Mikko Lehtonen · Florian Michahelles · Elgar Fleisch

Probabilistic Approach for Location-Based Authentication

Received: 8.6.2007 / Accepted: 30.6.2007

Abstract This paper explores location-based product authentication in a situation where only the past locations of products that flow in a supply chain are known. We transform location-based authentication into a pattern recognition problem and investigate different solutions based on machine-learning techniques. The proposed solutions are studied with computer simulations that model the flow of genuine and counterfeit products in a generic pharmaceutical supply chain. The results suggest that machine-learning techniques could be used to automatically identify suspicious products from the incomplete location information.

Keywords clone detection · data mining · tracking

1 Introduction

Ubiquitous computing technologies such as Radio Frequency IDentification (RFID), but also GPS and GSM, increase the location-awareness of products and other physical objects. Location-awareness allows for new applications and services, including location-based authentication. For example, if track and trace data tells that *product P is in warehouse x*, one can conclude that a product that claims to be P but is not in warehouse x is in fact not P but a cloned one. In many cases, however, one only knows where a product has been, but not where it currently is. It means that the location of a product is known only at discrete points of time and the track and trace data tells, for example, that *product P was observed at location x at time t.* This case calls for different authentication methods.

Product authentication is needed to distinguish counterfeit products from genuine ones in licit supply chains. Legislation in several states in the U.S. [18] as well as in Italy [4] requires that all pharmaceutical products are equipped with a unique pedigree record for authentication purposes, showing the need for technical countermeasures. This paper explores how the incomplete location information can be used to authenticate products that flow in a supply chain. We investigate the case where only a part of the supply-chain players share the location information of the products with a product authentication system. In this case, one can only know where a given product has been but not where it currently is. If all the custodians shared the location information, the product authentication system would have complete visibility and detecting clones would be straightforward. However, it is unlikely that the complete visibility will be always available since companies are careful what information they disclose. We study machine-learning techniques that can be used to automatically train the decision rules for clone detection. Without an automated way to generate the detection rules and apply them to large amounts of data, the solution won't scale to large and complex supply chains. Moreover, our goal is not to find a solution that provides the best possible results for a given problem, but to study how the location information should be used to detect clones in different cases.

This paper is organized as follows. Section 2 reviews the related work. Section 3 presents how we apply machine-learning techniques to location-based authentication. Section 4 describes our simulator study and the findings are presented in Section 5. Section 6 discusses the findings and we finish with conclusions and future work.

2 Related work

Product authentication is most often addressed in the related literature in terms of RFID-based approaches.

Mikko Lehtonen
ETH Zurich, Information Management
E-mail: mlehtonen@ethz.ch

Florian Michahelles
ETH Zurich, Information Management
E-mail: fmichahelles@ethz.ch

Elgar Fleisch
ETH Zurich & University of St. Gallen
E-mail: elgar.fleisch@ethz.ch

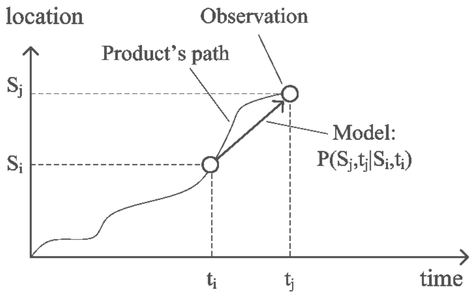

Fig. 1 An illustration of location-based product authentication approach based on estimating the transition probabilities between the observations.

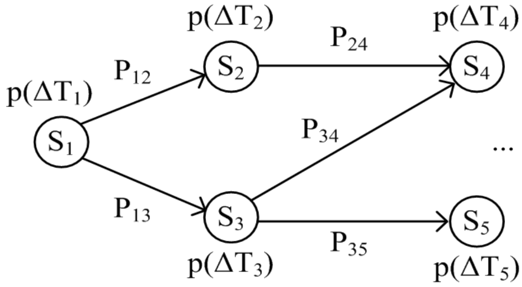

Fig. 2 Our stochastic supply chain model that is used to estimate the transition probabilities between observations.

RFID allows for different ways to authenticate products, including track and trace based plausibility check (e.g., [4]), cryptographic tag authentication (e.g., [2]), and the use of object-specific security features that link the RFID tag to a unique product (e.g., [11]). Different plausibility checks include verification that the product under study has a valid ID number, verification that the product under study has a valid trace (data on tag), and verification that the product under study is not a cloned one (data on network). The cloned products are detected by knowing where the product should be based on complete visibility in the chain of custody, and the case where the location information is incomplete, to our best knowledge, is not formally addressed in the related literature. Juels [2] has acknowledged, however, that already the unique numbering of objects, even in the absence of resistance to RFID tag cloning, can be a powerful anti-counterfeiting tool, in terms of clone prevention and clone detection.

Finding cloned products from the track and trace data can be seen as intrusion detection. Intrusion detection means the process of identifying and responding to malicious activity targeted at computing and networking resources [16]. Intrusion detection techniques are traditionally classified as anomaly- or signature- based. Signature-based systems act similar to virus scanners and look for known, suspicious patterns in their input data. Anomaly-based systems watch for deviations of actual from expected behavior and classify all "abnormal" activities as malicious. Intrusion detection techniques have been applied to RFID data, though so far not in supply chain applications. Mirowski [5] applied intrusion detection techniques to detect cloned RFID access cards, but the method is prone to false alarms.

Also credit card fraud detection deals with similar problems than location-based authentication. There the problem is to detect fraudulent transactions, which corresponds to the detection of copied credit cards, by looking for specific transaction patters in a large amount of data. Data mining techniques such as pattern recognition and classification have been successfully applied to detect fraudulent transactions (e.g., [6],[7]), and fraud-detection systems are currently in use to protect credit card companies and their customers.

3 Location-based product authentication

In this study we assume that a product authentication system is secure if it can detect traces generated by cloned products. A cloned product stands for a counterfeit product that carries the ID number of a genuine product. This means that even when the authentication system knows that a cloned product is present, it might not know for sure which of the observed products is the genuine and which is the cloned one. In this section, we describe two approaches how we apply machine-learning techniques to location-based product authentication.

3.1 Stochastic supply chain model (SSCM) approach

Detection of cloned products is easy if the locations of the genuine products are known; for instance, if the track and trace data says that the genuine product is in Switzerland while a product in Japan claims to be the genuine one, the system can conclude that the product in Japan is a cloned one. However, if the track and trace data says that the genuine product was observed in Switzerland one week ago but makes no statement of where it currently is, the authentication becomes non-trivial. We argue that the location information can still be used in authentication, though the authentication results become inheritably less certain in the absence of complete location information.

As a solution, we propose that the product authentication system estimates the transition probability that the genuine product has moved from Switzerland to Japan between the observations. This location-based authentication method is illustrated in Fig. 1. If the transition probability is low, for example because the product should have moved faster than any commercial jet plane, the observed product in Japan is likely to be a cloned one. When we denote the previous observed location of a subject with S_i and the current location with S_j, the observation times with t_i and t_j, respectively, this location-based authentication method can be formalized as follows. *Subject is authentic, if*:

$$P(S_j, t_j | S_i, t_i) > \epsilon \quad (1)$$

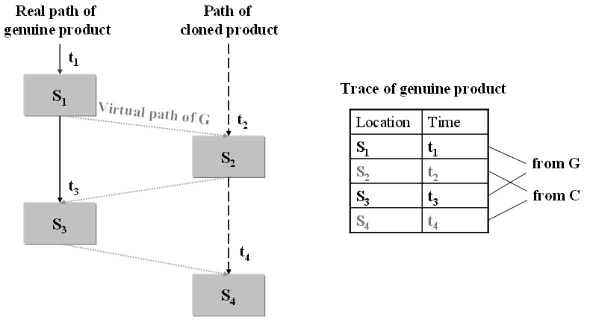

Fig. 3 Illustration how a copied product corrupts the trace of a genuine product (G), creating a virtual path for the genuine product that is seen from the trace ($t_1 < t_2 < t_3 < t_4$)

To estimate the probability in equation 1, we train a discrete-time stochastic supply chain model (SSCM) that has N distinct states, $S_1, S_2, ..., S_N$. When a product enters a state in the model, it will generate an observation that is analogous to sending a notification of receiving the product to the product authentication system. This is also when the authenticity check is performed. The time, measured in number of steps, that a product waits in a state is given by a probability density function (PDF) specific to each state. For state i, this PDF is denoted as $p(\Delta T_i)^1$. After time ΔT the product will enter a new state according to a set of state transition probabilities. The state transition probabilities are time independent and denoted as $P_{ij} = P(q_t = S_j | q_{t-1} = S_i)$, where q_t is the state of the product at time t. The state transition probabilities have the following properties:

$$\sum_{j=1}^{N} P_{ij} = 1, \qquad 0 \leq P_{ij} \leq 1 \tag{2}$$

In our model, the state transition probabilities from a state to itself (P_{ii}) are not defined because the time a product spends in a state depends on the state's waiting time distribution (cf. Fig. 2). When $S = \{S_1, ..., S_n\}$ is the sequence of states where the product has been observed, we consider the set of state transition probabilities, $L = \{P(S_2|S_1), ..., P(S_n|S_{n-1}\}$, and the set of probabilities of waiting times between the observations, $T = \{P(\Delta T_1), ..., P(\Delta T_{n-1}\}$. We study two different confidence values that estimate the transition probability in equation 1. When S stands for a subset of probabilities L and T, these confidence values are:

$$c_1(S) = \prod_{\forall P_i \in S} P_i \tag{3}$$

$$c_2(S) = \min_{\forall P_i \in S} (P_i) \tag{4}$$

We study the performance of all six different confidence values that can be derived from the SSCM, i.e., $c_i(S)$ with $i \in \{1, 2\}$ and $S \in \{L, T, L \cup T\}$. Using these confidence values, equation 1 is transformed into the following form[2]. *Subject is authentic, if*:

$$c_i(S) > \epsilon, \qquad i \in \{1, 2\} \tag{5}$$

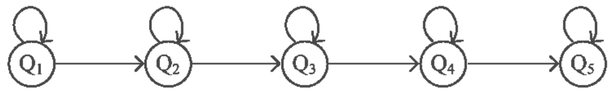

Fig. 4 An example (left-to-right) hidden Markov model.

3.2 Hidden Markov model (HMM) approach

When a genuine product (G) and a copied product (C) flow in a supply chain where location information is (partially) shared with a product authentication system, the events generated by G and C are appended to a single trace. In other words, the trace of G is *corrupted* by events that origin from C. This is illustrated in Fig. 3. In this way, location-based product authentication can be presented as a classification problem where the traces of products are classified into traces that are generated by only one, genuine product, and traces that are generated by genuine and (one or more) copied products.

Dividing the time axis into equally-spaced discrete steps allows us to formulate *observation vector v* whose elements tell the last observed location of a product for a given time step. Hence, the observation vector effectively presents the known trace of a product. When there are no copied products in the supply chain, v is generated by a single product G that moves in the supply chain. When there are copied products in the supply chain, v is generated by multiple products, G and (one or several) C. Using two models, M_1 and M_2, that represent uncorrupted and corrupted traces, respectively, we formulate the location-based authentication as follows. *Subject is authentic, if*:

$$P(v|M_1) - P(v|M_2) > \epsilon \tag{6}$$

To estimate the probabilities in equation 6, we train a hidden Markov model (HMM) classifier. The HMM (cf. Fig. 4) is a very powerful statistical method of characterizing observed data samples of a discrete-time series and it has been successfully applied in many classification problems such as keyword spotting [12]. The underlying assumption of the HMM is that the data samples can be well characterized as a parametric stochastic process, and the parameters of the stochastic process can be estimated in a precise and well-defined framework [8].

A Markov chain can be extended to a HMM by introducing a non-deterministic process for each state for generating the output observations [8]. We build the classifier by training two models, M_1 and M_2, that represent the cases where there are one or, respectively, multiple products in the supply chain. The number of hidden

[1] We use upper case $P(.)$ to denote a probability mass function and lowercase $p(.)$ for a probability density function

[2] When using c_1, the threshold ϵ is normalized with the number of terms in the multiplication n to ϵ^n

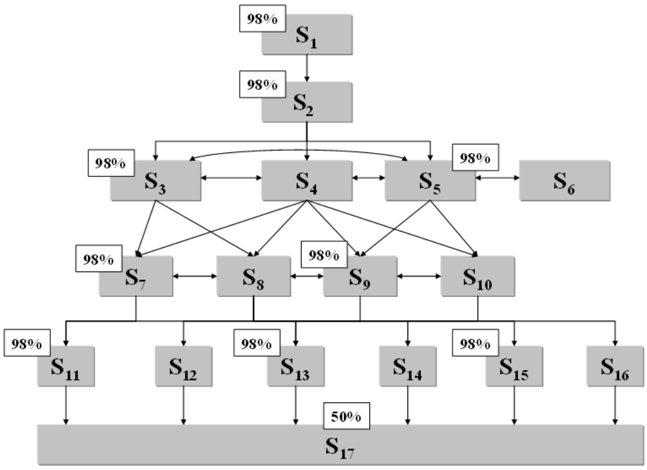

Fig. 5 The simulated pharmaceutical supply chain: nodes 1-2 represent the manufacturing level, 3-10 the wholesale level, 11-16 the retail level, and 17 the patient. The percentages represent reading rates in corresponding nodes.

states of both models is set to five that seemed to give the best results. The models are trained using the standard Baum-Welch expectation-maximization algorithm (e.g., [1]) and the used features are observation vectors v cut after the first element of the last observed location.

4 Simulation settings

We study the proposed authentication approaches by simulating flow of products in a generic pharmaceutical supply chain in a discrete event environment. Our pharmaceutical supply chain starts from the manufacturing level and ends to the patient who gets the drug product from retail level which consists of pharmacies and hospitals. Between these levels there are wholesalers who buy and sell drug products and can repackage them. Pharmaceutical supply chains are complex and the products can change hands up to half a dozen times [13], opening many opportunities to introduce counterfeits.

Our simulated supply chain is presented at Fig. 5. Node 1 presents the production line where the products obtain unique ID numbers and node 2 the manufacturer's warehouse. Nodes 3-10 present the wholesale level, including repackaging (node 6). We consider two levels of pharmaceutical wholesalers, central warehouse level (nodes 3-5) and regional warehouse level (nodes 7-10) [14]. Arrows in Fig. 5 indicate the possible ways how products can flow among the different players. When a product enters a node, it waits a random time between minimum and maximum waiting times that are specific to each node, and moves to a new node according to the node's state transition probabilities. The average lead time from manufacturer to patient in our model is one and a half months. We assume that the manufacturer and half of the supply chain partners share the location information with the product authentication system. In reality this can be done, e.g., by reading the products' RFID tags and by sharing the data in EPC network [15]. Since in reality not all the products are always read (though that is the goal), the probability that an observation is generated when a product enters a data sharing node is set to 98%. In addition, we assume that 50% of products that enter the final node are observed, corresponding capturing and sharing the point-of-sales and point-of-use data.

Table 1 Hit rates for different lots of counterfeit products for different methods (genuine products are produced during month 2; ϵ is set to achieve 1% false alarm rate; results are averaged from 10 Monte-Carlo simulations)

Approach	Lot A (month 1)	Lot B (month 2)	Lot C (month 3)	All
SSCM:				
$c_1(L)$	17%	8%	16%	14%
$c_1(T)$	31%	16%	9%	18%
$c_1(T \cup L)$	35%	20%	20%	25%
$c_2(L)$	**60%**	**46%**	37%	**47%**
$c_2(T)$	44%	29%	40%	38%
$c_2(T \cup L)$	46%	27%	**42%**	38%
HMM	72%	18%	26%	38%

We run simulations during a four month period (120 days), one time-step corresponding to one day. The manufacturer produces 30,000 genuine products during month two, 1000 products a day. We inject 300 copied products randomly in the wholesale level during month 1 (Lot A), month 2 (Lot B), and month 3 (Lot C). All 900 copied products have different identities, copied from the genuine products. In about 10% of the cases, a genuine product and its clone are observed in two different locations during one time step. Though in these cases the clones could be detected by defining that a product cannot be in two locations during one day, these collided products are omitted from the results.

We train our models by data simulated by the same supply chain model that is used in testing phase. The results are presented as the hit rate (ratio of corrupted traces detected) versus the false alarm rate (ratio of uncorrupted traces classified as corrupted). The results are averaged from 10 Monte-Carlo iterations that include training the models. The simulator is implemented in Matlab using the HMM toolbox[3].

The parameters of the SSCM are trained from 300 products. The state transition probabilities are estimated by *a priori* state transition probabilities of the training data, and the waiting time distributions of each node are estimated by uniform distributions matched between the smallest and largest observed waiting times. A very small probability is given to state transitions and waiting times that are not observed in the training data.

[3] Kevin Murphy. Hidden Markov Model (HMM) Toolbox for Matlab. http://www.cs.ubc.ca/ murphyk/Software/HMM/hmm.html

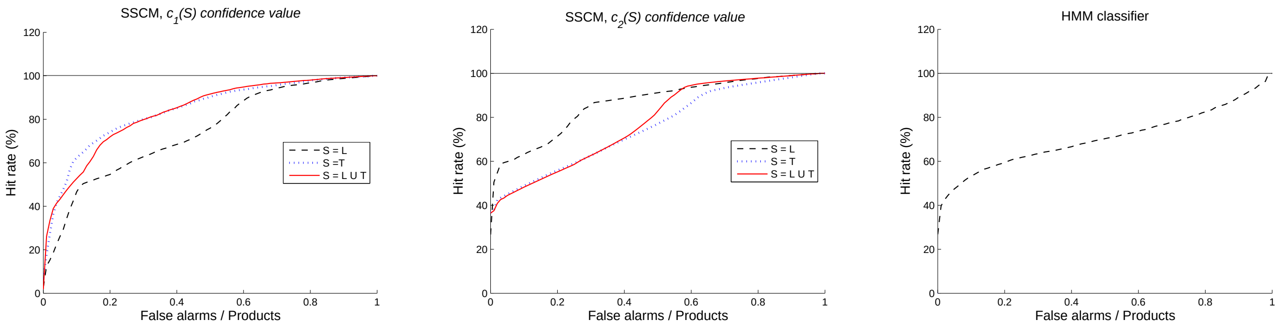

Fig. 6 Hit rate versus false alarm rate for the studied product authentication methods. The curves are obtained by varying the probability threshold ϵ. The results are averaged from 10 Monte-Carlo simulations.

The HMM classifier is trained from 600 traces, including 300 uncorrupted traces (M_1) and 300 traces (M_2) corrupted by clones injected before, simultaneously, and after the production of the genuine products. 40% of the training data of HMM classifier is used in validation to find the optimal amount of training iterations [1].

5 Findings

The results of the simulator study are presented in Table 1 and Fig. 6. They show that the proposed methods can detect some of the cloned products from incomplete location information when the underlying process that generates the observations is known and modeled.

In the SSCM approach, the confidence value that uses the minimum of the set of state transition and waiting time probabilities (equation 4) performed better than the confidence value that chains these probabilities (equation 3). This indicates that it's better to search for cloned products by looking for single suspicious observations than multiple unlikely observations. Moreover, looking for single unlikely state transitions, $c_2(L)$, was the best way to detect clones in the simulations. Combining location and time information to estimate the transition probabilities yielded somewhat better results with the chained confidence value than using only location or time, but overall it appears that the optimal way to combine the complementary location and time information was not found in this study.

The HMM approach was not able to outperform the SSCM approach. Overall, the HMM classifier does not give satisfactory results given its complexity and amount of training data. Table 1 reveals that the relative good average performance of the HMM classifier was given mostly by its capability to distinguish the cases where a cloned product enters the supply chain before the genuine product is manufacturer (Lot A). We observed that the HMM classifier gave best results with five hidden states per model, suggesting that hidden states correspond to the levels of the modeled supply chain.

All the studied methods are somewhat prone to false alarms; achieving very high hit rates appears to be possible only through high false alarm rates. This result is in concordance with how intrusion detection techniques perform in RFID access control application [5]. Table 1 shows that up to 47% hit rate can be achieved with a 1% false alarm rate. Overall, detection of clones that are injected in the supply chain before the genuine products (Lot A) seems to be somewhat easier than of those that are injected simultaneously (Lot B) or after (Lot C) the genuine ones. The probable reason for this is that when the first observation for a product is not at the manufacturing level but at the wholesale or retail level, the models can easily classify the trace as abnormal. It is important to note that the case where a cloned product is injected in the supply chain before the genuine product can be solved in the SSCM approach by introducing a state of "non-existence" and accepted transitions from that state. The case where a clone is injected after the genuine product has reached the consumer can be solved in a similar way. This would restrict the problem to the case where the genuine and copied product are simultaneously in the supply chain.

6 Discussion

The level of security of the studied methods, in terms of probability to detect the clones, is relatively low but still has a big impact on the counterfeiter's expected return from repeated criminal activities because the repeated use of copied ID numbers is made infeasible. In the case of static, cloning resistant security labels like watermarks and cryptographic RFID tags, a counterfeiter needs to break the security label only once (though it is hard) but can then copy it to virtually unlimited number of counterfeit products. In the case when copied products can be detected, however, counterfeiter has to forge several ID numbers to inject a large number of counterfeit products. An important advantage of location-based product authentication in supply chains is low variable cost to secure one product - only a unique identifier is needed, assuming that the data capturing and sharing infrastructure is in place. This is cheaper than protecting products using cryptographic RFID tags, for example, that provide cloning resistance. Location history also allows for pinpointing the players who inject counterfeits to the supply chain. Disadvantage of location-based product authentication is that it requires custodians to capture and share the location information.

Our assumption that business partners across the supply chain share the item-level data with a product authentication system is rather strong because today companies disclose such data only rarely. However, we can assume that sensing technologies will increase the location-awareness of products, and therefore it is important to know what the value of this information in different applications is. As already mentioned, if all the supply chain players would agree to share the location information with the product authentication system, detection of clones would become substantially easier. However, every time when the location information is incomplete, the traces of genuine products can be corrupted by cloned products and detection of cloned tags is needed.

Our simulated pharmaceutical supply chain is a heuristic model. Accordingly, many real world phenomena that also affect the results, e.g., erroneous shipments, shrinkage, recalls, and changes in demand need to be studied in future. In addition, our simulated supply chain doesn't change in time, though in reality the day of the week and season, for instance, affect the conditions, and also the supply chain partners can change. A practical solution should also address these changes. Overall, accurate modeling of supply chains as presented in [10], for instance, is not addressed. Therefore the results of the simulation study need to be validated using more realistic data. We acknowledge that especially real corrupted traces are hard to obtain in practice, so it is likely that they need to be at least partially simulated.

7 Conclusions and future work

This study suggests that machine-learning techniques can be used in location-based product authentication in supply chain applications when sufficient visibility can be provided. The studied methods are somewhat prone to false alarms and therefore serve best as identifying suspicious products. Our study shows that both time and location of observations carry information that can be used in authentication. The future work includes finding optimal ways to combine this information, validating the concepts with realistic data, and a sensibility analysis to measure how deviations between the real and the estimated model parameters affect the results. Also the assumptions regarding data sharing among the supply chain partners need to be verified to identify prominent real-life cases for the studied methods. Overall, we believe that the studied methods have potential for refinements and tuning. Finally, in addition to products in a supply chain, the presented methods can be applied to authenticate subjects also in other cases where the underlying process that generates the location observations is not completely random but respects some parametric model, e.g., vehicles, access cards, credit cards and other payment cards, mobile computing devices, and luggage.

Acknowledgements This work was supported in part by the European research project BRIDGE (Building Radio frequency IDentification solutions for the Global Environment, 033546). The Authors would like to thank Jasser Al-Kassab, Alexander Ilic, John Jenkins, Thorsten Staake, and the anonymous IWSSI reviewers for their help with the paper.

References

1. R. Duda, P. Hart, and D. Stork, Pattern Classification. Second Edition. New York: John Wiley Sons, 2001.
2. S. Dominikus, E. Oswald, and M. Feldhofer, "Symmetric authentication for RFID systems in practice," in ECRYPT Workshop on RFID and Lightweight Crypto, Austria, 2005.
3. A. Juels, "RFID Security and Privacy: A Research Survey," IEEE Journal of Selected Areas in Communications, vol. 24, pp. 381-394, February 2006.
4. R. Koh, E. Schuster, I. Chackrabarti, and A. Bellman, "Securing the Pharmaceutical Supply Chain," Auto-ID Labs White Paper, 2003.
5. L. Mirowski, Detecting Clone Radio Frequency Identification Tags. Bachelor's Thesis, School of Computing, University of Tasmania, November 2006.
6. P. Chan, W. Fan, A. Prodromidis, and S. Stolfo, "Distributed Data Mining in Credit Card Fraud Detection, " in IEEE Intelligent Systems, vol. 14, pp. 67-74, November/December 1999.
7. S. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. Chan, "Credit card fraud detection using meta-learning: Issues and initial results," in AAAI-97 Workshop on Fraud Detection and Risk Management, 1997.
8. X. Huang, A. Acero, and H.W. Hon, Spoken Language Processing; A Guide to Theory, Algorithm, and System Development. New Jersey: Prentice-Hall PTR, 2001.
9. L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Appliactions in Speech Recognition," in Proceedings of IEEE, vol. 77, no.2, pp. 257-286, February 1989.
10. J. Swaminathan, S. Smith, and N. Sadeh, "Modeling Supply Chain Dynamics: A Multiagent Approach," in Decision Sciences, vol. 29, pp. 607-632, 1998.
11. Z. Nochta, T. Staake, and E. Fleisch, "Product Specific Security Features Based on RFID Technology," International Symposium on Applications and the Internet Workshops (SAINTW'06), 2006, pp. 72-75.
12. R. Rose, "Keyword Detection in Conversational Speech Utterances using Hidden Markov Model based Continuous Speech Recognition" in Computer Speech and Language, vol. 9, pp. 309-333, October 1995.
13. P. Jaret, "Fake drugs, real threat." Los Angeles Times, 9 February, 2004.
14. A. Weis and A. Josten, "Effective Brand Protection in the Pharmaceutical Industry Needs Efficient Supply Chain Management." Pharmaceutical Manufacturing and Packing (PMPS), Autumn 2002.
15. EPCglobal. (2005, July). EPCglobal Architecture Framework Version 1.0.
16. E. Amoroso, Intrusion Detection: An introduction to internet surveillance, correlation, trace back, traps, and response. First Edition. Intrusion.Net Books, 1999.
17. Organization for Economic Co-operation and Development (2006). The Economic Impact of Counterfeiting.
18. Verisign. (2005). Electronic Drug Pedigree ("E-Pedigree"): Considerations in Choosing a Partner for the Drug Pedigree Race. [Online]. Available: http://www.verisign.com/