un contexte de crise de confiance généralisée envers les systèmes en lignes. Cette crise étant en grande partie liée à la multiplication des vols d'identité sur Internet (120000 victimes d'usurpation d'identité par an en France [2]). Pour lutter contre ce type de fraude, il apparait nécessaire de consolider les systèmes de sécurité traditionnels mais sans détériorer le niveau d'usage des services et tout en respectant la vie privée des internautes. On appelle authentification implicite une authentification basée sur l'étude du comportement de l'internaute. A titre d'exemple, l'adresse IP d'une machine ou la géolocalisation d'un utilisateur peuvent être interprétées comme une forme de signature. Ainsi la mise en place d'un tel système, accepté par l'utilisateur et ensuite transparent pour lui, est un moyen original pour lutter contre les fraudes, pour restaurer la confiance et ainsi favoriser l'usage des services en lignes.
Dans notre étude, l'authentification implicite se présente comme un élément de sécurité complémentaire aux éléments de sécurité traditionnels. Le rôle de notre système est de détecter le plus tôt possible qu'un internaute n'est pas celui qu'il prétend être et/ou de valider le plus longtemps possible son identité. Ce système automatique peut être appelé à la demande - mode ponctuel -

1. https ://www.ebeeoffice.ca/ebee-home/public
2. http ://www.lepopulaire.fr/limousin/actualite/2013/02/25

pour permettre l'accès à une fonctionnalité plus critique par exemple ou en mode continu pour élever le niveau de sécurité global de la plateforme d'accès.

Dans la section suivante nous donnons un aperçu de l'état de l'art, et une approche à notre problème est décrite dans la section 3.

2 État de l'art

L'authentification implicite sur les téléphones mobiles a été étudiée par (Shi *et al.*, 2011) en se basant sur des caractéristiques propres aux smartphones tels que les appels, les sms, la navigation entre les applications du smartphone et la localisation. Les expériences qu'ils ont faites à partir des données de 50 utilisateurs sur 12 jours sont prometteurs malgré l'insuffisance des données. De même, l'étude de (Abramson & Aha, 2013) applique son modèle aux données de 10 utilisateurs sur 1 mois. Les auteurs utilisent des données de navigation web multi-sites pour faire de l'authentification. Ils se basent sur la date/heure des rêquetes et l'url visitée. Une étape de pré-traitement leur permet d'extraire d'autres types de données pour leur étude tels que le genre de la page visitée. Le travail de (Yang, 2010) permet de faire de l'identification en se basant sur des données de navigation web multi-sites également. Elle utilise comme éléments caractéristiques du comportement, le nom du site visité, le nombre de pages vues, l'heure de démarrage d'une session et la durée d'une session. Pour construire le profil de l'utilisateur, son comportement passé est étudié à l'aide de modèles comportementaux. La probabilité qu'il soit celui qu'il prétend être est calculé en comparant le comportement récent, lors d'une session par exemple, avec le profil de l'utilisateur. Les modèles comportementaux sont des modèles statistiques ou des modèles qui s'appuient sur des algorithmes de fouille de données. La classification bayésienne est utilisée dans (Ullah *et al.*, 2011) pour construire le profil des utilisateurs de streaming vidéo afin de prédire l'identité de ces derniers mais les résultats peuvent être améliorer. L'étude de (Abramson & Aha, 2013) utilise la machine à vecteurs de support de LibSVM sous le logiciel Weka pour construire le profil des utilisateurs. Cette étude conclut que les caractéristiques utilisés ne sont pas suffisants pour authentifier ou distinguer les utilisateurs. (Yang, 2010) utilise des calculs de fréquence d'apparition des itemsets construits par l'algorithme Apriori (Agrawal *et al.*, 1996). Sa méthode devient inefficace pour de très grands ensemble de données. Les 300 premières sessions de 2798 utilisateurs suivis sur une année constituent les données de (Yang, 2010). Il est donc nécessaire de bien choisir les éléments qui caractérisent le comportement ainsi que le modèle comportemental permettant de construire le profil.

3 Approche proposée

Une brique d'authentification implicite évalue le comportement de l'utilisateur dans notre plateforme de services numériques. L'authentification implicite peut être continue (l'utilisateur est authentifié à chaque requête demandée) ou ponctuelle (l'utilisateur est authentifié à chaque demande d'accès à une fonctionnalité critique). La criticité de chaque fonctionnalité est évaluée par un expert métier. Cet expert se base sur le niveau de sensibilité de la fonctionnalité. Plus une fonctionnalité est sensible, plus l'impact d'une attaque sur cette fonctionnalité est importante. Pour accéder à une fonctionnalité critique, la probabilité qu'il soit celui qu'il prétend être doit être maximum. Pour la première utilisation de l'authentification implicite, il sera nécessaire

de demander à l'utilisateur s"il souhaite activer le module d'authentification implicite afin de respecter les recommandations de la CNIL [3]. Nous construisons notre profil utilisateur à partir de données de connexion et de navigation sur une plateforme de services. Nous avons moins de données que si nous avions utilisé l'ensemble des données de navigation à travers le web, mais l'étude de (Shi *et al.*, 2011) montre que nous pouvons construire des profils prometteurs avec peu de données. Les fonctionnalités sont regroupées par service sur la plateforme de services. Les éléments qui caractérisent le comportement de l'utilisateur sont le nom de la fonctionnalité demandée, date/heure de la demande, le début de la session, la durée de la session, le service auquel appartient la fonctionnalité demandée et l'adresse IP (anonyme). Nous regroupons nos instances par utilisateur et par moment de la journée. Nous proposons trois moments de la journée : "Matin", "Après-midi" et "Soir". Dans notre base d'apprentissage, nous cherchons les itemsets fréquents avec l'algorithme Apriori pour chaque utilisateur et pour chaque moment de la journée. Un itemset de taille et de support maximal est considéré comme un profil. Un utilisateur a un profil pour chaque moment de la journée. Pour évaluer notre approche, nous construisons une matrice de confusion en utilisant notre base de test. La matrice est remplie en regardant si le profil est inclus dans les instances de la base de test. Nous testerons notre approche sur les logs de 30 utilisateurs sur 15 jours.

4 Conclusion

Nous sélectionnons des éléments caractérisant le comportement qui sont spécifiques à notre contexte. Notre modèle comportemental est intuitif et simple, par rapport aux modèles cités dans la littérature. Une comparaison de nos résultats avec ceux de (Yang, 2010) permettra d'évaluer la qualité de notre approche. Nous allons prendre en compte nos contraintes les plus importantes à savoir le respect de la vie privée des utilisateurs ainsi que la détection très rapide d'un imposteur.

Références

ABRAMSON M. & AHA D. W. (2013). User authentication fromweb browsing behavior. In *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference*, p. 268–273.

AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H. & VERKAMO A. (1996). Fast discovery of association rules. In *Advances in knowledge discovery and data mining 12 (1), AAAI Press*, p. 307–328.

SHI E., NIU Y., JAKOBSSON M. & CHOW R. (2011). Implicit authentication through learning user behavior. In *M. Burmester et al. (Eds.) : ISC 2010, LNCS 6531, Springer-Verlag Berlin Heidelberg*, p. 99–113.

STOCKINGER T. (2011). Implicit authentication on mobile devices. In *the Media Informatics Advanced Seminar on Ubiquitous Computing*.

ULLAH I., BONNET G., DOYEN G. & GAÏTI D. (2011). Un classifieur du comportement des utilisateurs dans les applications pair-à-pair de streaming vidéo. In *CFIP 2011 - Colloque Francophone sur l'Ingénierie des Protocoles*.

YANG Y. C. (2010). Web user behavioral profiling for user identification. In *Decision Support Systems, Elsevier*, number 49, p. 261–271.

3. http ://www.cnil.fr/