



Authentification de documents par la vérification des informations textuelles

Chloé Artaud

► To cite this version:

Chloé Artaud. Authentification de documents par la vérification des informations textuelles. Forum Jeune Chercheur INFORSID, May 2016, Grenoble, France. hal-01376891

HAL Id: hal-01376891

<https://hal.science/hal-01376891>

Submitted on 5 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Authentification de documents par la vérification des informations textuelles

Chloé Artaud

*L3I (Laboratoire Informatique, Image, Interaction) et DGA
Avenue Michel Crépeau
17042 La Rochelle cedex 1
chloe.artaud@univ-lr.fr*

MOTS-CLÉS : recherche d'information, extraction d'information, ingénierie des connaissances, faux documents

KEYWORDS: Information Retrieval, Information Extraction, Knowledge Engineering, Fraudulent Documents

ENCADREMENT : Antoine Doucet (PR) et Jean-Marc Ogier (PR)

1. Contexte

En 2015, 68% des entreprises françaises interrogées dans le cadre d'une étude de PricewaterhouseCoopers¹ déclarent avoir été victimes d'une fraude au cours des 24 derniers mois (changements de RIB, fraudes comptables, fraudes aux achats...). Dans un autre registre, les administrations publiques sont également victimes de fraudes à travers de fausses déclarations. La Sécurité sociale a ainsi détecté en 2013 un préjudice de 350,5 millions d'euros lié à la fraude aux prestations sociales². Les institutions ne sont pas les seules à être victimes de fraudes : les particuliers eux-mêmes peuvent être victimes d'arnaques. Lors d'un achat de voiture, certains fraudeurs peuvent par exemple falsifier la carte grise en changeant la date de la voiture, pour l'achat d'une maison, les diagnostics énergétiques...

Les documents frauduleux jouent souvent un rôle important dans les fraudes, quel que soit le domaine ou le type de document, et doivent donc être particulièrement contrôlés dans le cadre de la lutte contre les fraudes. Si des solutions existent pour sécuriser et contrôler les documents d'identité, les autres types de documents, qu'ils soient électroniques ou papier, souvent en format A4 et simplement imprimés, sont facilement modifiables. Le sujet de notre travail s'inscrit

¹ Global Economic Crime Survey 2016 "La fraude explose en France. La cybercriminalité au cœur de toutes les préoccupations", PwC

² <http://www.securite-sociale.fr/La-fraude-sociale>

donc dans ce contexte : il s'agit de détecter les documents faux ou falsifiés, notamment en vérifiant les informations qu'ils contiennent.

2. État de l'art

La détection des fraudes sur les documents est très peu étudiée, malgré des besoins importants. Nos recherches traversent donc plusieurs domaines, de l'analyse de document à l'ingénierie des connaissances en passant par l'analyse d'image, la recherche d'information et les *digital forensics* (que l'on pourrait traduire par « informatique légale », par analogie à « médecine légale »).

La plupart des travaux sur la détection des faux documents concernent des indices graphiques, comme la différence d'inclinaison, de taille, d'alignement ou de bruit d'un caractère par rapport aux autres (Bertrand *et al.*, 2013), la différence de police ou d'espacements des caractères au sein d'un mot (Bertrand *et al.*, 2015), le décalage d'une ligne par rapport aux marges (van Beusekom *et al.*, 2010) ou encore l'inclinaison différente des lignes les unes par rapport aux autres (van Beusekom *et al.*, 2009). Dans ces travaux, l'hypothèse de départ est que les fraudeurs doivent souvent modifier les éléments du document dans la précipitation et que la modification n'est donc pas toujours parfaite, ce qui permet de la détecter.

D'autres travaux en analyse d'image permettent de détecter ou de suspecter des modifications, comme une étude sur les imprimantes et scanners utilisés dans la fabrication du document (Elkasrawi *et al.*, 2014) ou la vérification des tampons (Micenková *et al.*, 2015). (Poisel *et al.*, 2011) dresse un état de l'art des recherches de fraudes sur les données multimédias, dont celles sur les images. Plus largement, de nombreux travaux existent sur les modifications des images de scènes naturelles.

Parallèlement à l'analyse d'image, nous nous intéressons aux informations contenues dans le document. Si les journalistes s'intéressent de plus en plus au *fact checking* (vérification des faits), il est difficile de trouver des travaux scientifiques sur la vérification d'information. Nous pouvons cependant citer (Goasdoué *et al.*, 2013) et leur outil FactMinder, à la croisée de l'extraction d'information, de la recherche d'information sur le Web et de l'ingénierie des connaissances. Nous pourrions également nous inspirer des travaux sur la détection de plagiat.

3. Problématique

Si la recherche sur les faux documents se concentre essentiellement sur des analyses graphiques, nous pensons qu'il est possible de lier leur détection à celle de faux contenus ou de contenus incohérents. En effet, le contenu d'un document est composé d'informations diverses, textuelles et graphiques, ayant toutes un sens. Ces informations sont liées sémantiquement les unes aux autres et peuvent être vérifiées, seules et collectivement. Par exemple, l'adresse d'une entreprise est liée à son nom, et souvent à un numéro de téléphone, un numéro de SIRET...

Aujourd'hui, lorsqu'une personne veut vérifier une information, son premier réflexe est de faire une recherche sur Internet, à l'aide de moteurs de recherche. Généralement, lorsque les mots de la requête sont bien choisis, la page de résultats (SERP, pour *Search Engine Results Page*) donne une idée de la véracité de l'information. Sinon, il faut explorer les liens ou reformuler la requête. Nous envisageons d'automatiser complètement ce processus afin de le simplifier et de l'intégrer à notre système de détection des faux documents. Nous pensons qu'il serait intéressant de créer un système capable d'extraire automatiquement les informations du document et de les comparer avec l'information accessible en ligne.

4. Actions réalisées

Nous avons élaboré le plan d'un processus qui permettrait de vérifier les informations simples. Il s'agit tout d'abord d'extraire les informations du document et de les annoter afin d'en peupler une ontologie. Nous pouvons pour cela créer des règles d'extraction pour chaque information récurrente des documents. Cette extraction peut se faire avec des outils comme Unitex.

Une fois les informations extraites et identifiées, il s'agit d'en faire des requêtes pour un moteur de recherche. Les requêtes peuvent être composées d'une seule information, qui correspondra alors à une instance de l'ontologie (comme « 12345678901234 », instance de la classe « numéro SIRET »), ou de plusieurs mots combinant instances et classes (comme « numéro SIRET Exentrep » qui associe la classe « numéro SIRET » et l'instance « Exentrep » de la classe « nom société »). Nous avons choisi d'utiliser une ontologie pour pouvoir justement associer des informations grâce aux propriétés.

Ces requêtes permettent de récupérer une SERP dont on peut extraire les informations, en utilisant les mêmes règles d'extraction que précédemment. Il faut ensuite vérifier que les informations extraites sont identiques en les normalisant. Si elles sont identiques, elles deviennent une information unique que nous pouvons comparer avec celle du document à vérifier. S'il y a plusieurs résultats différents, il faudra déterminer lequel est le plus fiable. On peut vérifier les différents résultats en les requêtant à leur tour et en analysant les résultats de la recherche. Nous pouvons ensuite comparer les informations du document et celles de la SERP en utilisant des méthodes de calcul de similarité et affecter un indice de fiabilité à chaque information.

5. Actions futures

Le processus proposé se fera en deux phases : nous chercherons d'abord à évaluer chaque étape en les réalisant manuellement, puis nous automatiseront intégralement le processus, si la première phase se révèle concluante. Il s'agira donc à terme d'automatiser le processus tout en laissant à l'utilisateur la possibilité de le superviser et d'intervenir pour corriger d'éventuelles erreurs. Il faudra également

évaluer les performances de ce système en comparant les résultats avec ceux obtenus manuellement.

Nous pensons également explorer les liens proposés dans la SERP : nous pouvons pour des informations plus complexes avoir besoin de fouiller le web plus en profondeur. Ainsi, nous pouvons imaginer un système qui tenterait d'extraire les informations sur chaque page explorée et continuer à explorer les liens tant qu'il n'a pas trouvé suffisamment d'informations pertinentes.

Dans un dernier temps, nous chercherons à associer ces indices sémantiques aux indices graphiques décrits dans l'état de l'art pour réaliser un système de détection des fraudes complet où l'utilisateur n'aura qu'à entrer le document à vérifier pour que le système lui donne un taux de fiabilité pour ce document.

6. Bibliographie

- Bertrand, R., Gomez-Kramer, P., Terrades, O. R., Franco, P., Ogier, J. M., “ A system based on intrinsic features for fraudulent document detection ”, *International Conference on Document Analysis and Recognition (ICDAR)*, 2013, p. 106-110.
- Bertrand, R., Terrades, O. R., Gomez-Kramer, P., Franco, P., Ogier, J.-M., “ A Conditional Random Field model for font forgery detection ”, *13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, p. 576-580.
- Elkasrawi, S., Shafait, F., “ Printer Identification Using Supervised Learning for Document Forgery Detection ”, *11th IAPR International Workshop on Document Analysis Systems*, 2014, p. 146-150.
- Goasdoué, F., Karanasos, K., Katsis, Y., Leblay, J., Manolescu, I., Zampetakis, S., “ Fact checking and analyzing the web ”, *International Conference on Management of Data (SIGMOD '13)*, 2013, p. 997-1000.
- Micenkova, B., van Beusekom, J., Shafait, F., “ Stamp verification for automated document authentication ”, *Computational Forensics*, 2015, 8915, p. 117-129.
- Poisel, R., Tjoa, S., “ Forensics investigations of multimedia data: A review of the state-of-the-art ”, *6th International Conference on IT Security Incident Management and IT Forensics (IMF 2011)*, 2011, p. 48-61.
- van Beusekom, J., Shafait, F., Breuel, T., “ Automatic Line Orientation Measurement for Questioned Document Examination ”, *Computational Forensics*, 2009, 5718, p. 165-173.
- van Beusekom, J., Shafait, F., Breuel, T. M., “ Document inspection using text-line alignment ”, *9th IAPR International Workshop on Document Analysis Systems*, 2010, p. 263-270.