The Age of Confidentiality: A Review of the Security in Social Networks and Internet

Antonio Juan Sánchez[1] and Yves Demazeau[2]

[1] Universidad de Salamanca, Salamanca, Spain
anto@usal.es
[2] Laboratoire d'Informatique de Grenoble - CNRS, Grenoble, France
Yves.Demazeau@imag.fr

Abstract. Security based on content analysis in social networks has become a hot spot as a result of the recent problems of violations of privacy by governments to international security agencies. This article is an approach to the implementation of programs for extraction and analysis of the web information, this includes the technical difficulties and the legal consequences involved.

Keywords: Social networks analysis, security, data meaning, information fusion, wrappers, protection data laws, organizations, agents.

1 Introduction

Today, security has become a fundamental point in the information era. Just as all eyes are on this critical point and it affects all levels from the individual human being to international corporations. The cyber-crime has grown exponentially since the massive use of Internet and mobile devices, always protected by the anonymity of the criminals. The new technologies and their ability to analyse large amounts of information (Big Data) have made a breakthrough, the anonymity of aggressors is no longer such. Every step we take is reflected on the Internet in different ways, through different databases scattered around the globe, is what has been called digital fingerprint.

Every adverse comment we left on Facebook, every tweet that may contain traces of a past or future crime, every comment in a blog that may damage the self-image, all of them can be now analysed and linked with our personal data. These data can be extracted from the same sources where can be found documents with sensitive content (photos, videos) or information about our life and work (LinkedIn, Google+, telephone records on-line yellow pages extracted as, listings in public entities, etc.). In most cases, user data from mass use websites are data that can be accessed freely. Throughout this article the different strategies and tools that can be used to get personal content on Internet will be discussed as well as the legal issues involved, and finally a platform that supported all this strategies will be proposed.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_47,

2 Extracting Information from the Web

The semantic web has tried to simplify the process of extracting information, but the fact is that today the semantic capacity has not yet been implemented in most of the websites.

Scraping tools (download web content most used) today are Wget[1], cURL[2] or Pillow[3], although there are numerous libraries for most languages as Scrapy[4] and Beautiful Shop[5] (both for Phyton), Html Agility Pack (.NET)[6], JSoup (Java)[7], htmlSQL (PHP)[8]. Beautiful Shop is the most complete of all bookstores because it supports even malformed HTML.

The crude extract of the contents of the web is the most developed stage of the analysis and involves minor problem. The difficulty increases when we talk about information processing to get useful information obtained.

There is no doubt that the nature of the information obtained depends on the type of paper used, but also we must consider that in each document can be found associated information does not correspond to the nature of this: for example, a photograph is mainly graphical information, but the metadata of the image file may also contain information about where it was made (many GPS mobile devices leave a mark) or the date and time it was made. In addition, passing an OCR to the picture, we can find texts, which could then be processable by a computer.

There are many different types of tools to extract information from the obtained data. To process the text got on web pages, first of all it's necessary to select and then label content that may be useful to us. Wrappers are used to select [2][3] (data mining programs specialized in information extraction). There are five different approaches to generate wrappers [6]:

- **Wrapper induction** need specific rules. These rules can be determined by a user or by a training-induced. In both cases supervised algorithms [5]. Within this category are the NLP (Natural Language Processing) [7].
- **Model based Wrappers** operate inversely to wrapper induction. They try to find underlying structures within the web content to serve as a model to extract data [8].

[1] GNU Wget. `http://www.gnu.org/software/wget/`
[2] cURL. Command line tool for transferring data with URL syntax. `http://curl.haxx.se`
[3] Pillow. Tool for Phyton for transferring data with URL syntax `https://pypi.python.org/pypi/Pillow/`
[4] Scrapy. Screen scraping and web crawling framework `http://scrapy.org/`
[5] Beautiful Soup. Scraping framework for Python. `http://www.crummy.com/software/BeautifulSoup/`
[6] Html Agility Pack L. Scraping framework for PHP. `https://github.com/hxseven/htmlSQL`
[7] JSopup. Scraping framework for Java. `https://github.com/hxseven/htmlSQL`
[8] htmlSQL. Scraping framework for PHP. `https://github.com/hxseven/htmlSQL`

- **The self-extracting wrappers** select relevant content using regular expressions and dictionaries [4]. Among the most comprehensive tools in this section we find Camaleon [20] that supports HTML, Javascript and cookies.
- **The ontology based** seek constant within the content and build objects around them. The most important work in this aspect has been made in the Brigham Young University [23] and it's based on a RDF extraction system [21] and DAML ontology[9].
- **Automatic design languages** are definition languages wrappers, generally allow the definition of grammars through EBNF or regular expressions with help of the typical exceptions to procedural languages. Notable in this kind of languages: Minerva, Web-OQL and WebL.
- **The HTML-based generators** are strictly based on the knowledge of how this language tag to select the important parts. Among the most important tools are W4F, XWRAP, Road Runner or LIXTO.

The optimum solution for generate wrappers would be a mixed technique of Wrappers based in Model and Ontology, generating data structures based in the objects found in the ontology.

The most widely model used to classify information properly is an ontology. Ontologies can generate a vocabulary with the data, specify the meaning of terms and formalize the relationships between them using a suitable logic [9]. Some wrappers facilitate the process of step to ontology, by providing the ontology formed.

The fundamental difficulty of step to ontology comes from the scalability, and it is essential if we think that the extraction and storage of the web information is taken continuously once the process is started. The higher an ontology becomes, the greater the difficulty in applying an algorithm for reasoning about it, is a problem of type 2NEXPTIME-complete [9]. The proposed solutions to this problem agree on the division of the ontology groups, through a clustering algorithm [25], which always grow at much less speed than the set [10].

3 Is It Legal to Extract Data from Internet for Purposes Related to Security?

The data protection laws and copyright is an important point to deal since they mark the limit of the data that may be extracted from Internet. In Spain the legislative rules that regulate are “Ley 15/1999 de Protección de Datos Personales” or LOPD, “Ley 34/2002 de Servicios de la Sociedad de la Información y el Comercio Electrónico” (LSSICE), and “Ley 56/2007 de Medidas de Impulso de la Sociedad de la Información”.

Many websites, such as Twitter, Facebook o Google+, require to leave public information from the information extractor (usually through a registration in the web developer) and require that the information obtained can’t be used outside of the social network. On the other hand, the screen scraping is approved in Spain as legal

[9] DAWL-OIL. OWL. Web Ontology Language
http://www.w3.org/2001/sw/wiki/OWL

technique if it is done under certain assumptions [11], ie, the extraction of sources such as blogs, newspapers or news pages is only protected by copy-laws of intellectual character.

In Spain, Ley 25.326 protects the data of citizens, and their possibility to freely disclose their personal data. Transfers of data to companies to obtain services usually involve the protection of these data. Furthermore, according to the data protection law 15/1999, the personal data collected -with consent- of a certain website may be used only for purposes directly related to the functions of the transferor and transferee and must be a connection between the activities and purposes of the communication of the transferor and transferee companies.

Thus, initially, the extraction of data from social network users for studies or statistics outside the network is prohibited by law, but according to the Spanish Law of Criminal Procedure 579, a judge may grant the interception of communications in cases of national security, avoiding also Article 8 of the European Convention for the protection of Human Rights and Fundamental Freedoms affecting privacy.

4 Data Analysis

4.1 How to Analyze the Data Obtained

Each of the sources available on the Internet gives us 5 different perspectives fingerprint left by web users. For your personal data is necessary to merge information from different sources, such as social networking, government listings, company records, state news bulletins or telephone directories, and delete those erroneous data.

From the comments in conventional social networks can be drawn about the personality characteristics of the subjects: hobbies, interests, and political, sexual and moral preferences.

Of employment-oriented social networks like LinkedIn or Infojobs can get your professional profile, professional curriculum and academic data.

Merging data from their networks of professional and personal contact can get your closest contacts. If we consider these networks of relationships between people are graphs, you can apply all the related theory and small-world networks to establish communities and circles of trust. There are many practical examples of this, to highlight Anheier, Gerhards and Romo [12] Wouter De Nooy [13] and Burgert Senekal [14], based on the theories of Ronald S. Burt [15].

Through photographs and multimedia documents can plot profile of the subject and get a map and a calendar of your visited websites. The retrieval of this information is not too problematic, there are specialized tools and techniques in it. To extract information automatically from the metadata multimedia files can be used as Tika libraries[10], MediaInfo[11] or Exif Tags Collection[12]. To extract information from sound

[10] Tika Project. `http://tika.apache.org/`
[11] MediaInfo Project. `http://mediainfo.sourceforge.net`
[12] Exif Tags Collection Project. `https://www.codeproject.com/script/Members`

content type and pass it to text would require other libraries like Microsoft Speech API for .NET[13], Cmu Sphinx[14] for Java o Julius[15] , Kalidi[16], Simon[17], iATROS[18] for C. To extract textual information from videos and photos there is a wide range of OCR type solutions both free and paid license.

Crowd of on-line applications are dedicated today to make this information collection and create comprehensive profiles of users of social networks, so far these profiles have always responded to public properties and / or objective of subjects. The greatest interest in security and privacy do not provide these features, but those more personal that allow make a profile more focused on the psychological and personal.

4.2 Data Groupings

As discussed above managing the wealth of information that involves extracting data from the web takes to divide it to make it operatively computable.

The partition can be accomplished by different techniques at different times of the process. The nature of the data makes these may be divided at least when perform design ontologies and networking. For the division of space relations of the individual, detection techniques of communities are commonly used [16], while for the ontologies, besides the graph analysis procedures, cluster analysis (either hierarchical or partition) can be applied.

4.3 Extracting Results

After the whole process of clustering of terms in the ontology and the distinction of communities only would analyse the "psychological" data of individuals, ie verifying that the texts employees doing their comments or reviews not show signs of offense (whether physical or virtual).

In this case, special ontologies will be required very tight to the terminologies searched cybercrimes, since the semantics used is substantially different from the ordinary natural language.

These ontologies can be created from the application of various supervised algorithms on texts.

The self-learning system for detecting future crimes or from changes in ontologies, is based on anomaly detection in texts. In David Guthire studies [18] about significant deviations from the context different detection methods can be found.

---

[13] Microsoft Speech API `http://www.microsoft.com/downloads/details.aspx?FamilyID=5e86ec97-40a7-453f-b0ee-6583171b4530`

[14] Cmu Sphinx. `http://cmusphinx.sourceforge.net/sphinx4/`

[15] Julius in Sourceforge. `http://julius.sourceforge.jp`

[16] Kalidi in Sourceforge `http://kaldi.sourceforge.net`

[17] Simon. `http://simon-listens.blogspot.com/`

[18] iATROS. `http://prhlt.iti.es/software/iatros/doc/speech/`

Moreover, another problem is the detection of fake profiles. Studies of the National Institute of Technology of Roukela [19] prove that by techniques of SVM (Support Vector Machine) 95% of the fake profiles can be detected taking into account the following attributes of each profile: number of friends, education / current job, amount of text on own information, marital status, number of images, number of comments in other people's profiles, percentage among friends of the same gender and friends in total, percentage between applications of "friendship" sent and received, number of groups to which it belongs and the number of "I like you" (or similar).

5 Conclusions

5.1 Case Studies of Implementation of Cybersecurity Systems

Both companies and public entities have begun in recent years to promote cybersecurity projects. Large intelligence agencies internationally as CIA, Interpol, DEA or FBI have their own software but for safety information they are inaccessible.

Other projects due to its link with research teams are a little more open to analysis.

Riot is a project of the U.S. defense company Raytheon, for monitoring individuals from storage and fusion of data on social networks like Facebook, Twitter or FourSquare. Riot uses the contents of comments and the information contained in the metadata of media relations for plotting graphs, maps the movement of individuals, and predict future movements.

In August 2012, the University of Sydney released version 2.0 of the tool GEOMI (maximum penetration geometry) for data visualization. GEOMI allows police and security agencies to visualize and analyze complex relationships in social networks, email and phone records [22].

5.2 Proposal for a System with Agents

Automated search of crime through the content analysis in online social environments still involves technical difficulties that will soon be overcome:

- Detect false profiles
- Find a optimal rime dictionary
- Automated the information extraction from different systems
- Simulated future scenarios of crime

As has been shown, some government cybersecurity programs in different countries already operate with reduced automation. The main problem of application software for the cyberdetection of crimes is the maintenance required, since it is very difficult to ensure effective probability of success in generating new dictionaries criminal terms.

In any case, the results should never be understood as significant evidence of a crime but cyberdetection software should be better understood as an example of decision support system.

As the process of collection, storage and analysis of information is complicated, to work in the future an approach with a society of distributed agents [24] is proposed (see Figure 1).

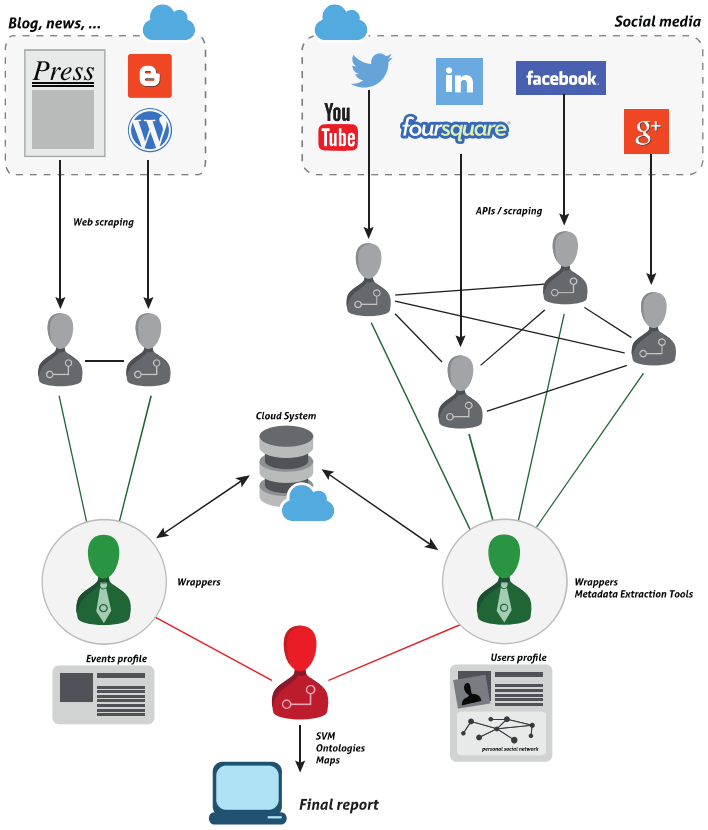

Fig. 1. Extraction and analysis process with agents

Acknowledgement. This work has been partially supported by the MINETUR project PIAR, TSI-100201-2013-20.

References

1. W3C Semantic Web Activity. World Wide Web Consortium (W3C) (November 7, 2011) (retrieved November 26, 2011)
2. Kushmerick, N., Weld, D.S., Doorenbos, R.: Wrapper Induction for Information Extraction. In: Proceedings of the International Joint Conference on Artificial Intelligence (1997)

3. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer (2007)
4. Dalvi, N., Kumar, R., Soliman, M.: Automatic wrappers for large scale web extraction. Proceedings of the VLDB Endowment 4(4), 219–230 (2011)
5. Kushmerick, N., Weld, D.S., Doorenbos, R.: Wrapper Induction for Information Extraction. In: Proceedings of the International Joint Conference on Artificial Intelligence (1997)
6. Laender, A., Ribeiro-Neto, B., Silva, A., Teixeira, J.: A Brief Survey of Web Data Extraction Tools. SIGMOD Record 31(2) (June 2002)
7. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009) ISBN 978-0-596-51649-9
8. Alderberg, B.: NoDoSe – A Tool for semi-automatically extracting structured and semistructured data from text-documents. In: Proceddings of ACM SIGMON International Conference on Management of Data, Seattle
9. Glimm, B., Horrocks, I., Motik, B., Shearer, R., Stoilos, G.: A Novel Approach to Ontology Classification. J. of Web Semantics 14, 84–101 (2012)
10. Stuckenschmidt, H., Klein, M.: Structure-based partitioning of large concept hierarchies. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 289–303. Springer, Heidelberg (2004)
11. Sentencia del Tribunal Supremo de España 9 de Octubre de, número 572/2012 (2012)
12. Anheier, H.K., Gerhards, J., Romo, F.P.: Forms of capital and social structure of fields: Examining Bourdieu's social topography. American Journal of Sociology 100, 859–903 (1995)
13. De Nooy, W.: Fields and networks: Correspondence analysis and social network analysis in the framework of Field Theory. Poetics 31, 305–327 (2003)
14. Senekal, B.A.: Die Afrikaanse literêre sisteem: 'n Eksperimentele benadeing met behulp van Sosiale-netwerk-analise (SNA), LitNet Akademies (2012)
15. Burt, S.R.: Structural Holes: The Social Structure of Competition. Havard University Press (1992)
16. Coskun, G., Rothe, M., Teymourian, K., Paschke, A.: Applying Community Detection Algorithms on Ontologies for Identifying Concept Groups. In: Proc. of the Fifth International Workshop on Modular Ontologies (WoMO 2011), pp. 12–24 (2011)
17. Leskovec, J., Lang, K.J., Mahoney, M.W.: Empirical Comparison of Algorithms for Network Community Detection. In: WWW 2010: ACM WWW International Conference on World Wide Web (2010)
18. Guthrie, D.: Unsupervised Detection of Anomalous Text. Department of Computer Science University of Sheeld (July 2008), http://nlp.shef.ac.uk/Completed_PhD_Projects/guthrie.pdf
19. Barrientos, F., Ríos, S.A.: Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicaciones. Revista de Ingeniería de Sistemas 27 (Septiembre 2013)
20. The Camaleon Web Wrapper Engine, http://chameleon.readthedocs.org/en/latest
21. RDF. Semantic Web Standards, http://www.w3.org/RDF/
22. Ahmed, A., et al.: GEOMI: GEOmetry for Maximum Insight. In: Healy, P., Nikolov, N.S. (eds.) GD 2005. LNCS, vol. 3843, pp. 468–479. Springer, Heidelberg (2006), http://sydney.edu.au/engineering/it/~cmurray/geomi.pdf

23. Liddle, S.W., Embley, D.W., Scott, D.T., Yau, S.H.: Extracting Data Behind Web Form. In: Olivé, À., Yoshikawa, M., Yu, E.S.K. (eds.) ER 2003. LNCS, vol. 2784, pp. 402–413. Springer, Heidelberg (2003)
24. Garijo, F., Gómes-Sanz, J.J., Pavón, J., Massonet, P.: Multi-agent system organization: An engineering perspective. In: Pre-Proceeding of the 10th European Workshop on Modeling Autonomous Agents in a Multi-Agent World, MAAMAW 2001 (2001)
25. Castellanos-Garzón, J., García, C., Novais, P., Díaz, F.: A visual analytics framework for cluster analysis of DNA microarray data. Expert Systems With Applications 40(2), 758–774 (2013) ISSN: 0957-4174