Utilizing Document Classification for Grooming Attack Recognition

Dimitrios Michalopoulos
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
dimich@uom.gr

Ioannis Mavridis
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
mavridis@uom.gr

Abstract— **Grooming attack recognition is a complex issue that is difficult to address using simple word matching in order to identify potential hazard for minor users. In this paper, the utilization of document classification to create patterns from real dialogs is proposed. Furthermore, a decision making method that results in generating proper warning signals based on the classification results is introduced. The decision making method is then applied using the best ranked algorithm with a comparative evaluation which conducted on seven document classification algorithms.**

Keywords-component**: Grooming Recognition; Document Classification;* **Naive** *Bayes; Cyber-Crime; Sexual Exploitation

I. Introduction

As the World Wide Web (WWW) has penetrated well into social living and connected people in different geographic regions, it has also elevated new forms of communications. Such forms include from instant messaging and chat rooms to social networking (e.g. Facebook) and blogs. During this communication procedure, concern has been how to secure this virtual interaction and thereby increase safety of the individual [1, 2]. The reason behind this concern lies on the fact that Internet provides applications where users can masquerade themselves and act malicious based on personal interests. Pedophilies, people who feel emotionally and sexually attracted with children, use this anonymity to attract victims that either have not sense of who they interact with neither the scope of interaction [3, 4]. A term that is used to describe such malicious actions with a potential aim of sexual exploitation or emotional connection with a child is referred as “Child Grooming” or “Grooming Attack” [5]. The effects from such actions not only could distract a child from a normal behavior [6] but also cause psychological damages that may never cure in during time [7]. Such effects can be divided in two main categories: the psychological and the physical damage; with the first to be the most difficult to handle [8].

Parents concern as the hazards their children are exposed to while they are online is in increasing rate [9]. Parental control software, which monitors Internet traffic and detects potential misuse, is regarded as a temporary solution that provides static security control. The characteristic of such products is the parental authorization which enables a warning signal or blocks the connection to the Internet when a text or web content is found to be inappropriate. However, there are significant disadvantages of using such kind of tools. On the one hand, such software interrupts with minors’ communication privacy. On the other hand, simple word matching does not provide accurate recognition of potential hazards. For example, an automatic attack recognition system is much harder to understand the meaning of a dynamic dialog instead of a static text document. Indeed, during a written communication, the content of the conversation cannot be efficiently extracted from each word separately but usually based on a type of summary of exchanged words and phrases. Actually, instead of monitoring, parental control software systems block chat conversations containing inappropriate words and suffer from lack of identifying alternative ways of managing forbidden WWW material.

Related research work in the field of protecting minors from WWW hazards by the means of technical solutions is quite limited. Specifically, in [10] the authors proposed an approach based on integration between communication theory and computer science methodologies for protecting minors from online hazards. They also proposed an analysis of the stages that a predator follows through a grooming procedure, which consists of three major stages: gain access to the victim, involve the victim in a deceptive relationship and launch and prolong a sexually-abuse relationship. Similarly, the authors of [11] collected chat dialogs, separated those of victim from those of predator, and created word unigrams, bigrams and trigrams. Subsequently, they processed them with the classification algorithms Support Vector Machine (SVM) and k-nearest neighbors (k-NN) and concluded that the k-NN algorithm with k equal to 30 provides the most effective classification resulting in a value of f-measure that equals to 0.943. Moreover, using natural language processing (NLP) the authors of [12] created a chat corpus which can be used for complex NLP applications.

In the context of our current research work, which is mainly focused on protecting young Internet users from online hazards [13], we are developing a system capable of recognizing grooming attacks, called GARS (Grooming Attack Recognition System). GARS will monitor Internet based dialogs sourced from instant messaging, chat or social network traffic in full respect of communication privacy. In this paper, we propose a decision making method to be used for recognizing potential grooming threats by extracting information from captured dynamic dialogs. However, extracting content information is much more complicated in dynamic chat dialogs rather than in static documents. Therefore, we utilize document classification [14] to create pattern bases from real dialogs in English language. To decide on the most appropriate document classification algorithm for grooming recognition, we conducted a comparison evaluation of seven different algorithms using specific criteria and based on a rich set of real captured dialogs.

The structure of the paper is as follows. Section 2 presents the decision making method and the two possible topologies of the grooming recognition system. The comparative evaluation of the document classification algorithms is presented in section 3. The application of the proposed decision making method with the best ranked document classification algorithm is demonstrated and discussed in section 4. The paper concludes in section 5.

II. GROOMING RECOGNITION SYSTEM

A. *Decision making method*

The proposed decision making method makes use of the following three classification classes, which are similar to [10] and contain dialog parts that:

- Gaining Access (*ga*): indicate predators intention to gain access to the victim
- Deceptive Relationship (*dr*): indicate the deceptive relationship that the predator tries to establish with the minor, and are preliminary to a sexual exploitation attack.
- Sexual Affair (*sa*): clearly indicate predator's intention for a sexual affair with the victim.

The classification process, which is utilized in the context of the decision making method, computes the probability that a captured dialog belongs to each one of the above classes. Additionally, the grooming recognition system has actually to decide on sending a warning signal or not, indicating each decision with 'Yes' or 'No', respectively.

The next step is mapping impacts considering each decision. Wrong estimated impacts could lead to the undesired situations of either undetected sexual exploitation attempt, which would make the system ineffective, or false warning signals, which would make it irritating and unreliable. Function *I(x,d)* in table 1, with *x* standing for each classification class (*ga*, *dr* or *sa*) and *d* for each decision ('Yes' or 'No'), yields the impact of each decision. Impacts are evaluated with numbers scaled from 0 to 9 as depicted in table 1. For example, the impact of being in the *ga* class and sending a warning signal is estimated to 6 (as annoying), whereas for the same class and not sending a signal the impact is 0.

TABLE I. MAPPING IMPACT VALUES

I(*ga*, 'Yes')	6
I(*ga*, 'No')	0
I(*dr*, 'Yes')	5
I(*dr*, 'No')	2
I(*sa*, 'Yes')	6
I(*sa*, 'No')	9

Impact values are mapped with the following declarations. The class that is most possible of leading in a sexual exploitation attack is the *sa*. Indeed, most of the words that indicate this class are rude, containing sexual innuendo. The other two, and especially the *ga*, indicate the conversation that predators use without revealing any malicious intention but collect information about the potential victim [9]. Therefore, it is less possible for the *dr* to lead to a grooming attack and even less for the *ga*. Furthermore, normal dialogs that minors write can indicate *dr* or *ga* categories by false. This false categorization is less possible to be categorized in *sa* class.

The grooming recognition system decides according to the lowest value of each decision on equation (1), which defines the risk function *U(d)* used for this purpose:

$$U(d) = P(ga)I(ga,d) + P(dr)I(dr,d) + P(sa)I(sa,d) \quad (1)$$

Where:

- *P(sa)*,*P(dr)*,*P(ga)* are classification probabilities, results of the document classification process,
- *I(ga,d)*, *I(dr,d)*, *I(sa,d)* are the impact values,
- *d* stands for each decision ('Yes' or 'No').

Figure 1 below presents distributions for U('Yes') and U('No') for all classification probabilities. Values in vertical axis present the risk value whereas values on horizontal axis define the probability distribution for all possible probability values of the three classes.

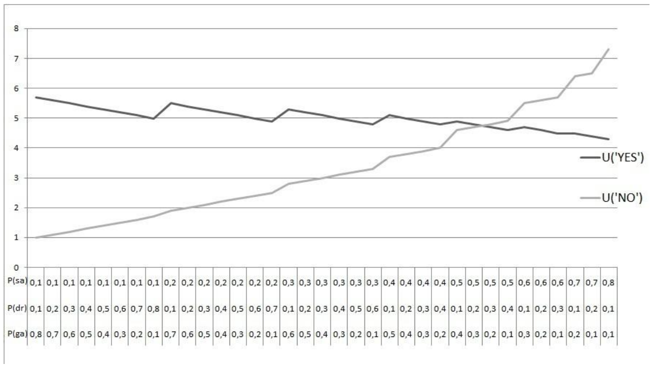

Figure 1. *U('YES')* and *U('NO')* distribution diagram.

Therefore, the warning signal will be sent when the *U('Yes')* is lower than the *U('No')*. As depicted in figure 1, a warning signal is sent when $P(sa) \geq 0.5$.

Considering that hazards for minors vary through age and sex, impact values in table 1 are adjustable controlling that way the sensitivity of the system. Actually, a more sensitive system is needed to minors under 13 years old rather than minors aged 16 to 18 years where a less sensitive system is needed. For instance, by changing the impact values in table 1 for *I(sa,'No')* and *I(dr,'No') to 10 and* 3, respectively, from equation (1) it can be extracted that the warning signal is sent when $P(sa) \geq 0.4$ and $P(dr) \geq 0.4$ resulting in a more sensitive system. Similarly, by mapping *I(sa,N)* = 7 in table 1, the warning signal is sent when $P(sa) \geq 0.6$ and $P(dr) \geq 0.3$ resulting in a less sensitive system. The distributions of *U('Yes')* and *U('No')* for the less sensitive system are presented in figure 2.

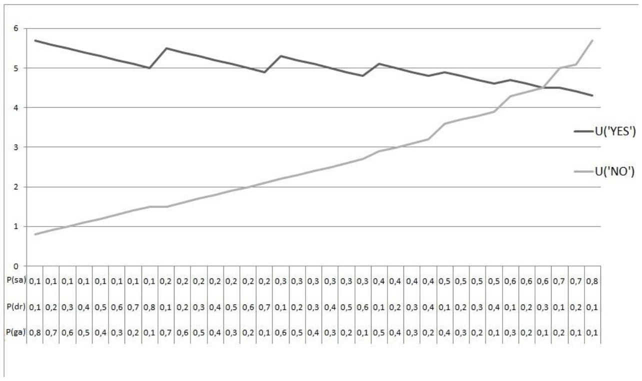

Figure 2. *U('YES')* and *U('NO')* distribution diagram of a less sensitive system.

B. *System topologies*

The proposed grooming recognition system can be implemented in two proposed topologies: the local and the distributed one. The main difference is the location where the recognition process takes place. Both topologies can be implemented either on a desktop or laptop computer to monitor instant messaging or e-mail conversations, or on a mobile device to monitor Short Message Service (SMS) messages.

On the one hand, in a local system topology the whole system is installed at the client's device. However, in order for the GARS system to operate transparently, there is a demand for increased system resources. On the other hand, in a distributed topology, clients send the captured communication data directly to the server, wherein the recognition process is performed. In case of a potential grooming attack, the alarm signal is sent directly from the server to the parent's device. The advantages of the distributed topology consist of the following ones:

- There is no need to install additional software at client side, which demands for extra processing power. This is important especially in mobile devices with low processing power, although client software is needed on mobile devices for capturing dialogs and sends them to servers.
- Any updates or upgrades at recognition patterns or software can be implemented directly on the server, as opposed to each client separately in case of local topology.
- Central monitoring of distributed topology provides the opportunity to clients to exchange information about potential threats. Such a process of combining information from different sources can result in providing early and more accurate detection. For example, in case a predator is detected through communication with one client, the rest of clients can be informed about predator's attempts to establish communication with another client.

However, there are disadvantages as well:

- Distributed topology demands for network connectivity. However, whenever our system is implemented on a mobile device, network coverage is not always available. In addition, cell phone providers charging policies usually increase the cost of mobile device implementation.
- The transmission of captured dialogs to servers includes a security risk of personal data loss. A malicious user may install network sniffer and capture the data that are transmitted from the client to the server. This security risk can be reduced by implementing efficient cryptographic algorithms.
- Recognition performance depends on network conditions. Potential network failure or high round trip response time can obstruct the whole process or even interrupt it.

III. Comparative Evaluation

The comparative evaluation of seven document classification algorithms in order to identify the most suitable for grooming recognition was conducted as follows. The tested algorithms are [15] : k-nearest neighbors (k-NN), Naive Bayes, SVM, tf-idf, Maximum Entropy, EM and EMSIMPLE. Given the focus on real time transparent classification [13], we identified the following evaluation criteria:

- Accuracy: Grooming recognition requires high accuracy in classification. Thus, accuracy is the most important criterion.
- Speed: Having in mind that classification process is running in parallel with a chat dialog, only few seconds are available after each incoming dialog for classification before the new text arrives and demand classification again.
- Low demand on system resources: Classification process should run transparently with other system processes and should not slow down the whole system performance.

A. *Test conditions and preparation*

The first step included the creation of the classification patterns. For that purpose we used 73 dialogs from perverted-justice website material [16]. This website consists of volunteers who enter chat rooms as underage individuals and if they become manipulated for sex by others, they cooperate with police and drive predators to justice. In addition, visitors can rank published dialogs according to how malicious they appear. The extracted dialogs include both homosexual and heterosexual exploitation. Similarly, dialogs include results of (fake) victim attacks. All dialogs were parted and these parts were sorted into three categories that correspond to the classification classes mentioned above (gaining access - *ga*, deceptive relationship - *dr*, sexual affair – *sa*).

After analyzing the dialogs, we found that predators leading a conversation, choose the subject and most of the times force the victim to answer unethical questions. Victims in most of the cases follow the conversation subject with typical answers like 'yes', 'no' , 'maybe', 'we will see'. Therefore, in an attempt to simplify the classification procedure only predators' chats were analyzed. Furthermore, another issue considered was the matters of natural language processing because the captured dialogs are not in formal English but in some cases idioms and sort expressions are used. For example the 'asl' refers to 'age sex location' used in the begging of a dialog. At this point, such idioms and expressions are analyzed having in mind that this classification procedure is towards grooming recognition from chat dialogs, as opposed to formal text. Only slight grammatical errors were corrected, like 'sholl' to 'school', for example.

The software that was used is a Rainbow classification tool [17]. This is a front to end document classification tool that performs instant classification of a given text. It was developed in C language and can be executed in most operating systems (that support console C compilers). In our test it was installed on FreeBSD 8.1-PRERELEASE implemented on i386 CPU architecture.

As a result, 219 text files were created (73 for each class, 3 for each dialog), indexed into the three classes. During the tokenization process, alphabetic sequences of characters were transformed in tokens [17], except those that were added to the stop list. The stop list includes 524 common words which have no importance for classification, as for example the words ‘the’, ‘is’, 'it' and similar.

B. Particular tests

The aim of the comparison of document classification algorithms is to identify the most effective for grooming attack recognition, according to the three criteria defined below. Therefore, a number of tests were performed, as follows.

Documents have been collected randomly from classes, initially 10% from each class, then 20% until finally 80% of the existing document set. The extracted documents were then classified having as classification base those that have not been extracted in each class. The classification tests were performed 2000 times for each test (10 to 80%) and the classification results were statistically collected and reported.

In particular, we got 219 documents to examine, 73 for each class. Initially, 20 of them (approximately 10%) were taken out and classified using as pattern the rest of them (199). The test was performed 2000 times for each algorithm. Test result is an average percentage which represents the successfully classified documents in the class from which they were taken off. Afterwards the test was repeated for 40 documents (approximately 20%) until the final 170 documents (approximately 80%).

a) k-NN

The above tests were performed for the k-NN classification algorithm from k=2 to k=34 [8]. The main result derived from the application of the algorithm is that the deviation between results is narrow. Indeed, the best results are with k=24 with average percentage of 95.12625%.

b) tf-idf

The tf-idf algorithm [18] was tested scoring an average 94.7725%. Specifically, there was medium deviation in scoring between tests.

c) SVM

The Support Vector Machine (SVM) classification algorithm is considered popular and widely used [19]. In tests, information gain function was used for setting the weights of each documents’ words [20]. The average percentages were 71.53% far lower than the other algorithms. Indeed, the lowest percentages are surprising far lower than the other algorithms reaching 27.57% and the deviation between test results was the highest among all algorithms. Moreover, the time needed for performing the above tests was much more than the other methods. Clearly, SVM is not suitable for grooming attack recognition.

d) Naive Bayes

The naive Bayes classification algorithm [21] was tested having the highest average scores. In algorithm details, words that occurred 7 or less times in all documents were ‘smoothed’, meaning that the probability of occurrence were not zero. In addition equal prior probabilities were set for each class calculating information gain and scoring. The average percentages were the highest than all other algorithms achieving 96.02% with narrow deviation between results. In addition, tests with naive Bayes were the fastest than all in terms of time.

e) Maximum Entropy

The Maximum Entropy classification algorithm [22] was tested resulting average percentage of 90.51% with narrow deviation between tests.

f) EM and EMSIMPLE

EM and EMSIMPLE [23] classification algorithms were tested resulting and average percentage of 95.13% for EM and 95.04 for EMSIMPLE with narrow deviation between tests. Besides, classification speed was also remarkably fast.

C. Compatison results

Having analyzed the results of the performed classification tests, we conclude that the most appropriate document classification algorithm for grooming attack recognition is the naive Bayes. Naive Bayes, not only has the highest percentages of accuracy, but also is the faster classifier than all the others. Alternatively, EM achieved high percentages results, even higher than these of k-NN. k-NN is classified in third place. Similarly, EM SIMPLE is in the fourth position. tf-idf is in fifth place and Maximum Entropy comes sixth with high classification speed but lower accuracy results. In contrast, SVM had the lowest percentages of accuracy while it required much more time for classification. The latter is crucial for grooming attack recognition because there is a demand for real time analysis of Internet dialogs. Therefore, the classification process has to be performed instantly in order the decision making algorithm to decide if a warning signal should be sent or not. All the algorithms expect from SVM, were fast enough with Naive Bayes in first place slightly above the others. In contrast, the SVM was the slowest. naive Bayes performed each test from the 2000 tests in less than 10 minutes, EM required approximately 12-13 minutes, EM SIMPLE and maximum entropy approximately 15 minutes, tf-idf and k-NN and the SVM algorithm over 1 hour. All tests were performed in the same computer equipped with Intel Pentium III processor. Final comparison results are presented in table 2.

TABLE II. COMPARISON RESULTS

Rank	**Algorithm**	**Accuracy (%)**	**Classification time**
1	Naive Bayes	96.0275	9 min 38 sec
2	EM	95.1375	12 min 14 sec
3	k-NN	95.12625	14 min 58 sec
4	EM SIMPLE	95.04625	12 min 2 sec
5	tf-idf	94.7725	15 min 9 sec
6	Maximum Entropy	90.5125	13 min 22 sec
7	SVM	71.53714	62 min 19 sec

IV. EXPERIMENTAL APPLICATION

In this section we demonstrate an experimental demonstration of the proposed decision making by utilizing the Naive Bayes as the most suitable document classification algorithm.

The Rainbow classification tool [17] is an open source software that we adopted to be used as part of the GARS system, due to the fact that it demands for low system resources. Patterns are the same with 73 dialogs which was created for the classification test. The classification process computes the probability of each class and then the decision making method concludes on sending or not a warning signal. For example, the phrase ‘what did you wear at school today’ returns classification results:

- *ga* 0.5849047392
- *dr* 0.4044349669
- *sa* 0.01066029387

This phrase can be written in a conversation between a predator and a victim, however it can be written in a conversation between two teenagers. From equation (1) using impact values in table I, decision method results in sending a warning signal. In contrast, the phrase ‘I want to lick you’ returns classification results:

- *sa* 0.889622576
- *ga* 0.06260139678
- *dr* 0.04777602723

The classification results of the second phrase denote that the probability of *sa* class is approximately 0.89 and similarly the decision becomes to send a warning signal. Indeed, the warning signal even if the adjustment is for less sensitive system (Figure 2).

Warning signal’s form is another critical aspect. As our system works transparently to minor user, any warning signal is directly sent to the parent who is full responsible of child’s protection. The direct warning signal address one drawback of parental control software which blocks online communication in case of a detected threat as it is discussed previously. In addition, in respect of communication privacy, warning signals should not contain any sensitive information about the conversation or the content, just the results of the recognition method as it is described in this work.

Nevertheless, the demonstrated application revealed a number of drawbacks, as follows. First of all, the decision making system is memoryless. In case one user’s dialogs generate multiple warning signals, current system implementation does not support feedback mechanisms for putting that user to a quarantine space. Therefore, other minor users who are monitored by this application will not be informed about the potential threat. In addition, classification patterns are dedicated for dialogs in English language. If GARS is needed to be used with another language, patterns in that language need to be created using parts from real dialogs that lead to sexual exploitation attack.

V. CONCUSIONS

Various document classification algorithms can be utilized for the purposes of grooming attack recognition. According to the comparative evaluation we conducted and presented in this paper, we concluded that Naive Bayes is the most appropriate, not only by reaching the highest average classification score of 96%, but also by being the fastest than all. In contrast, SVM reported the lowest than all average classification score. In parallel, the SVM algorithm was found to be the slowest than all the others and demanding for more processing time and computer resources as well.

As a future work extension, we intend to increase GARS performance and accuracy by implementing, except from document classification, more processes like emotion recognition. Similarly, we intend to increase the pattern base including more dialogs. Furthermore, we intend to test our system with various gender and age groups (e.g. males under 13 or girls over 13) and research on the most effective sensitivity adjustment of the decision method on each group, as described in section 2.

REFERENCES

[1] K. Nash. "A Peak inside Facebook". Available: http://www.pcworld.com/businesscenter/article/150489/a_peek_inside_facebook.html accessed 4 Octomber 2010.

[2] "Billions Connected, Global Instant Messaging Market Share – Open Data". Available: http://billionsconnected.com/blog/2008/08/global-im-market-share-im-usage/ accessed 9 November 2010.

[3] R. J. Estes. "The Sexual Exploitation Of Children". Center for Youth Policy Studies. Available: http://www.sp2.upenn.edu/~restes/CSEC_Files/CSEC_Bib_August_2001.pdf accessed 9 November 2010.

[4] C. Marcum, "Sexual Predators: How to recognize them on the Internet and on the street -- how to keep your kids away" Silver Lake Publishing 2007 ISBN:1-56343-794-5.

[5] C. Crosson-Tower, "Understading child abuse and neglet" Addison-Wesley 2008 6th edition ISBN 020540183X.

[6] "Child Sexual Abuse", US National Library of Medicie Available: http://www.nlm.nih.gov/medlineplus/childsexualabuse.html accessed 15 Octomber 2010.

[7] I. Berson, "Grooming Cybervictims: The Psychosocial Effects of Online Exploitation for Youth," Journal of School Violence, vol. 2, 2003.

[8] J. Briere, "Methodological issues in the study of sexual abuse effects," Journal of Consulting and Clinical Psychology, vol. 60, pp. 196-203, 1992.

[9] L. N. Olson, J.L Daggs, B.L. Ellevold, and T.K. Rogers, "Entrapping the Innocent: Toward a Theory of Child Sexual Predators Luring Communication," Communication Theory, vol. 17, pp. 231-251, 2007.

[10] A. Kontostathis, L. Edwards, A. Leatherman, "Text Mining and Cybercrime," in Text Mining: Application and Theory, E. Michael W. Berry and Jacob Kogan, Ed., ed: John Wiley & Sons, 2009.

[11] N. Pendar, "Toward Spotting the Pedophile: Telling victim from predator in text chats," in IEEE International Conference on Semantic Computing, Irvine California USA, 2007, pp. 235-241.

[12] E. N. Forsyth and C.H Martell, "Lexical and Discourse Analysis of Online Chat Dialog," presented at the Semantic Computing, 2007. ICSC 2007. International Conference on, Irvine USA, 2007.

[13] D. Michalopoulos, I. Mavridis and V. Vitsas "Towards a Risk Management Based Approach for Protecting Internet Conversations," presented at the 9th European Conference on Information Warfare and Security, ECIW 2010, .

[14] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34 (1), pp. 1-47, 2002.

[15] L. Busagala, "Automated Document Classification: Methods and Algorithms" VDM Verlag Dr. Müller, 2010, ISBN: 3639270509

[16] "Perverted-Justice.com Perverted Justice". Available: www.perverted-justice.com accessed 2 Octomber 2010.

[17] A. K. McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classi_cation and clustering," ed, 1996.

[18] H.C. Wu, R.W.P. Luck, K.F. Wong, K.L. Kwok, "Interpreting TF-IDF term weights as making relevance decisions," ACM Transactions on Information Systems, vol. 26, 2008. DOI: 10.1145/1361684.1361686.

[19] T. Joachims, "Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms", 1 ed.: Springer, 2002. ISBN: 079237679X.

[20] R. B. Wells, "Applied Coding and Information Theory for Engineers" Prentice Hall, 1998. ISBN: 9780139613272.

[21] I. Rish, "An empirical study of the naive Bayes classifier," presented at the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001.

[22] Y. Kitamura, "Empirical Likelihood Methods in Econometrics: Theory and Practice" Cowles Foundation Discussion Papers 1569. Available: http://cowles.econ.yale.edu/P/cd/d15b/d1569.pdf accessed 5 Octomber 2010.

[23] C. Gilles and G.Govaert "A Classification EM algorithm for clustering and two stochastic versions," Computational Statistics & Data Analysis - Special issue on optimization techniques in statistics, vol. 14 (3), 1992.