Contemporary Management Research
Pages 93-108, Vol. 9, No. 2, June 2013
doi:10.7903/cmr.3845

Handheld Usage Data Mining for Handheld Data Protection

Wen-Chen Hu
University of North Dakota
E-Mail: wenchen@cs.und.edu

Naima Kaabouch
University of North Dakota
E-Mail: naima.kaabouch@engr.und.edu

Lei Chen
Sam Houston State University
E-Mail: chen@shsu.edu

Hung-Jen Yang
National Kaohsiung Normal University
E-Mail: hjyang@nknucc.nknu.edu.tw

ABSTRACT

Mobile handheld devices such as smart cellular phones are easily lost or stolen because of their small sizes and high mobility. Personal data such as addresses and messages stored in the devices may be revealed when the devices are lost. Handheld devices must include rigorous and convenient handheld data protection in case the devices are lost or stolen. This research proposes a novel approach for handheld data protection by using handheld usage data mining, which consists of five steps: (i) usage data gathering, (ii) usage data preparation, (iii) usage pattern discovery, (iv) usage pattern analysis and visualization, and (v) usage pattern applications. Handheld usage data is collected before applying this method. Usage patterns are discovered and saved by using finite automaton, which is then used to check device usage. When an unusual usage pattern such as an unlawful user trying to access the handheld data is detected, the device will automatically lock itself down until an action, such as entering a password, is taken. Experimental results show this method is effective and convenient

for handheld data protection.

Keywords: Handheld Security, Mobile Handheld Devices, Smartphones, Data Mining, Usage Mining, Usage Pattern Discovery, and Identification

HANDHELD USAGE DATA MINING FOR HANDHELD DATA PROTECTION

Worldwide PC and mobile phone sales are given in Table 1 according to various market research reports (BNET, 2004; Canalys, 2007 & 2010; CNET, 2006; Gartner, 2005–2011; GsmServer, 2004; IDC, 2008). The number of smartphones shipped worldwide has passed the number of PCs and servers shipped worldwide in 2011, and the gap between them is expected to increase. People can carry handheld devices anytime, anywhere, and use them to perform daily activities such as making phone calls, checking schedules, and browsing the mobile web. However, they are easily lost or stolen because of their small size and high mobility. Personal data such as addresses and messages stored in the devices may be revealed when the devices are lost (Ghosh & Swaminatha, 2001) or used by unauthorized persons.

This research proposes a set of novel approaches to solve the problems of handheld data protection by using usage pattern identification. The proposed method is divided into five steps:

- Usage data gathering, which is to collect the device usage data;
- Usage data preparation, which removes noises from the raw usage data;
- Usage pattern discovery, which finds valuable patterns from the prepared usage data;
- Usage pattern analysis and visualization, which is to analyze and display the discovered patterns for finding hidden knowledge; and
- Usage pattern applications, one of which is handheld data protection used in this research.

A usage finite automaton is constructed based on the handheld usage data collected and prepared. The automaton is then used to check against any possible unauthorized uses. Experimental results show this method is effective and convenient for handheld data protection, but the accuracy may need to be further improved.

Table 1 Worldwide PC and cellphone sales

Year	Number of Units Shipped (Million)				
	Mobile Phones	PCs and Servers	Smartphones	PDAs (without phone capability)	Tablet PCs
2002	432	148	—	12.1	—
2003	520	169	—	11.5	—
2004	713	189	—	12.5	—
2005	813	209	—	14.9	—
2006	991	239	64	17.7	—
2007	1153	271	122	—	—
2008	1220	302	139	—	—
2009	1221	306	166	—	1
2010	1609	346	286	—	17
2011	1775	353	486	—	73
2012 (Estimated)	1900	368	702	—	119

LITERATURE REVIEW

This research is related to two themes: handheld security and handheld usage mining. Related research of these themes will be discussed in this section.

Handheld Security

Many companies such as the device manufacturer HP (Hewlett-Packard Development Company, L.P., 2005) and the embedded database vendor Sybase Inc. (2006) propose practical handheld security methods (e.g., the owners of lost devices can call the centers to remotely lock down the devices). Those methods are normally workable, but not particularly innovative. Susilo (2002) identifies the risks and threats of having handheld devices connected to the Internet and proposes a personal firewall to protect against the threats. Argyroudis et al. (2004) present a performance analysis focused on three of the most commonly used security protocols for networking applications; namely SSL, S/MIME, and IPsec. Their results show that the time taken

to perform cryptographic functions is small enough not to significantly have an impact on real-time mobile transactions and that there is no obstacle to the use of quite sophisticated cryptographic protocols on handheld mobile devices. Public keys are used to encrypt confidential information. However, limited computational capabilities and the power of handheld devices makes them ill-suited for public key signatures. Ding et al. (2007) explore practical and conceptual implications of using server-aided signatures (SAS) for handheld devices. SAS is a signature method that relies on partially trusted servers for generating (normally expensive) public key signatures for regular users. Digital watermarking is particularly valuable in the use and exchange of digital media on handheld devices. However, watermarking is computationally expensive and adds to the drain of the available energy in handheld devices. Kejariwal et al. (2006) present an approach in which they partition the watermarking embedding and extraction algorithms and migrate some tasks to a proxy server. This leads to lower energy consumption on the handheld without compromising the security of the watermarking process.

Handheld Usage Mining

This research applies the method of usage mining to handheld data protection. Two major methods are commonly used to find usage patterns:

- Sequential pattern generation (Agrawal & Srikant, 1995), which is to discover sequences of maximal length that appear more frequently than a given percentage threshold over a collection of transactions ordered in time.
- Association rule discovery (Agrawal & Srikant, 1994), which is used to find unordered correlations between items found in a set of database transactions. In the context of handheld usage mining, association rules refer to sets of applications that are accessed together with a support value exceeding some specified threshold.

Usage mining can be applied to many areas, especially recommender systems. Two of the systems are explained next:

- Intelligent web recommendations derived from frequent user web-access patterns can help typical mobile users efficiently navigate standard websites. Zhou et al. (2006) propose an implicit server-side approach using intelligent web recommendations that can significantly enhance the mobile-browsing experience.
- Recommender systems rely on relevance scores for individual content items; in particular, pattern-based recommendation exploits co-occurrences of items in user

sessions to ground any guesses about relevancy. To enhance the discovered patterns' quality, Adda et al. (2007) propose using metadata about the content that they assume is stored in a domain ontology. Their approach comprises a dedicated pattern space built on top of the ontology, navigation primitives, mining methods, and recommendation techniques.

THE PROPOSED SYSTEM

This research proposes the following steps to protect sensitive data in a handheld device from unauthorized accesses (Hu, et al., 2005):

- Usage data gathering,
- Usage data preparation,
- Usage pattern discovery,
- Usage pattern analysis and visualization, and
- Usage pattern applications for handheld data protection.

Figure 2 shows the steps and data flows among them. The steps will be introduced next, and Steps C and E will be detailed in the next section. If the system detects a different usage pattern from the stored patterns, it will assume the users are unlawful and block their accesses. The users need to verify their identities such as entering passwords or answering a question in order to continue their operations. This approach has the advantages of convenience and vigorous protection compared with other approaches such as password protection and fingerprint recognition.

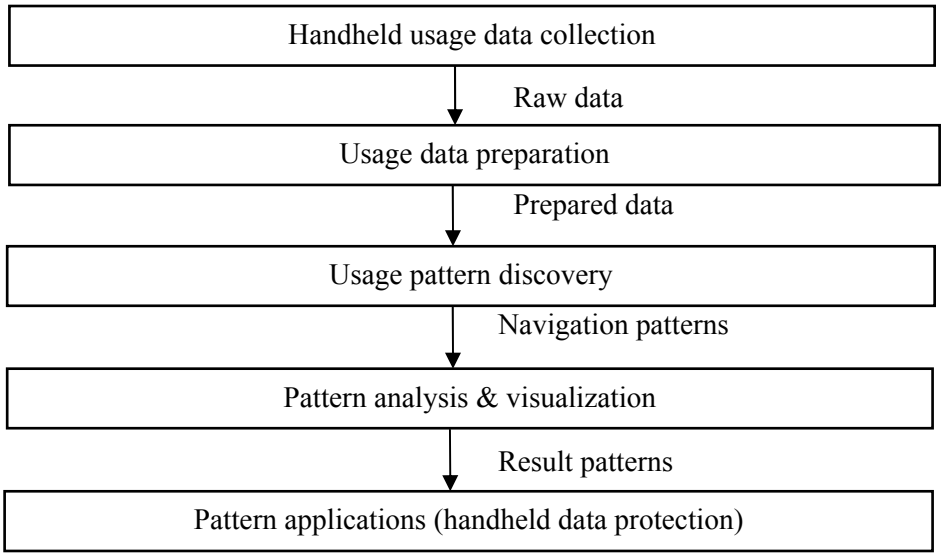


Figure 2 The structure of the proposed system

Usage Data Gathering

This stage focuses on collecting data of defined categories in order to construct a user usage profile. The data should include the user's unique characteristics of using that handheld device. Our research is based on the assumption that every user has a set of distinguishable and identifiable usage behaviors, which can separate one user from another. This assumption has been verified and applied to other information security applications, including intrusion detection. For example, a cell phone user may follow the patterns below to operate his/her phone the first thing in the morning:

- Turn on the cellular phone.
- Check phone messages.
- Check address book and return/make phone calls.
- Check instant messages.
- Reply/write messages.
- Check schedule book.
- Write notes.
- Turn off the cellular phone.

The above steps are one example of handheld usage patterns. Other patterns exist for the user, and each user has his/her own unique usage patterns. To collect usage data, users click on the icon "Pattern" on the interface in Figure 3.a to bring up the interface in Figure 3.b, which asks users to enter a number of days of usage data collection. The collection duration could be a week or a month depending on the use frequencies. The interface as shown in Figure 2.a is re-implemented, so when an application is clicked, it is recorded, and the application is then activated.

(a) (b)

Figure 3 (a) The user interface of a device re-implemented to gather usage data and (b) user entry of data collection time

Usage Data Preparation

The data collected from the Step A is usually raw and therefore cannot be used effectively. For example, the usage patterns should not include an event of alarm-clock operation if the user rarely uses the alarm clock. Data preparation may include the following tasks (Mobasher, Cooley, & Srivastava, 2000):

- Delete the event whose frequency is less than a threshold value such as 5. For example, if the usage data is collected for a month, data synchronization can be ignored if it is performed twice during that period.
- Remove the event if its duration is less than a threshold value such as 10 seconds. An event lasting less than 10 seconds is usually a mistake.
- Repeatedly performing the same action is considered performing the action one time. For example, making three phone calls in a row is treated as making one call.

The interface in Figure 4.a allows users to decide whether or not to modify the default threshold values. If the user clicks the button "Yes," the interface in Figure 4.b allows him/her to enter two new threshold values.

(a) (b)

Figure 4 (a) Users deciding whether or not to modify the threshold values and (b) two input fields for threshold values

After the raw usage data is prepared, a usage tree is created. Figure 4 shows a sample simplified usage tree, where the number and letter inside the parentheses represent the number of occurrences and the shorthand of the event for the use of the next section, respectively. For example, (P: 20) means the event "making phone calls" occurs 20 times. This usage tree is only a simplified example. An actual usage tree is much large and complicated. Ideally, a directed graph instead of a tree should be used to describe the usage data. However, a directed graph is more complicated and therefore is difficult to process. Using a tree can simplify the processing, but it also

creates duplicated nodes (e.g., the event "making phone calls" appears four times in the usage tree of Figure 5).

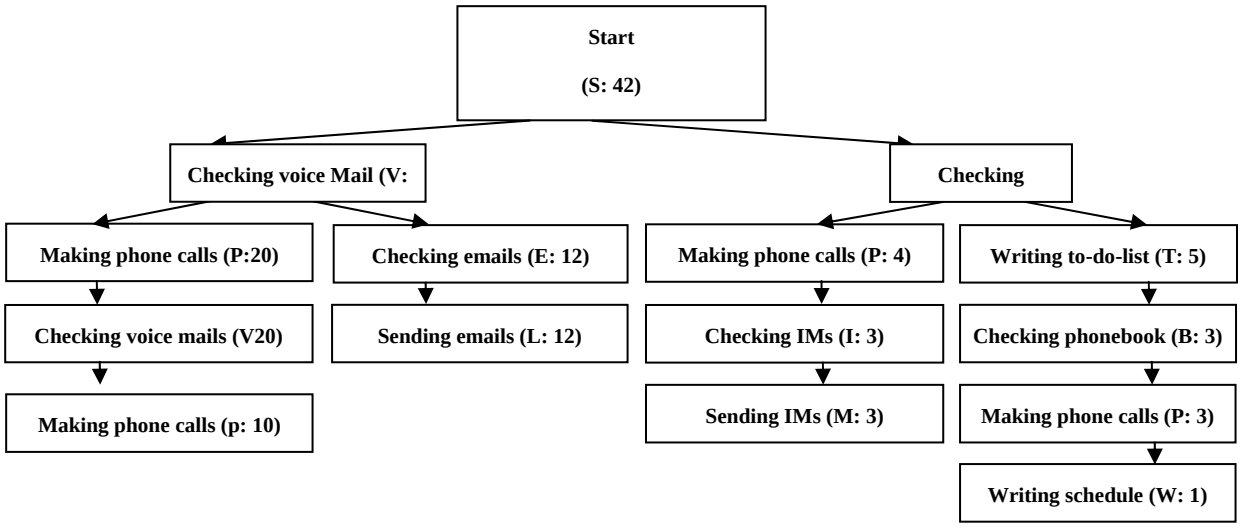


Figure 5 A sample simplified usage tree

Usage Pattern Discovery

This stage focuses on identifying the desired usage patterns. Given the complexity and dynamic nature of user behaviors, identified usage patterns could be fuzzy and not that apparent. Advanced AI techniques such as machine learning, decision tree, and other pattern matching and data mining techniques can be applied in this stage. Many data mining algorithms are applied to usage pattern discovery. Among them, most algorithms use the method of sequential pattern generations (Agrawal & Srikant, 1995), while the remaining methods tend to be rather ad hoc. The problem of discovering sequential patterns consists of finding inter-transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. This research converts the usage tree into a deterministic finite automaton (DFA). Pattern identification is achieved by feeding the automaton with usage sequences. Nothing happens if a sequence ends at an accepting state. Otherwise, a security action will be activated. Details of usage pattern discovery will be given in the next section.

Usage Pattern Analysis and Visualization

The major task of this step is to pick useful ones from the discovered patterns and display them. If the figure of the usage tree and the usage DFA in Section IV can be displayed on the device screen, it will greatly help the mobile users to better manage the proposed methods. However, creating and displaying complicated figures takes much computation time and consumes valuable resources such as memory from the device. Therefore, this research allows users to check the usage data, which may be too complicated to be used by users, but not the usage figures.

Usage Pattern Applications

Usage patterns can be applied to various applications such as recommendation systems (Adda, et al., 2007) and web page reorganization (Eirinaki & Vazirgiannis, 2003). This research uses the handheld usage pattern identification to find any illegal uses of the device. Details of pattern applications of handheld data protection will be given in the next two sections.

Usage Finite Automata

Finding a sequence from the usage tree is costly because the running time of the matching is at least $O(|V_1||V_2|)$, where V_1 and V_2 are the node sets of the sequence and tree, respectively. To speed up the searches, this research applies finite-automaton technologies (Aho, et al., 2006) to usage-pattern matching. A usage finite automaton M is a 5-tuple $(Q, q_0, A, \Sigma, \delta)$ where

- Q, which is a finite set of states,
- $q_0 \in Q$, which is the start state,
- $A \subseteq Q$, which is a distinguished set of accepting states,
- Σ, which is a set of events, and
- δ, which is a function from $Q \times \Sigma$ into Q, called the transition function of M.

For a prepared usage tree from Part B of the previous section, a usage DFA (deterministic finite automaton) M can be constructed by following the steps:

- Each path starting at the root and ending at a leaf is a regular expression. For example, the regular expression of the path

  Checking schedule (H) → Making phone calls (P) → Checking IMs (I)

  → Sending IMs (M)

  is "HPIM" where the letters are shorthand of the events in Figure 4.

- Combine all regular expressions into a regular expression by using the "or" operator '|'. For example, the result regular expression of the usage tree in Figure 4 is "VPVP|VEL|HPIM|HTBPW."
- Convert the regular expression into an NFA (nondeterministic finite automata).
- Convert the NFA to a DFA where
  1. An edge label is an event such as making phone calls.
  2. An accepting state represents a match of a pattern.

For example, the DFA of the usage tree in the Figure 5 is given in Figure 6, where the nodes of double circles are the accepting states.

Using a DFA to store usage patterns and search for patterns is an effective, convenient method, but this approach also suffers the following shortcomings:

- The DFA may accept more patterns than the usage tree does. For example, the pattern

  Checking schedule → Making phone calls → Checking voice mails

  → Checking emails → Sending emails

  is accepted by the DFA according to its DFA path:

  0 → 1 → 4 → 2 → 5 → 8

where the final state 8 is an accepting state, but the pattern does not exist in the tree. However, this feature may not be considered harmful because it may accept more "reasonable" patterns. For example, the above pattern is legitimate (i.e., the users may as well operate their devices by using the pattern "checking schedule, making phone calls, checking voice mails, checking emails, and sending emails").

- This approach misses an important piece of information: the event frequency. The Step B, Usage Data Preparation, of this method removes events with frequencies lower than a threshold value. Otherwise, this DFA does not use the frequency information, which could be useful.

The pattern discovery is virtually not used in this research because the DFA uses all paths from the usage tree. Without using much pattern discovery, the usage tree and DFA may grow too large to be stored in the device.

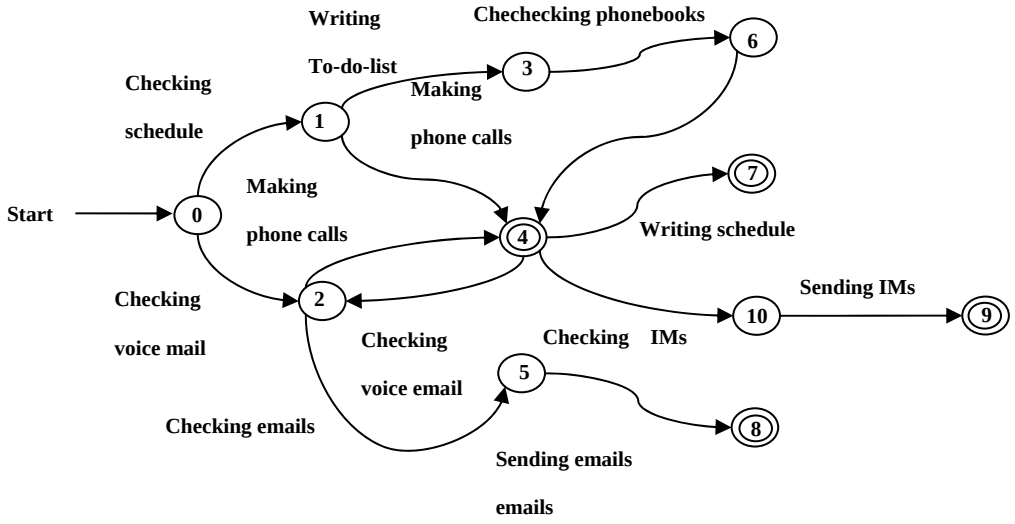


Figure 6 A deterministic finite automaton of the prepared usage tree in Figure 4

Experimental Results

The DFA can be used to find unlawful uses, which stop at the nonending states or could not reach to the final states. The finite automaton begins at the state 0 and reads the user actions one at a time. The automaton accepts the user actions if the actions end at any of the accepting states. For example, a sequence of handheld actions is as follows:

Checking voice mails → Making phone calls → Checking voice mails

→ Making phone calls → Checking IMs → Sending IMs

Its DFA path is as follows:

0 → 2 → 4 → 2 → 4 → 10 → 9

which ends at the accepting state 9, and no security action is taken. Another sequence of actions is as follows:

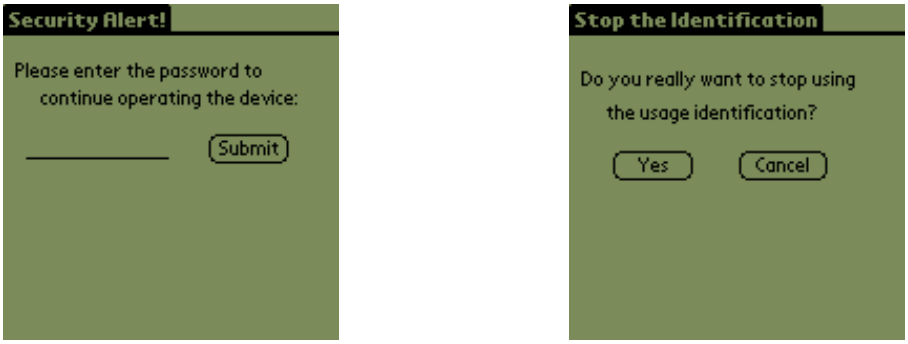
Checking voice mails → Checking emails → Checking schedule

→ Making phone calls → Checking IMs

Its DFA path is as follows:

0 → 2 → 5 →

which stops at the nonaccepting state 5, and a security action is taken. A typical action is to ask the user to enter a password (as shown in Figure 7.a) before he/she is allowed to continue the operations. If the password submitted is incorrect, the interface in Figure 7.a stays. Otherwise, the system displays the interface in Figure 7.b, which allows the user to decide whether to continue applying the methods, in case the user does not want the proposed methods to keep interrupting his/her works. The interface in Figure 3.a is then displayed and the user resumes his/her operations.



(a) (b)

Figure 7 (a) A security alert after detecting suspicious handheld uses and (b) user entry of whether to continue using this method

This method is convenient and effective. These searching automata are efficient; they examine each user action exactly once. The time used—after the automaton is built—is therefore $O(m)$ where m is the number of user actions in the sequence. However, the time used to build the automaton can be large. One major disadvantage of this method is the accuracy problem. For example, many times the owner's operations were interrupted because he/she was trying new functions/patterns. Sometimes presumed unlawful uses were undetected because the usage DFA includes too many patterns.

CONCLUSION

Mobile commerce is a promising trend of commerce, and mobile handheld devices are the mandatory tools for performing mobile commerce transactions. However, handheld devices may be easily lost or stolen because of their small sizes. When people lose their devices, they worry that private data such as messages and addresses stored in the devices may be revealed to strangers. This research proposes a novel approach of handheld usage pattern identification for protecting handheld data.

Handheld usage data is first collected, and usage patterns are then discovered and saved. When an unusual usage pattern such as an unlawful user trying to access the handheld data is detected, the device will automatically lock itself down until a further action like entering a correct password is taken. The proposed method includes the following advantages:

- Convenience: It is convenient compared with the password-based method because no user intervention is required after the usage DFA is built unless suspicious actions are detected.
- Efficiency: The time used for the usage finite automaton construction could be long, but once the automaton is built, the usage identification is efficient, and the automaton could be used for a long time until the users want to change the patterns.
- Accuracy: The accuracy of owner identification should be higher than fingerprint or retina recognition, which is still in an early stage of development. Further experiments may be required to support this claim.
- Flexibility: The users are able to adjust the level of security by trying various durations of data collection or entering different threshold values.

Experimental results show this method is effective and convenient for handheld data protection. However, the accuracy problem must be solved before it can be put to actual uses. This problem is related to the questions such as

- Usage data collection: How much time should be spent on data collection or how much or what kinds of usage data should be collected?
- Data preparation: What is the frequency threshold value for removing trivial events or duration threshold value for a valid event?
- Usage finite automata: This research uses a deterministic finite automaton to store usage patterns. It is effective, but patterns stored may be too many or may not be optimal, and some usage data, such as frequency, is not used.
- Usage pattern discovery: Will the sequential pattern discovery, the most popular pattern-discovery method, or other pattern-discovery methods outperform our methods? Finding answers for these questions will be our future research directions.

REFERENCES

Adda, M., Valtchev, P., Missaoui, R., & Djeraba, C. (2007). Toward recommendation based on ontology-powered web-usage mining. *IEEE Internet Computing*,

11(4), 45-52. http://dx.doi.org/10.1109/MIC.2007.93

Agrawal R. & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *Proceeding of 1994 Interference Conference Very Large Data Bases (VLDB'94)* (pp. 487-499). Santiago, Chile: Morgan Kaufmann.

Agrawal, R. & Srikant, R. (1995). Mining sequential patterns. *Proceeding 1995 International Conference Data Engineering (ICDE'95)* (pp.3-14). Taipei, Taiwan: IEEE Computer Society. http://dx.doi.org/10.1109/ICDE.1995.380415

Aho, A. V., Lam, M. S., Sethi, R., & Ullman, J. D. (2006). Chapter 3. *Compilers—Principles, Techniques, and Tools* (2nd ed., pp. 109-190). Boston, USA: Addison-Wesley. http://dx.doi.org/10.1109/MIC.2007.93

Argyroudis, P. G., Verma, R., Tewari, H., & D'Mahony, O. (2004). Performance analysis of cryptographic protocols on handheld devices. *Proc. 3rd IEEE International Symposium on Network Computing and Applications* (pp. 169-174). Cambridge, MA: Springer-Verlag. http://dx.doi.org/10.1109/NCA.2004.1347774

Canalys. (2007). 64 Million Smart Phones Shipped Worldwide in 2006. Retrieved from http://www.canalys.com/pr/2007/r2007024.htm

Canalys. (2010). Majority of Smart Phones now Have Touch Screens. Retrieved from http://www.canalys.com/pr/2010/r2010021.html

CBS Interactive. (2004). Gartner says worldwide PDA industry suffers 5 percent shipment decline in 2003 - Top Stories. [BNET]. Retrieved from http://findarticles.com/p/articles/mi_m0NZB/is_2_6/ai_113888610/

Ding, X., Mazzocchi, D., & Tsudik, G. (2007). Equipping smart devices with public key signatures. *ACM Transactions on Internet Technology* 7(1). http://dx.doi.org/10.1145/1189740.1189743

Eirinaki, M. & Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology*, 3(1), 1-27. http://dx.doi.org/10.1145/643477.643478

Gartner. (2005). *Gartner Says Mobile Phone Sales Will Exceed One Billion in 2009*. Retrieved from http://www.gartner.com/press_releases/asset_132473_11.html

Gartner. (2006). *Gartner Says Worldwide PDA Shipments Reach Record Level in 2005* . Retrieved from http://www.gartner.com/it/page.jsp?id=492242

Gartner. (2007). *Gartner Says Worldwide PDA Shipments Top 17.7 Million in 2006*. Retrieved from http://www.gartner.com/it/page.jsp?id=500898

Gartner. (2008). *Gartner Says Worldwide Mobile Phone Sales Increased 16 Per Cent in 2007.* Retrieved from http://www.gartner.com/it/page.jsp?id=612207

Gartner. (2009). *Gartner Says Worldwide Smartphone Sales Reached Its Lowest Growth Rate with 3.7 Per Cent Increase in Fourth Quarter of 2008* [Gartner]. Retrieved from http://www.gartner.com/it/page.jsp?id=910112

Gartner. (2010). *Gartner Says Worldwide Mobile Phone Sales to End Users Grew 8 Per Cent in Fourth Quarter 2009; Market Remained Flat in 2009*. Retrieved from http://www.gartner.com/it/page.jsp?id=1306513

Gartner. (2011). *Gartner Says Worldwide Server Shipments Grew 7 Percent; Revenue Increased 5 Percent in the Third Quarter of 2011*. Retrieved from http://www.gartner.com/it/page.jsp?id=1859415

Ghosh, A. K. & Swaminatha, T. M. (2001). Software security and privacy risks in mobile e-commerce. *Communications of the ACM*, 44(2), 51-57. http://dx.doi.org/10.1145/359205.359227

GsmServer. (2004,). *Mobile Phone Sales in 2003*. Retrieved from http://gsmserver.com/articles/sales2003.php

Hewlett-Packard Development Company, L.P. (2005). *Wireless Security*. Retrieved from http://h20331.www2.hp.com/Hpsub/downloads/Wireless_Security_rev2.pdf

Hu, W.-C., Yang, H.-J., Lee, C.-w., & Yeh, J.-h. (2005). World Wide Web usage mining. In J. Wang (Ed.). *Encyclopedia of Data Warehousing and Mining* (pp. 1242-1248). Hershey, PA: IGI Global. http://dx.doi.org/10.4018/978-1-59140-557-3.ch234

IDC, Task Force. (2008). *Handheld Devices Sink 53.2% During Fourth Quarter But Protracted Decline Appears to Be Slowing, Says IDC*. Retrieved from http://www.idc.com/getdoc.jsp?containerId=prUS21083408

Kanellos, Michael. (2006, February 10). Mobile Phone Sales Pass 800 Million *CNET*. Retrieved May 12, 2009, from http://news.cnet.com/ Mobile-phone-sales-pass-800-million/2100-1039_3-6037984.html

Kejariwal, A., Gupta, S., Nicolau, A., Dutt, N. D., & Gupta, R. (2006). Energy efficient watermarking on mobile devices using proxy-based partitioning. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 14(6), 625-636. http://dx.doi.org/10.1109/TVLSI.2006.878218

Mobasher, B., Cooley, R., & Srivastava, J. (2000, August). Automatic personalization based on Web usage mining. *Communications of the ACM*, 43(8), 142-151.

Spooner, John G. (2003, November 7). Gartner Ups Estimate for 2003 PC Shipments. *CNET*. Retrieved May 12, 2009, from http://news.cnet.com/ Gartner-ups-estimate-for-2003-PC-shipments/ 2100-1003_3-5104019.html

Susilo, W. (2002). Securing handheld devices. *Proceeding of 10th IEEE International Conference Networks* (pp.349-354). Grand Copthorne Waterfront, Singapore: IEEE Computer Society. http://dx.doi.org/10.1145/345124.345169

Sybase Inc. (2006). *Afaria—The Power to Manage and Secure Data, Devices and Applications on the Front Lines of Business*. Retrieved from http://www.sybase.com/files/Data_Sheets/Afaria_overview_datasheet.pdf

Zhou, B., Hui, S. C., & Chang, K. (2006). Enhancing mobile web access using intelligent recommendations. *IEEE Intelligent Systems and Their Applications*, 21(1), 28-34. http://dx.doi.org/10.1109/MIS.2006.5