Natural Language Processing of Rules and Regulations for Compliance in the Cloud

Nick Papanikolaou

Cloud and Security Lab
HP Labs, Bristol, UK
nick.papanikolaou@hp.com

Abstract. We discuss ongoing work on developing tools and techniques for understanding natural-language descriptions of security and privacy rules, particularly in the context of cloud computing services. In particular, we present a three-part toolkit for analyzing and processing texts, and enforcing privacy and security rules extracted from those texts. We are interested in developing efficient, accurate technologies to reduce the time spent analyzing and reasoning about new privacy laws and security rules within the enterprise. We describe the tools we have developed for semantic annotation, and also for information extraction - these are specifically intended for analysis of cloud terms of service, and therefore designed to help with self-compliance; however, the techniques involved should be generalizable to other relevant texts, esp. rules and regulations for data protection.

1 Introduction

Cloud service providers compete on many fronts, with offerings that differ in terms of service availability, cost, security capabilities, flexibility and control over the location of data and virtual machines, to name a few. Security concerns remain the #1 issue for CIOs and CISOs responsible for deciding to purchase a cloud service, and there is much to be done on this front in order to satisfy the needs of enterprise customers. For individual customers, cloud services must be designed so as to offer privacy protection; the consequences of privacy breaches are hard to ignore, and include obligations to pay hefty fines, while also leading to loss of customer trust.

The core problem that needs to be addressed is how to keep up with changes in the legislation, regulators' rules regarding customer data, industrial standards (cf. the standards set forth by the Cloud Security Alliance, to take an example), and corporate guidelines, while minimizing the overall cost, effort and manpower required.

The way a cloud service provider handles its customers' data is usually specified in a written contract or agreement which comprises the ToS (Terms of Service) and SLA (Service Level Agreement). No commonly accepted standard exists for the format or content of cloud ToS and SLAs, nor any consensus about the expected security and privacy practices of service providers.

adfa, p. 1, 2011.

This poses difficulties for customers and providers alike, who have expectations (and duties) with regards to a given service offering. End-users require clarity and understanding on issues such as:

- how long a provider keeps data which has been stored or exchanged through its cloud services;
- how and when such data is destroyed;
- what remediation procedure exists in case of data loss and in case of data breach,
- to what extent data will be shared with parties external to the service provider and for what purpose (e.g. targeted advertising).

Enterprise customers typically require assurances regarding:

- service availability (e.g. estimated downtime per calendar month);
- cost of basic services versus added-value offerings;
- how data stored by a provider is kept isolated from other customers' data (particularly for multi-tenancy arrangements);
- encryption methods used, if any, and authentication technologies;
- backup methods and regularity of backup;
- remediation procedures and compensation offered in cases of data loss and data breach.

Although the field of cloud computing still lacks well-defined standards and best practices, they are actively being developed, and it is likely that cloud service providers will have a business need to adopt them in the future. This introduces another level of compliance and, unless cloud service providers are equipped with appropriate controls and tools, much manual effort may be required to achieve it.

There is also a need for tools that ensure what we might call **self-compliance**, namely compliance of a cloud service provider with its own stated policies. To date there is no obvious way of ensuring that the Terms of Service stated by cloud service providers are actually adhered to fully in practice.

This paper presents ongoing work on developing software tools to automate compliance in the cloud, particularly natural-language processing of cloud terms of service.

2 Previous and Related Work

In previous work the authors have developed technical mechanisms for controlling the flow of data in an IT infrastructure, notably through the use of privacy controls [7], sticky policies [8], and policy enforcement [6]. Although the cited works do not specifically focus on cloud computing scenarios, we expect these techniques to be readily extendable and adaptable to suit the needs of a cloud service provider.

- Comparison of policies and decision support
- Automated enforcement of security and privacy rules

Related work in the context of website privacy policies includes May and Gunter's formalism of policy relations, which are formal relationships defined over the intend-

ed semantics (or the authors' interpretation thereof) of P3P [5]. In a previous paper [6] we developed a mapping from P3P to CSP, enabling direct comparison of privacy policies using the model-checker FDR.

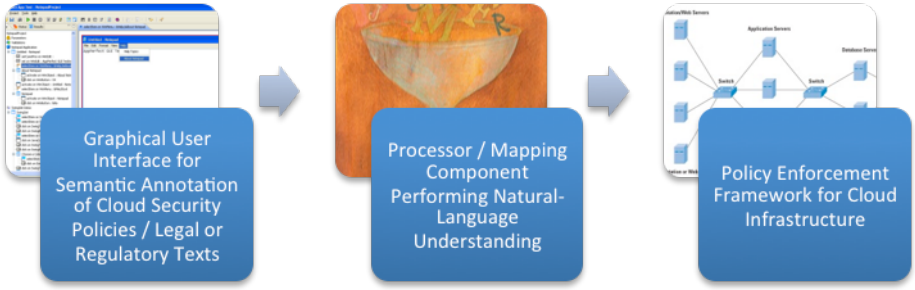
The EnCoRe research project is developing a platform for expressing and enforcing privacy preferences for personal data; recent case studies include a system for managing data held within an enterprise's HR systems, and health data stored about individuals and tissue samples in a biobank. Through the use of a suitable policy enforcement architecture, legal and regulatory privacy rules, along with individuals' privacy preferences, can be automatically enforced so that unauthorized and/or unsuitable access to data is prevented. In [7] we proposed a simple conceptual model for representing privacy rules, which can be directly mapped to technically enforceable access control policies (expressed e.g. using XACML).

3 A Toolkit Enabling Cloud Service Providers to Automate Compliance

We are developing a toolkit to automate more of the compliance processes in a cloud business than is currently the case. In particular, rather than just developing automated methods for monitoring and tracking data in the cloud, we should be developing natural-language processing tools to analyze and extract information from legal and regulatory texts automatically. We are not interested in doing this for arbitrary texts, as this largely unreasonable; our work focuses on texts that specifically describe policies – rules, prohibitions, necessary measures that need to be put in place to provide assurance in cloud computing. In some cases, the texts to be analyzed are large pieces of legalese; this would include, for instance, national data protection legislation in Europe, or the Health Insurance and Portability and Accountability Act in USA. Other examples of texts that our techniques could be applied to include the Sarbanes-Oxley Act (SOX), the Consumer Data Security and Notification Act, or the Gramm-Leach-Bliley Act. In the context of cloud computing, there are cloud security standards, as well as competitors' Terms of Service that are particularly worthy of analysis and comparison to our own.

Natural-language processing techniques have been developed for many years in academia and industry; our contribution is the development of a framework and toolset for integrating some of these techniques and applying them to a real-world problem and a central business need. We do not believe HP's competitors currently have the technology to readily adapt and respond to changes in security and privacy requirements for cloud services. Through investment in this research, we believe HP can gain significant competitive advantage and provide service offerings that far exceed the expectations of customers in terms of data protection.

Our solution comprises three parts, depicted in the figure below. We describe the functions of each component in turn.



Graphical User Interface for Semantic Annotation: We are currently developing a visual editor to enable domain experts in security and privacy to perform semantic annotation of texts containing rules that must be applied within a cloud infrastructure. Human experts are needed for this stage, even though the central objective of this work as a whole is to automate the analysis of texts as much as possible; the experts will use the editor to highlight and mark up portions of texts that are to be translated to machine-readable rules. This is the only stage where human interaction is required, and is essential in order to signpost the different parts of a rule (namely, the actor or subject of a rule, the action taken by the actor/obligation of the actor/restriction on the actor, the object of the rule, and the exceptions that apply).

It is important to note that the use of the editor is intended to attach semantic information to parts of the legal/ regulatory texts/cloud security policies in question, in a form that is understandable by the processor component, described next. The natural language processing algorithms that the processor implements handle most aspects of syntax, but assume that semantic information is also available.

Processor/Mapping Component: This component performs the 'natural-language understanding' and may be considered the 'intelligent' part of our solution. The automation argued for earlier in this paper is largely due to the functions of this component. The processor analyses source texts which contain security and privacy rules (assumed to be in ordinary English), detects patterns in the use of language that describe typical security features of cloud services, and extracts from the text entities and relationships between them (the cloud service provider, third parties, components of infrastructure, mechanisms, practices that are described in the texts). We consider developing a first version of this component by manually building a database of concepts and relationships that appear in cloud service providers' terms of service, but envisage the eventual use of machine learning algorithms to make an adaptive, self-modifying tool.

Policy Enforcement Framework: The final element of our solution is a policy enforcement framework, namely a system of Policy Decision and Policy Enforcement Points that can are deployed within a cloud service providers' infrastructure in order to implement the policy rules produced by the processor component above. Notice that the policies used in this framework are low-level, machine-readable policies expressed in a language such as XACML. The idea is that these policies will directly

implement the rules coming out of the legal, regulatory and other texts that have been passed through the above components. Furthermore, note that there will be a significant number of security rules, dictated for example by the law or by cloud security standards, which map directly to simple access-control policies that can be directly enforced in the cloud infrastructure. Such rules, which merely consist of restrictions or authorizations to access particular components or data within an infrastructure, are the simplest to enforce in practice, and for these the benefits of cost reduction and increase in efficiency due to automation will be most apparent. There has been a significant amount of previous work on policy enforcement in various past projects at HP Labs, including projects PRIME and EnCoRe, and we argue that we can build on and extend this work to the cloud computing context.

In the next section, we describe the tools we have developed for semantic annotation, and also for information extraction, namely, for the processor/mapping component described above. The tools we have developed are specifically intended for analysis of cloud terms of service, and therefore designed to help with self-compliance; however, the techniques involved should be generalizable to other relevant texts, esp. rules and regulations for data protection.

4 Implementation: A Tool for Semantic Annotation and Information Extraction for Cloud Terms of Service

The figure above presents our current model for analysing cloud computing terms of service. We are developing a tool for marking up and extracting information from cloud terms of service, namely, the contract documents that describe a customer's relationship with a cloud service provider. Our tool is not fully automated as it requires, as a first step, a human user to indicate which sections of such documents describe which types of rules; this process is referred to as semantic annotation. Our tool provides a text editor with functions to highlight portions of text that describe restrictions, obligations, and other types of constraint with a particular colour. Output from the tool includes a marked-up version of the original contract, with semantic tags. This output can then be fed into a separate processor, which is work in progress, whose functions include information extraction and rule generation.

Key characteristics of cloud Terms of Service include:

- Cloud ToS are almost always formatted as rich-text web documents with headings and numbered paragraphs ("clauses" – in the legal sense, not the grammatical sense of the word).
- Significant portions of these texts contain disclaimers, enabling the service provider to refuse being held accountable in certain cases (these parts of the ToS actually state what the provider will not be expected to do, rather than what the provider's actual practices are).

- If a service provider has several similar offerings (e.g. in the case of AWS) there will typically be two documents of interest – (i) a core agreement which sets out the main terms of service, and (ii) separate ToS for each of the different offerings (e.g. in the case of AWS offerings include: EC2, S3, EBS, SQS, SNS, SES, VPC, FWS, SimpleDB, GovCloud).

A recent legal research paper [9] documented the features and caveats of different cloud service level agreements, including discussions of both the general service descriptions and the terms and conditions available online.

While a cloud service provider may employ legal experts to draw up their terms and conditions in writing, it is the developers and system administrators that are responsible for making sure these terms are indeed enforced in the infrastructure used for a particular cloud offering. It is in the interest of the latter to have machine readable rules that are in a one-to-one correspondence with the statements made in the written ToS.

Natural-language analysis of the written ToS can certainly assist in the creation of such rules; if the written style of an ToS is very prescriptive, enforceable rules are easier to generate automatically. Otherwise human intervention will be required to ensure that generated rules are:

- **correct:** namely, that they express what actions a system needs to implement to make sure the requirements of the ToS are fulfilled on a constant basis;
- **as complete as possible:** namely, that the machine readable rules capture all those aspects of the ToS that can be enforced automatically.

We are not aware of any previous work that addresses the whole lifecycle of natural-language analysis of privacy texts with the goal of enforcing suitable rules, e.g. in an enterprise setting (although the EU CONSEQUENCE project does take an holistic approach it does not involve natural-language analysis). As stated in the Introduction, achieving compliance with privacy legislation and regulations is a central concern in enterprises, and means of automating compliance are highly desirable. Since new privacy rules are almost exclusively expressed using natural-language, means of automatically analysing the appropriate texts and extracting rules from them necessary – the resulting rules can then be incorporated into existing enterprise rule-bases, such as those used in compliance checkers or information governance (GRC) platforms.

The most naïve analysis seeks to find in the text of an ToS occurrences of particular verbs, namely verbs which are prescriptive by nature; examples include:

"The Provider will provide a backup of data […]";
"The User will not upload pornographic images to the service"

since these typically arise in statements expressing duties and obligations (see also [3]). Certain verb groups appear in phrases expressing rights, typically rights of the customer but not necessarily:

"The Customer may request in writing a full copy of data held [...]"

"The Provider can refuse to provide access to the service at any time [...]"

In the case of simple prescriptive sentences it is possible to represent the information given by a triple *(verb, subject, object)*. Such a representation says nothing of the nature of the rule or (legal) clause appearing in the ToS, but may assist a service provider in automatically generating a set of access control rules for enforcement within its infrastructure. Our tool uses a form of markup referred to as a formal requirements specification language (RSL); the RSL we are using is due to Breaux and Gordon [10].

Our tool is designed to detect delimiters and punctuation, so that long-winded sentences of legalese may be separated into their constituent parts. In a given sentence, those secondary clauses, which serve only to explicate and/or amplify the main thrust of the sentence, may be ignored (subject to interpretation and the judgment of a human user, of course; this suggests the process cannot be completely automated), and a semantic representation can be built of the remaining constituents of the sentence.

An interesting toolkit that we are considering to use to automate part of this task is GATE ("General Architecture for Text Engineering") [4], whose user interface provides a helpful facility for tagging and colour-coding portions of text of particular semantic relevance. The technique that applies here is known as semantic annotation. We believe that such an approach is highly beneficial for the visual representation of the terms and conditions contained in a given cloud Terms of Service document.

5 Conclusions and Future Work

Our main contribution in this paper has been to describe our current work on developing software tools for automated information extraction of cloud terms of service. Further work is required on interpreting the semantic information provided through the editor component. We are at the moment experimenting with the use of the semantic markup formalism in [10], but fully expect more work will be required to interpret and make use of the semantic information within the processing component. Most natural-language processing algorithms are concerned with syntactic issues, and the novelty here is to interpret rules correctly and resolve ambiguities when they arise; however, the problem of interpretation is not intractable or exceedingly complex, given that we are restricting the type of analysis to texts that have quite similar content and general characteristics.

References

[1] V. Ong. An Architecture and Prototype System for Automatically Processing Natural-Language Statements of Policy. Thesis, Naval Postgraduate School, Monterey, California, 2001.

[2] C.D. Manning, H. Schutze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[3] T. D. Breaux, M. W. Vail and A. I. Antón. Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with Regulations. In Proceedings of 14th IEEE International Requirements Engineering Conference (RE'06) (2006).

[4] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M.A., Saggion, H., Petrak, J., Li, Y., Peters, W. 2011. Text Processing with GATE (Version 6). Department of Computer Science, University of Sheffield.

[5] May, M., Gunter, C., Lee, I., Zdancewic, S. 2009. Strong and Weak Policy Relations. In Proceedings of the 2009 IEEE International Symposium on Policies for Distributed Systems and Networks (POLICY '09). IEEE Computer Society, Washington, DC, USA, pp. 33-36, 2009.

[6] Papanikolaou, N., Creese, S., Goldsmith, M. Refinement checking for privacy policies. Science of Computer Programming. Article in Press, DOI:10.1016/j.scico.2011.07.009.

[7] Casassa Mont, M., Pearson, S., Creese, S., Goldsmith, M., Papanikolaou, N. A Conceptual Model for Privacy Policies with Consent and Revocation Requirements. In Proceedings of PrimeLife/IFIP Summer School 2010: Privacy and Identity Management for Life, Lecture Notes in Computer Science, Springer (2010).

[8] Pearson, S., Casassa Mont, M., Kounga, G. 2011. Enhancing Accountability in the Cloud via Sticky Policies. Secure and Trust Computing, Data Management and Applications, Communications in Computer and Information Science, vol. 187, Springer Verlag, Heidelberg, pp. 146-155.

[9] Bradshaw, S., Millard, C., and Walden, I. (2010) *Contracts for Clouds: Comparison and Analysis of the Terms and Conditions of Cloud Computing Services.* Queen Mary University of London, School of Law Legal Studies Research Paper No. 63/2010. Available from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1662374 (Accessed 14 May 2012).

[10] Breaux, T.D., and Gordon, D.G. (2011) *Regulatory Requirements as Open Systems: Structures, Patterns and Metrics for the Design of Formal Requirements Specifications.* Technical Report CMU-ISR-11-100, Institute for Software Research, Carnegie-Mellon University. Available from reports-archive.adm.cs.cmu.edu/anon/isr2011/CMU-ISR-11-100.pdf (Accessed 14 May 2012).