VOLUME-2, ISSUE-1

SREEPATHY

JOURNAL OF COMPUTER SCIENCE & ENGINEERING

Published by

Department of Computer Science and Engineering

Sreepathy Institute of Management and Technology, Vavanoor

Palakkad - 679 533

July 2015

Contents

Maintenance Enabled Stability Enhanced Routing and Channel Assignment for Mobile Ad Hoc Cognitive Networks

Yadu P. Dev
Assistant Professor
Department of CSE, SIMAT, Vavannoor
Email: yadu.polpaya@gmail.com

Govindaraj E.
Associate Professor
M E S College of Engineering
Email: govindaraje@yahoo.com

***Abstract*—The under utilization of licensed spectrum and over crowding of unlicensed spectrum leads to spectrum shortage. Cognitive radio make use of the licensed spectrum by unlicensed users when the licensed users are not using that. Data communication in mobile ad hoc Cognitive Radio Networks(CRN) significantly suffers from link stability and channel interference. Routing in CRN must consider the mobility of nodes and interference of primary users. A lot of researches are being conducted to improve the route stability of CRN routing. Some such innovations are analysed in this work. A routing solution that produces stable routes in cognitive radio network is highly desirable as it is one of the main challenges of CRN due to the PU activities. Unstable routes will lead to frequently firing new re-routing events which consumes the network resources and degrades its performance. A new routing technique which maintains a stable route and channel assignment in multi flow and multi hop mobile ad hoc cognitive network is proposed.**

Keywords: Cognitive Radio Networks (CRN), Channel interference, Route stability, Mobility prediction, Channel assignment, MP-JSRCA.

I. Introduction

Spectrum assignment is based on an auctioning mechanism in which, frequency bands are assigned to users on a long term basis. Spectrum assigned to this users are called licensed spectrum. Due to the under utilization of the licensed spectrum and over crowding in the unlicensed spectrum, there will be spectrum scarcity problem. Cognitive Radio Networks (CRN) [1] will address this problem. In this network, there are two types of users. Licensed users are called Primary Users(PU) and unlicensed users are called secondary users(SU) or cognitive nodes(CN). Primary users have licensed spectrum for their communication. Cognitive radio network make use of this licensed spectrum for the secondary users, when primary user is not using that.

Each SU will search for any free licensed spectrum for the easiness of communication, called spectrum sensing. Set of frequency bands currently available for the secondary users, which is not occupied by primary users is called spectrum opportunity(SOP) or spectrum hole or white space. Dynamically accessing the sensed free spectrum, by changing it's frequency band is called dynamic spectrum access(DSA). But primary users will always have the priority for using the licensed spectrum. So if the primary user comes back for accessing it's spectrum, SU must vacate from that spectrum. SU must not interrupt the transmission of PU.

Two nodes in the traditional ad hoc network can communicate with each other if they are in the transmission range. But in cognitive radio networks, they can communicate only if a common channel is available between them. Difficult from traditional routing protocols in ad hoc networks, routing in CRN has to deal with a number of challenges:

- Traditional network challenges
- Spectrum awareness
- Primary user interference and channel switching
- Signalling and channel deafness problem

Both SU's and PU's are mobile. Routing in cognitive radio network must address the ad hoc nature of the nodes, primary spectrum sensing and DSA. Routing must overcome the traditional network challenges such as node's mobility, node's limited power, network life time, channel scarcity. While routing, each SU must analyse the spectrum in it's vicinity, whether it is available or not and check the type of spectrum. PU is the owner of the spectrum. When it is free, SU can start using that by dynamic spectrum access. As soon as SU detects the transmission of PU on the channel, SU must stop it's transmission and channel switching has to be done. Nodes tuning on one channel can not able to sense the signals transmitted on different channel band. So synchronization must be done or a common control channel(CCC) must be used.

A stable route has to be maintained between two nodes. Instability arises due to the interference of PU's. When interference occurs, a stable route has to be reconstructed. A routing metric [2] is to be selected to maintain quality routes between nodes. A routing solution that produces stable routes in CRN is highly desirable as it is one of the main challenges of CRN due to the PU activities. Unstable routes will lead to frequently firing new re-routing events which consumes the network resources and degrades the performance.

Various methods for stability enhanced routing in CRN has been analysed. A better approach is considered for ensuring stable routing in CRN. The proposed approach is efficient in maintaining a stable route through out the communication.

II. Literature Survey

Researches are always been conducted to improve routing stability in cognitive radio networks by establishing new routing metrics. This chapter briefly presents some of such innovative approaches to maintain quality routes in CRN's.

A. STOD RP

STOD RP [3] is a spectrum tree based on demand routing protocol which establishes a spectrum tree for each spectrum band. Routing is done with the help of a new routing metric and the spectrum tree for each band. The new routing metric reflects the route quality and spectrum availability. Cost of a single link is calculated as follows:

$$\mathrm{C}_i = \left[O_{ca} + O_p + \frac{P_{kt}}{r_i}\right] \frac{1}{1-e_{pti}} \frac{1}{T_i} \qquad (1)$$

- O_{ca} : Channel Access Overhead
- O_p : Protocol Overhead
- P_{kt} : Size of packet
- r_i : Link rate
- e_{pti} : Packet error rate
- T_i : Time duration during which a spectrum band is available to link i

Cost of the end to end route is the summation of the link costs and the switching cost. Spectrum tree is formed for each spectrum band. Node with largest number of available spectrum bands and longest available time of spectrum band is selected as root. Root announcement message is send to other nodes. Every root keeps an inter spectrum nodes list. Nodes that are in two spectrum trees are called overlapping nodes. Route discovery combines tree based proactive routing with on demand route discovery. Source first sends request to the root through the tree and root identifies whether intra or inter spectrum routing has to be performed. Inter spectrum routing is achieved through overlapping nodes.

B. Coolest Path

In this technique [4], stability of the path is determined by frequency diversity and channel stability. Frequency diversity is the minimum number of sub links over all the hops on the path. Channel stability is the probability that the channel is claimed by a PU. PU activities are measured using the metric link temperature [LT], which measures the link stability. It is the primary channel utilization at each node. It is updated by the sensing capability of the SU. Temperature of the channel c between nodes i and j can be calculated as follows, where u_i^c and u_j^c are the channel busyness ratio of channel c caused by primary nodes detected at node i and j respectively.

$$\mathrm{T}_{ij}^c = 1 - \left(1 - u_i^c\right)\left(1 - u_j^c\right) \qquad (2)$$

Link temperature is defined as the temperature of the channel with lowest primary channel temperature between two nodes. Path temperature is the sum of link temperatures. It can defined in different ways. Accumulated spectrum temperature $\left(T_a^L\right)$ is the sum of link temperatures over the entire path. Path with lowest accumulated spectrum temperature is most stable path. Highest spectrum temperature $\left(T_{max}^L\right)$ is maximum link temperature of the links over the path. Path with minimum highest spectrum temperature will be the most stable path. This routing will take the path with low T_a^L and low T_{max}^L. It also uses the Dijkstra's algorithm for finding the minimum temperature path. Path with lowest path temperature is the coolest path, which is the most stable path.

C. Gymkhana

Gymkhana [5] is an approach where the stability of the path is determined using the connectivity of the network path. If the path is connected, it is a stable path. Connectivity is measured with the help of laplacian matrix and eigen spectrum. The second eigen value λ_2 is called the algebraic connectivity, which measures the stability and robustness of a complex network model. λ_2 is the routing metric used in this approach. Each SU measures on each available channel for the activity factor which is measured on the basis of active and silence period of PU. Each SU composes an influence vector based on the activity factors.

Source node broadcasts RREQ to discover all paths towards the destination. RREQ arriving at destination contains two lists, nodes encountered in that path and influence vectors of nodes encountered in that path. RREQ processing is done by creating a virtual graph V_k for each k paths and calculating the eigen values of the laplacian of the virtual graph. Virtual graph is created by associating N_p virtual nodes to each node in the path, where N_p is the number of primary nodes. Horizontal edge is formed if the nodes are consecutive nodes in the path and vertical edge is formed if these two nodes are virtualization of the same node on different nodes.

Laplacian matrix of the virtual graph and it's eigen spectrum is calculated. An utility function U_k to path k is calculated using λ_2 and destination selects the path with highest utility function.

D. Spectrum Aware High Reliable Routing

In this approach [6], two paths primary path and alternate path are created. Transmission starts through primary path. When SU is using the licensed channel in primary path and PU returns back, it waits for maximum waiting time T_m and after that it switches to the alternate path. Path stable time is one of the routing metric used in this routing. Stable time of a channel is the period of time before this channel is

failed. Link stable time is the stable time of a particular link. Multiple channels will be available for a single link and the best channel is selected for the link.

In this protocol, primary path and alternate path are disjoint paths. There are two types of disjoint paths. Node disjoint path does not have any common nodes except source and destination. Link disjoint path does not have any common links. Link divergence degree and node divergence degree are used to find the most disjoint path of the primary path which will act as the alternate path.

E. Minimum Maintenance Cost Routing

In minimum maintenance cost routing [7], route with minimum maintenance cost is selected as the most stable path. Maintenance cost represents the effort needed or penalty paid for maintaining end to end connectivity in dynamic multi hop CRN's. The maintenance of a route may involve link switching and channel switching as PU's become active. In the former case, one or more links along the route must be replaced by other ones which are not interfered by PU. In latter case, the same link can be maintained but the transmission must be carried over to another spectrum portion. In either case, signalling is required to coordinate with other SUs, which translates to a cost in terms of consumed power, and service interruption time while switching routes. Concept of epoch is introduced in this technique. Epoch is time period during which there is no change in PU activities.

Cost for the change of path is calculated first. It can be either link change cost or channel change cost. Route selection cost is the sum of initial route set up cost and all transition costs. Route selection variables and link cost variables are used for this calculation.

III. Joint Stable Routing and Channel Assignment

The stable routing is a fundamental issue in CRN's because a stable route can significantly reduce packet loss, network latency, and overhead due to less route reconstruction. Compared with traditional wireless networks, the frequently changing PU activities and CN mobility in CRN's make the optimal route set up more difficult. The connectivity between any pair of neighbouring CNs in CRN's depends on not only their relative movement but also the potential conflict to adjacent PNs and CNs.

In this technique, a mobility prediction-based joint stable routing and channel assignment (MP-JSRCA) [8] approach to solve the above challenging issues, aiming at maximizing network throughput. For the route selection, focus is on the stability of a path to enable the path duration as long as possible. For the channel assignment, consider an optimal solution that completely avoids the impact to PNs and at the same time, mitigates intra- and inter-flow interferences among CNs using minimal number of available channels.

A. Reference Network Model and Assumptions

A CRN is modelled as an undirected graph G = (V, E) where V is the union of the CN set and the PN set such that $V = V_c \bigcup V_p$. Number of primary nodes is M and cognitive nodes is N. M is very much lesser value than N. E is the union of the set of links among CNs and the set of links among PNs such that $E = E_c \bigcup E_p$. CN moves according to the Random Waypoint (RWP) model. PN does not communicate with any CN. Different CNs may have different available channels that change over time and over locations while each PN holds only one licensed channel. The available channel set of a CN u is defined as $C_u(t) = \{c_i^u | i = 1, 2, ...c_u(t)\}$ in a time slot t. A pair of CNs u and v can communicate only on a common channel $c_{u,v}(t) \epsilon C_{u,v}(t) = C_u(t) \bigcap C_v(t)$.

Every channel has the same transmission Range, R_T and interference Range, R_I. For any pair $u, v \epsilon V_c$, $d_{u,v}(t) = |u - v|$ denote the Euclidean distance between u and v in a time slot t. A link $l_{u,v}$ is available if, $|C_{u,v}(t)| \geq 1$ and $d_{u,v}(t) \leq R_T$.

B. Channel Assignment Problem

The objective of channel assignment scheme is to prohibit the affect to any PN, to minimize the interference to inter-flows and the intra-flow, and to use the minimal number of orthogonal channels. The main idea is to make full use of both channel diversity and spatial re usability. Nodes (links) involved in a flow as active nodes (active links). Accordingly, other nodes (links) are called as inactive nodes (inactive links).

Let a link $l_{u,v}$ in P_{sk}^{dk} is using $c_{u,v}$. Channels assigned to P_{sk}^{dk} is the set $C_{sk}^{dk} = \{c_{u,v} | u, v \epsilon P_{sk}^{dk}, v \epsilon N(u, R_T)\}$. To avoid the co channel interference, only one link can use a given channel in the λ-hop neighbourhood of the target receiver. λ is set to 2 in this technique.

C. Channel Interference Patterns

In CRN's, there are two categories of channel interferences: intra-flow interference among adjacent active nodes in the same flow and inter-flow interference among different concurrent flows. To avoid the intra-flow interference, assign each link with a channel different from that of two-hop neighbours in the same flow. In the case of the inter-flow interference, however, there are more links that interfere with each other.

- Node Link Interference : When one node is transmitting, adjacent two links in an inter-flow are interfered.
- Node Node Interference : Only two links in different flows interfere with each other.
- Link Link Interference : Two intermediate nodes on different flows (4 links) interfere with each other.
- Cross Flow Interference : CFI interference occurs when two flows cross at an intermediate node, where at most eight links two hops away from the crossed node interfere with each other.

D. Performance Metric

A data transmission in CRN's is possibly interrupted by the two factors: node mobility and channel interference. The latter may be caused by: 1) the appearance of PN communication and 2) cochannel interference among CNs. Accordingly, a new measure metric called data transmission cost (DTC) is developed, by capturing the above two factors as:

$$\text{DTC } (l_{u,v}) = \alpha C_M^{l_{u,v}} + \beta C_I^{l_{u,v}} \qquad (3)$$

where $C_M^{l_{u,v}}$ and $C_I^{l_{u,v}}$ are the mobility cost and channel interference cost of $l_{u,v}$ caused by the node mobility and channel conflict, respectively. From the node mobility point of view, it is preferred that a link between a pair of nodes can keep a longer communication duration. Maximal lifetime $MLT_{u,v}$ is used to measure the stability of $l_{u,v}$. It is defined as the time interval from a route setup until $l_{u,v}$ is lost due to $d_{u,v}(t)\ R_T$. Higher relative movement between u and v results in a shorter $MLT_{u,v}$. Moreover, a link with higher bandwidth can transmit a specified flow within a shorter period of time. So, the mobility cost as follows:

$$\mathrm{C}_M^{l_{u,v}} = \gamma MLT_{u,v} \times w_{u,v}^c \qquad (4)$$

- γ : Adjustable Coefficient
- $MLT_{u,v}$: Maximum Life Time of $l_{u,v}$
- $w_{u,v}^c$: Available bandwidth between u and v

Channel conflict cost $C_I^{l_{u,v}}$ results from the signal interferences to both primary nodes and to cognitive nodes, which can be calculated by

$$\mathrm{C}_I^{l_{u,v}} = C_{I:PN}^{l_{u,v}} + C_{I:CN}^{l_{u,v}} \qquad (5)$$

- $C_{I:PN}^{l_{u,v}}$: Interference of primary nodes
- $C_{I:CN}^{l_{u,v}}$: Interference of cognitive nodes

E. Routing Stability Prediction

Prediction of $MLT_{u,v}$ is based on the random waypoint model. Nodes move in Random Waypoint Model. $d_{u,v}(t)$ is the distance between u and v at time t. If $d_{u,v}(t) = R_T$, t reaches $MLT_{u,v}$. $d_{v,PN_m}(t)$ is the distance between v and PN_m at time t. If $d_{v,PN_m}(t) = R_I$, t reaches MLT_{v,PN_m}. MLT_{v,PN_m} defines how long v will interfere with PN_m.

F. Interference Avoidance to Primary Nodes

Let PN_m hold a licensed channel c_{PN_m}. MP-JSRCA requires that for all cognitive nodes cannot use c_{PN_m} when $d_{v,PN_m} \leq R_I$. A potential conflict set CS_v of a cognitive node v is the set of all primary nodes potentially interfered by v, i.e., $CS_v = \{PN_m | d_{v,PN_m} \leq R_I\}$ at that time. The longest duration that cognitive node v can communicate with u is $MLT_{u,v}$. If $MLT_{u,v} \leq MLT_{v,PN_m}$, $l_{u,v}$ will lose before $d_{v,PN_m} \leq R_I$. Add PN_m to CS_v if, PN_m is communicating on c_{PN_m} and $MLT_{u,v} \leq MLT_{v,PN_m}$. Available channel set (ACS_v) of any CN v can be formulated as $ACS_v = \{c_{PN_m} | \forall PN_m \epsilon CS_v\}$. Cognitive node v can only use a channel in ACS_v. If $ACS_v = \phi$, $C_{I:PN}^{l_{u,v}} = \infty$.

G. Interference-Avoiding Channel Assignment

Channel assignment takes charge of selecting a channel for each next-hop candidate and calculating its channel conflict cost. More specifically, a working node u, which is running MP-JSRCA protocol, assigns a channel $c_{u,v}$ to $l_{u,v}$ in the following approach, according to candidate v's node types (normal and critical nodes) and interference patterns. The focus is how to avoid the impact to any $PN_m \epsilon CS_v$ and to minimize the interference to two-hop CNs. Although different CNs have different available channels in a given time slot t, each CN can make sure available channels of its two-hop neighbouring CNs by means of extended periodical beacon mechanism.

MP-JSRCA assigns channels in two phases: 1) calculation of the common channel set $C_{u,v}$ and 2) assignment of an interference-avoiding channel $c_{u,v}$ to $l_{u,v}$. In the first phase, ACS_u and ACS_v are first calculated. If $ACS_v = \phi$, MP-JSRCA sets $C_{I:PN}^{l_{u,v}} = \infty$ so that v is completely excluded from the route selection and the MP-JSRCA directly transfers to selecting the next candidate. Otherwise, the MP-JSRCA sets $C_{I:PN}^{l_{u,v}} = 0$ and $C_{u,v} = ACS_u \bigcap ACS_v$ and carries out the second phase.

In the channel assignment phase, the MP-JSRCA firstly calculates available common channel set $C_{available}$ through removing channels that have been assigned to the last two hops in f_i from $C_{u,v}$ such that $C_{available} = C_{u,v} - \left\{c_{1hop}^{f_i}, c_{2hop}^{f_i}\right\}$. $c_{khop}^{f_i}$ refers to a channel assigned to the link $l_{k_hop}^{f_i}$, which is in f_i and is k hops away from the working node u in the source side. If a normal node does not have a two-hop or even a one-hop upstream node, $c_{1_hop}^{f_i}$ or $c_{2_hop}^{f_i}$ accordingly set as ϕ.

Critical nodes potentially interfere to both intra-flow and inter-flows. Channel assignment for these nodes also works in two phases. Essentially, the goal of the first phase is to avoid the channel conflict to primary nodes. This phase is similar to that in the channel assignment for normal nodes. The channel assignment phase is more complicated than the above. Firstly, to guarantee that $l_{u,v}$ does not interfere with inter-flows, available common channel set $C_{available}^{j}$ should exclude channels assigned to two-hop links in all inter-flows f_i such

that $C^j_{available} = C_{u,v} - \bigcup_j \left\{ c^{f'_j}_{2hop}, c^{f'_j}_{1hop}, c^{f_j}_{2hop}, c^{f_j}_{1hop} \right\}$. The calculation of $C^j_{available}$ is dependent on interference patterns. A CFI-interference link conflicts to at most four links in each f_i. An LLI-interference link potentially interferes to two links in each f_i. NLI-interference or NNI interference is a special case of LLI.

MP-JSRCA protocol can firstly avoid interferences with existing inter-flows and then considers conflict avoidance to the intra-flow. For this purpose, it sets a high interference cost for the inter-flow conflict ξ and a low interference cost for the intra-flow conflict. As a consequence, a candidate v with the inter-flow interference has a lower probability to be selected as the next hop than the node with only the intra-flow interference.

H. MP-JSRCA Protocol

During route setup, MP-JSRCA protocol carries out its local optimization at each hop towards a destination node, by always selecting the best link for jointly solving the stable routing and channel assignment. The best link between two CN's is the link with the minimal data transmission cost. Considering the node mobility, MP-JSRCA uses on-demand routing mechanism, which in general introduces the flooding. Any node can detect its location and velocity by a GPS device. Furthermore, each node knows location, velocity and available channels of its one-hop neighbors by means of the periodic beacon mechanism through MAC-layer hardware. With a little extension of the beacon mechanism, any node can also make sure available and used channels of two-hop neighbouring nodes. MP-JSRCA selects candidates of the next hop within a controlled sector region with an angle ψ , circled at the working node u, towards the destination. MP-JSRCA can adjust the angle ψ during the discovery of the next hop.

MP-JSRCA protocol runs hop by hop in a distributed fashion. The source node s_k and each relay node directly determines the next hop based on the lowest DTC. MP-JSRCA works in two phases: route discovery and then route acknowledgement, where u and v represent a working node and its next-hop node, respectively. At the beginning, u is set as $u = s_k$. The protocol stops when $v = d_k$.

Route discovery includes next hop selection and message forwarding. Node u calculates the DTC of each neighbouring node located in the sector region, and selects the node v with the lowest DTC as the next hop. During the calculation of DTC of each candidate, MP-JSRCA calls our channel assignment algorithms presented previously to estimate the channel interference cost, based on the node type and the interference pattern of v. Finally, u sends a route selection packet (RSP) to v, which consists of unique packet ID, unique flow ID, 2-D node coordinate, selected subpath , assigned channel , sender and destination. RSP packet is transmitted on the CCC channel.

On receiving a RSP packet, v becomes a new working node . It also selects its next hop in the above way, through running MP-JSRCA protocol, and then sends the updated RSP packet to the next hop. In this way, intermediate nodes carry out the route discovery hop by hop until the destination d_k. After receiving the RSP packet, d_k gets both the stable path P^{dk}_{sk} and assigned channels for every links in P^{dk}_{sk}. It then responds a routing response packet (RRP) with the P^{dk}_{sk} and the assigned channels to s_k via the inverse path of the P^{dk}_{sk}, still on the CCC channel.

IV. Maintenance Enabled Stability Enhanced Routing and Channel Assignment for Mobile Ad Hoc Cognitive Networks

PU interference is the main problem in the routing. At the time of routing, if a SU is using a primary channel and that channel is needed by the corresponding primary user, SU has to exempt that channel. Channel switching or other mechanisms has to be done at that time. Mechanism can be either channel switching or a re establishment of the route.

At present, the source SU will find the next node with lowest DTC and that node finds next node and continues until it reaches the destination SU. At the time of finding the new node, a RSP packet is send from the finding node to discovered node. When RSP reaches the destination, it will send a RRP packet back towards source. When RRP reaches the source, it starts sending the packets through the discovered path to the destination. Channels will be already jointly assigned in the previous phase itself. At the time of packet sending, nodes will not check whether the channel is free or not. If it is not free, communication has to be stopped there and channel must be exempted to PU. So inclusion of a maintenance mechanism will solve all the issues related to the PU interferences.

A. Waiting Time Maintenance Mechanism

When source starts sending the packets to the destination, each node must check whether the assigned primary channel is free or not. If the channel assigned is free, communication can be carry forwarded to the next node. If all the channels assigned for SU's are free, packets can be transmitted through the discovered path itself. At some SU, if the assigned channel is not free, ie. using by the corresponding primary user, that channel cannot be used by the SU and packet sending must be stopped immediately.

If such a situation occurs, the corresponding SU will send a route suspend message (RSUS) to the source through the upward path. When source receives the RSUS message, it will stop packet sending. The SU which sends the RSUS message will wait for a fixed time T_m. If within that time, the PU stops using the channel, the SU can continue using that channel. So SU will send a route continue (RCON) message to the source through the upward path. When RCON reaches the source, packet will send through the same path as discovered earlier.

If within the time T_m, the PU does not stop using the channel, the SU will send a route dismiss (RDIS) message to the source through the upward path. When RDIS reaches

the source, source will start the discovery of a new route to the destination using the same way as firstly done.

B. Waiting Time Algorithm

The detailed algorithm for the maintenance mechanism is as follows:

Algorithm 1 Waiting Time Algorithm

1: Initially, source node starts sending the packets in the discovered path
2: **for** each node in SU
3: **if** assigned channel is free
4: Send packets through the discovered path.
5: **else**
6: Send RSUS message back to source in the reverse path
7: Wait for a maximum time T_m
8: **if** channel becomes free within T_m
9: Send RCON to source and send packets in the previous path itself
10: **else**
11: Send RDIS message back to source and source rediscovers the path
12: **end if**
13: **end if**
14:**end for**

V. Results and Discussions

A. Simulation Parameters

The project is implemented using NS2. Nodes are mobile. There are 5 primary users and 25 secondary users in the network. Path or the next node from the current node is calculated based on the DTC value. Channels are assigned jointly assigned at the time of routing itself.

TABLE I. Network Simulation Parameters

Parameter	Value
Number of Nodes	30
Simulation area	500 x 500
Node type	Mobile
Traffic	CBR

VI. Results

The proposed maintenance enabled stability enhanced routing is compared with the existing joint stable routing. In the modification, a new algorithm is added with the present routing method for maintenance support. Basic routing procedures in the modification is same as that of joint stable routing. Throughput and packet delivery ratio are taken as the parameters.

Once the modification is done it is found that the packet delivery ratio has increased from that of the stability enhanced routing. In joint stable routing, at the time of packet sending, there is a chance for the PU channel is using it. It is not considered by the routing technique. So some packets

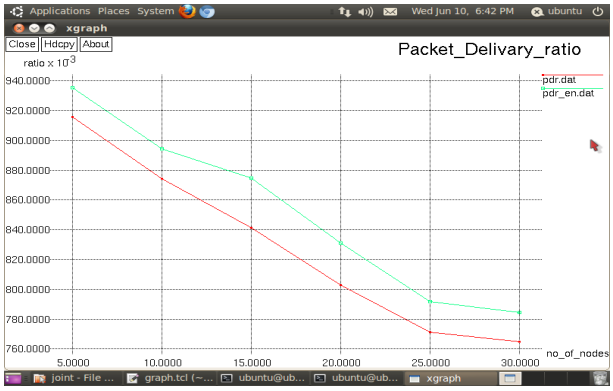


Fig. 1. Packet Delivery Ratio in Transmission(Existing versus Proposed)

may drop because of this reason. But in the modification, maintenance mechanism is included. So if the channel is not available, it will wait for some time to re avail it and send the corresponding packet. Therefore no packets become drop. And if within the waiting time, if the channel is not free, rediscovery takes place and packets are send again. So no packet loss will be happening. Fig 6.2 shows comparison of the packet delivery ratio between the existing joint stable routing and maintenance enabled joint stable routing.

Maximum data that can be send through the channel is defined as the throughput of the network. If the channel is not using by the primary user, then maximum data can be send through the channel and maximum throughput can be achieved. But if PU uses it's channel, throughput will be decreased. In the joint stable routing, at the time of packet sending, it never checks whether the PU using it or not. If it is using, the throughput decreases strongly because of the packet loss. But as a maintenance mechanism is included, the things become different.

Maintenance mechanism checks whether the channel is free or not at the time of packet sending. If the channel is free, throughput is maintained in the same level. If the PU channel is not free, it stop sending the packets. So that no packets are dropped. SU waits for a maximum waiting time and if the PU channel frees within that time, packets are transmitted. So it does not effect the throughput. Fig 6.3 shows comparison of the throughput between the existing joint stable routing and maintenance enabled joint stable routing. From the graph, it is clear that the throughput of the network increases when the maintenance mechanism is included in the routing.

From the above comparisons it is clear that the maintenance enabled routing definitely performs better than the joint stable routing. Even though the joint stable routing selects the path which has less interference, there is chance for the PU to use it at any time. So through this waiting time algorithm, correct measures are taken to tackle the PU interference.

Fig. 2. Throughput (Existing versus Proposed)

VII. Conclusion and Scope of Future Work

MP-JSRCA protocol for CRN's, which jointly take mobility prediction-based stable routing and interference-avoiding channel assignment into account. First, it designs a mobility prediction based measure metric DTC that captures the node mobility, the impact to PNs, and the cochannel interference among CNs. Next, it presents channel assignment approaches for different channel interference patterns. Finally, it develops the MP-JSRCA protocol that jointly selects stable routes and assigns channels by discovering the link with the lowest DTC, which significantly improves the network throughput. But a minimum delay channel switching criteria has to be established, when PU interference occurs. Absence of route repair mechanism was the main problem in the present routing technique. By introducing the waiting time algorithm, route repair or maintenance is done. So that the performance of the routing is increased.

But the disadvantage is that route has to be rediscovered if the PU does not becomes free after the waiting time. So some other mechanisms has to be included in it to avoid the rediscovery.

References

[1] Matteo Cesana, Francesca Cuomo, Eylem Ekici, "Routing in cognitive radio networks: Challenges and solutions," *www.elsevier.com/locate/adhoc* , 2010.

[2] Moustafa Youssef, Mohamed Ibrahim, Mohamed Abdelatif, Lin Chen, and Athanasios V. Vasilakos, "Routing Metrics of Cognitive Radio Networks: A Survey," *IEEE Communications Surveys Tutorials* , vol. 16, no. 1, 2014.

[3] Guo-Mei Zhu1, Ian F. Akyildiz, Geng-Sheng (G.S.) Kuo, "STOD-RP: A Spectrum-Tree Based On-Demand Routing Protocol for Multi-Hop Cognitive Radio Networks," *IEEE Global Telecommunications Conference, GLOBECOME*, 2008.

[4] Xiaoxia Huang, Dianjie Lu, Pan Li and Yuguang Fang, "Coolest Path: Spectrum Mobility Aware Routing Metrics in Cognitive Ad Hoc Networks," In *International Conference on Distributed Computing Systems*, 2010.

[5] Anna Abbagnale, Francesca Cuomo,, "Gymkhana: a Connectivity-Based Routing Scheme for Cognitive Radio Ad Hoc Networks," In *IEEE Conference on Computer Communications, INFOCOM* , 2010.

[6] Hua Song, Xiaola Lin, "Spectrum Aware Highly Reliable Routing in Multihop Cognitive Radio Networks," *Wireless Communications Signal Processing International Conference on. IEEE* , 2009.

[7] Ilario Filippini, Eylem Ekici, and Matteo Cesana, "New Outlook on Routing in Cognitive Radio Networks: Minimum-Maintenance-Cost Routing," *IEEE/ACM Transactions On Networking* , 2013.

[8] Feilong Tang, Leonard Barolli, and Jie Li, "A Joint Design for Distributed Stable Routing and Channel Assignment Over Multihop and Multiflow Mobile Ad Hoc Cognitive Networks," *IEEE Transactions On Industrial Informatics,* , vol. 10, no. 2, May 2014.

Bigdata Management using Semantic Web and Machine Learning Techniques

Manu Madhavan,
Dept. of Computer Science and Engg,
SIMAT, Vavanoor, Palakkad
manumadhavan@simat.ac.in

***Abstract*—The aim of proposed research is to solve the Knowledge Management problems with big data analysis, using semantic web approach. The automatic analysis of information and creation of Knowledge are the key ideas in the field of semantic web. The two wide-ranging challenges to meet this semantic vision is processing of large-scale text documents and well structured knowledge representation. In the age of information explosion, performing these tasks of big data become tedious and impractical. Ontology, provides the conceptual representation of domain information, which can be a solution for knowledge representation in semantic web. Analysis and predictions from this represented knowledge can be done by exploring machine learning algorithms.**

.

***Keywords*—Ontology, Bigdata, Semantic Web, MapReduce, Graph Database.**

I. Introduction

THE renaissance in the modern digital age comes up with information explosion. Most of the information available is coded in Natural Language text. The complexities of processing and managing this bulk amount of information obliged the researchers to think about mechanizing these processes. Natural language processing (NLP) is the area based on the science called computational linguistics, which aims to design computational models for language processing. Extracting information, even from a single line of text by a machine is challenging job. The process becomes tedious when the NLP has to dealt with Big Data.

As the information available in the web are unstructured, it is not machine readable. The concept of semantic web tries to solve this challenge by providing semantic of the document in the form of extra annotations. Semantic role labeling (SRL) is a task of the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles [4]. This extra knowledge is represented as concept ontologies. The conceptualization of the domain in the form of ontologies, will enhance the semantic search and concept mining in the domain.

For a system handling Big Data, the essential key is to scale up to the exceptionally large volume of data and provide treatments for the characteristics like heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data (HACE Theorem) [1]. So, annotating the documents manually will not be practical. Here, the capacity of Machine learning techniques can be utilized for automatic semantic understanding, by which system will automatically map the knowledge to corresponding domain ontology.

The automatic labeling can be rule based or statistical. In rule based systems, the conditions for labeling will be represented in conditional statements (if-else logic) or predicate logic[3]. In statistical system, it will use statistical techniques to predict the semantic label. The proposed method use an hybrid approach which involve a two pass learning. In pass-1, it will statistically label the data and in pass-2 it verify the semantic labels by specialized rules. So, this method improve the accuracy of automatic knowledge representation. Another major challenge in knowledge management is the lack of standardization. This work also proposed to reuse the existing big data (ontologies) like DBPedia, Geo-Ontology, Gene Ontology, etc for solving standardization issues[4].

II. Major Areas of Work

Major Areas The major areas covered in this work are:

- Bigdata analysis
- Natural Language Processing
- Semantic Web and Ontology Engineering
- Machine Learning and Datamining

III. Relevance of the Work

Ontology engineering and NLP are active research areas in Computer Science. The proposed work can be viewed as a NLP application towards Semantic web technology. The work can be further modified into a generic system, that can be applied to any domains, by supplying the vocabulary in the form of domain ontology. The Big data analytics play a vital role in todays situation. Traditional Relational Database Management System (RDBMS) may not be a suitable data model for the large variety and vast amount of unstructured or semi-structured data. New analytical tools for processing large amount of unformatted data are coming up now-a-days [3]. But representation of data is more important than these. Many new data models say NoSQL data models have come up to accommodate these large amounts of data. Appropriate data models have to be selected so that new knowledge can be discovered from the existing data by integrating them suitably.

IV. Methodology

Collecting Information from unstructured texts are done using statistical natural language processing tools like NLTK,

Stanford CoreNLP etc. The extracted information are then stored in a graph DB to form a knowledge base. Additional information are added to the knowledge base with the help of semantic web technology and by using dbpedia and geo-ontology. Reasoners like pellet are used to improve the reasoning capabilities of stored data in knowledge base. The query system (query given in natural language text) will search for absolute match in the stored knowledge base. A Natural language generation module is made in which the processed query result is adjusted with the template and the result is produced in natural language texts.

V. Applications

Some of the relevant areas where the proposed ideas find application are:

- Analysing Medical Records, images
- Natural Language Report Management Systems
- Analysing Forensic records for tracking crime pattern
- Business intelligence and point of sale
- Question Answering System

References

[1] Xindong Wu, et.al, "Data mining with big data, Knowledge and Data Engineering", IEEE Transactions on (Volume:26 ,Issue: 1), 2014.

[2] Anantharangachar R, Ramani S, Rajagopalan S, "Ontology Guided Information Extraction from Unstructured Text", Int. Jr. of Web & Semantic Technology (IJWesT),Vol.4, No.1, January 2013.

[3] Chiticariu L, Krishnamurthy R, et.al, "Domain Adaptation of Rule-based Annotators for Named-Entity Recognition Tasks", in Proc. of Empirical Methods on Natural Language Processing (EMNLP), 2010.

[4] Abirami, A. M., et al. "Ontology based ranking of documents using Graph Databases: a Big Data Approach."

[5] Li Kang, Li Yi, LIU Dong, "Research on Construction Methods of Big Data Semantic Model", Proceedings of the WCE2014, Vol I, July , 2014, London, U.K

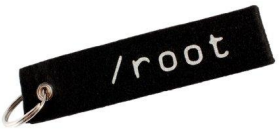
ROUTES TO THE ROOT

Jyothis K P,
Dept. of Computer Science and Engg,
SIMAT, Vavanoor, Palakkad
jyothis.kp@simat.ac.in

***Abstract*—This article shares the information about how to start working with Linux Kernal. The useful commands and concepts are described in detail. .**

ALMOST a decade back I started searching highs and lows for the routes to reach the Root. I tried many routes and tasted flavours from Slackware to Gentoo distros.As all we know in computing, the Root is a special user account used for system administration.On Unix-like systems the user with an user identification of zero is the superuser, regardless of the name of that account.

Anyway we should be 'zero' to access and transfigure Kernel.Most of the distros offers a Kernel which is sufficient.But in this article I am trying to reveal methods for compiling and installing a specialized kernel. You may have

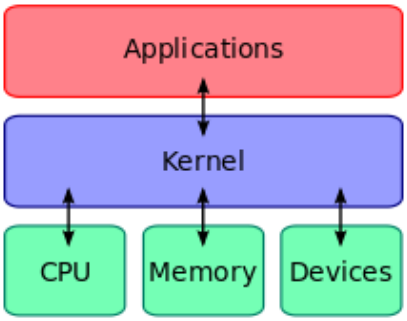


come across some system administrators saying that their web servers deliver high performance than others. The only reason behind this is that they have taken efforts in building their own kernel from source.There are some advantages of customizing the Kernel.

Whenever a program is customized for a specific architecture, usually the level of performance increases and becomes more stable version. Generic binaries need to function on various processors, which makes it inferior to your system.Another advantage is that you are able to eliminate the unnecessary modules and components from it. Removing excess modules and components will reduce the size of the compiled kernel and helps it to boot quickly.Building your own kernel allows you to add various features and optimize it for better performance.

Custom kernels can make your device that much better. If you happen to choose a kernel that is not right for you, you can always find another one and replace the currently installed one. Once you have found one that is right for you, then you just made your device that is much better for you.

The kernel is a computer program that manages input or output requests from software, and translates them into data processing instructions for the central processing unit and other electronic components of a computer. The kernel is a fundamental part of a modern computer's operating system.



The critical code of the kernel is usually loaded into a protected area of memory, which prevents it from being overwritten by other, less frequently used parts of the operating system or by applications. The kernel performs its tasks, such as executing processes and handling interrupts, in kernel space, whereas everything a user normally does, such as writing text in a text editor or running programs in a Graphical User Interface, is done in user space.

The first thing we need to have is the newly released Linux Kernel which is freely available in the official website of Linux Kernel Organization 'kernel.org'. The Linux Kernel Organization is a California Public Benefit Corporation established in 2002 to distribute the Linux kernel and other Open Source software to the public. Till date the stable release is 4.1.3.While doing this be sure to download the full source and it sould not be a patch of your existing Operating System.

The extracted file will be in the compressed format.The next thing to do is to extract the files from downloaded package. The `'tar xjvf kernel'` command can be used to extract it.

Next step is the configuration of Linux Kernel, which is one of the vital stages here.Usually this can be done by 4 different ways.

First one is the Make old configuration which is the most time consuming method of all.This will ask you one by one what the Kernel need to support.This will take not less that an hour to complete.So usually this method is not at all

recommended.But all the Geeks can make a try of this because you will get a boundless idea of what is going on while configuring.

The second is Make menuconfig.This is easier than the fist.It actually creates a menu where you can browse options on what the kernel supports.The menuconfig requires curses library, but that is likely already on your computer.

Curses is a terminal control library for Unix-like systems, enabling the construction of text user interface applications.The curses API is described in several places.Most implementations of curses use a database that can describe the capabilities of thousands of different terminals.

There are a few implementations, such as PD Curses, which use specialized device drivers rather than a terminal database.Curses has the advantage of back-portability to character-cell terminals and simplicity. For an application that does not require bit-mapped graphics or multiple fonts, an interface implementation using curses will usually be much simpler and faster than one using an X toolkit.

Using curses, programmers are able to write text-based applications without writing directly for any specific terminal type. The curses library on the executing system sends the correct control characters based on the terminal type. It provides an abstraction of one or more windows that maps onto the terminal screen.

Each window is represented by a character matrix. The programmer sets up each window to look as they want the display to look, and then tells the curses package to update the screen. The library determines a minimal set of changes needed to update the display and then executes these using the terminal's specific capabilities and control sequences.

In short, this means that the programmer simply creates a character matrix of how the screen should look and lets curses handle the work.

The third try is the clone of second but you will get a graphical based menu.Make qconfig/xconfig/gconfig"qconfig".

The command will work only if your system have QT library pre-installed.It is a cross-platform application framework that is widely used for developing application software that can be run on various software and hardware platforms with little or no change in the underlying codebase, while having the power and speed of native applications. Qt is currently being developed both by the Qt Company, a subsidiary of Digia, and the Qt Project under open-source governance, involving individual developers and firms working to advance Qt.

The fourth method is configuration of current kernel. Run this from your kernel source folder.Copy it by the command ``cp /boot/config-`uname -r` .config ''. This saves a lot of time, but you may want to change version number of the kernel to be compiled kernel to avoid replacing your current kernel. 'General setup' Local version - append to 'kernel release'.

When the config window is opened, you can see that a specific type of configuration is already selected like support for essential drivers, files systems like ext3 or ext4 etc. Further, you may customize the options like adding support for your specific type of device/controller/driver like you may add support for ntfs file system from Filesystem ¿¿select ntfs file system support, thereby taking full advantage of custom kernel.

During configuring the kernel, you can see a section known as kernel hacking, where different types of options are given for hacking into kernel and learning it.This is nothing but ethical.

An ethical hacker attempts to bypass system security and search for any weak points that could be exploited by malicious hackers. This information is then used by the organization to improve the system security, in an effort to minimize or eliminate any potential attacks.

If you want to use it then you may add further options, otherwise you may disable the option 'kernel debugging', as it makes the kernel a lot heavier and and may be improper to use in the production environment.

Now it is all set, well configured, it is time to compile and install the kernel. You can run needed commands in one line by separating them with double ampersand symbols as written below. This may take a long time.

```
'make && make modules_install && make
install'
```

You may want to use -j option with make. This allows to fork additional processes for compiling kernel.

```
'make -j 3'
```

In this the number 3 represents the total number of processes which has to be created.

Well, now our new Kernel is been installed.Our next step is to make the Kernel bootable.To do this run this command with Root privillages.

```
'mkinitrd -o initrd.img-<kernelversion>'
```

Now we can point the boot loader at the new kernel so it can be started. Use the tool that came with your distro to configure your bootloader. Add a new entry for the new kernel.Now you can reboot your system to our newly customized Kernel.

Basically you will surely gain three things from this. The fist one is faster boot time.Since the kernel will only probe the hardware on the system, the time it takes the system to boot can decrease.

Secondly lower memory usage.A custom kernel often uses less memory than the generic kernel by omitting unused features and device drivers. This is important because the kernel code remains resident in physical memory at all times, preventing that memory from being used by applications. For this reason, a custom kernel is useful on a system with a small amount of Random Access Memory.

The third and most important thing is that you will get additional hardware support. A custom kernel can add support for devices which are not present in the generic kernel.

Always remember before building a custom kernel, consider the reason for doing so and know the pros and cons.It is better to know how to customize the kernel, if you do things wrong, it could lead your baby to an intensive care unit.

“ So think twice, code once... ”

Happy coding...

REFERENCES

[1] Robert Love, *Linux System Programming*
[2] Daniel P. Bovet, Marco Cesati *Understanding the Linux Kernel*
[3] Robert Love, *Linux Kernel Development*
[4] Craig Hollabaugh, *Embedded Linux.*

Web Search Log Segmentation Techniques - A Survey

Hima G.
Computer Science and Engineering Dept.
MES college of Engineering
Kuttippuram, India
Email: himagnair@gmail.com

Jasila E.K.
Computer Science and Engineering Dept.
MES college of Engineering
Kuttippuram, India
Email: jasilaabhilash@gmail.com

***Abstract*—Web log is a pouch of valuable information that records users search queries and related actions on the internet. By mining the recorded information, it is possible to exploit the users underlying goals, interests and search behaviors. In order to mine information from web logs, the web logs should be segmented into sessions or tasks by clustering the queries. In this work, Task Trail is introduced to understand user search behaviors. A Task can be defined as set of semantically relevant queries issued to satisfy an atomic user information need. A task trail represents all user activities within the particular task, such as query reformulations, URL clicks. In most of the previous works, web search logs have been studied mainly at session or query level where users may submit several queries within one task and handle several tasks within one session. Although previous studies have addressed the problem of task identification, little is known about the advantage of using task over session or query for search applications. Instead of analyzing Session Trails or Query Trails, Task Trails can be analysed to determine the user search behaviour much more efficiently. By separating different task trails from a session, it can be used in several search applications such as determining user satisfaction, predicting user search interests, and suggesting related queries.**

***Keywords*-Query clustering, Search engine, Task- based clustering,User search interest, Web log.**

I. INTRODUCTION

Web logs[1] are a pouch of valuable information that records search queries and related actions of a user on internet. Web logs can be categorized into two types such as Search logs and Browse logs. Search logs are collected from search engines and record the interaction details between search engines and users. These details include queries submitted to search engines, search results returned to users, and clicks made by users. Browse logs are usually collected from client-side browser plug-ins or proxies of Internet Service providers. They record all URLs visited by users, irrespective of search engines and web servers. Web log query clustering is a technique for discovering similar queries on a search engine. The driving force of the development of query clustering techniques comes from the requirements of modern web searching.

The web log query clustering techniques can be mainly of Query-level, Session-level and Task-level. The Query-level clustering analyses each query in the web log separately. The session-level query clustering technique clusters a set of queries issued by the user of a web search engine within a particular time period. Task-level query clustering clusters a set of non-contiguous queries issued by a user to carry out a particular task. After clustering the queries into sessions or tasks, the web log can be analysed and required knowledge can be extracted. The need of web log analysis is to determine the user behaviour such as user satisfaction, to predict user search interest and to suggest related queries on internet.

Time	Event	Value	Task
09:03:26	Query	facebook	1
09:03:39	Click	www.facebook.com	1
09:06:34	Query	amazon	2
09:07:48	Query	faecbook	1
09:08:02	Click	facebook.com/login.php	1
09:10:23	Query	amazon kindle	2
09:10:31	Click	kindle.amazon.com	2
09:13:13	Query	gmail log in	3
09:13:19	Click	mail.google.com/mail	3
09:15:39	Query	amazon kindle books	2
09:15:47	Click	amazon.com/Kindle-eBooks...	2
09:15:59	Click	astore.amazon.com/Amazon..	2
09:17:51	Query	i'm picking up stones	4
09:18:54	Query	i'm picking up stones lyrics	4
09:19:28	Query	pickin' up stones lyrics	4

Figure 1. A sample session from web search log

II. LITERATURE SURVEY

Web log segmentation can be done at different levels such as query-level, task-level or session level. The query-level segmentation method segments the web log into independent queries. And each queries are considered as independent segments. In web logs, a single query is often followed by a sequence of browse behaviors before the next query is submitted by the same user. Thus, the simplest search log segmentation is to treat one query plus its followers as an independent query trail. Session-level segmentation method segments the web log into various sessions. A session can be defined as a set of several TCP connections generated while surfing the web during a given time frame by a single user. Task-level segmentation method segments the web log into different tasks. A task can be defined as a Set of semantically relevant queries issued to satisfy an atomic user information need.

A. Threshold-based Algorithm

Neha Bagoria, and Nirmala Huidrom proposed a threshold based algorithm[2] for Web log clustering. It is a session-level clustering of web log. Threshold based algorithm is the simplest and very basic method that was used for the identification of user-session. The algorithm depends on the proper selection of the threshold value. Based on the threshold value, the sessions are identified. The algorithm works as follows:
The procedure starts with the proper selection of threshold value. Based on this value, the following condition is checked:
• If the inter-arrival time between the consecutive queries from a web log is less than the threshold value, then the two queries are considered to be in the same session.
• If the inter-arrival time between the consecutive queries is greater than the threshold value, then the two queries are considered to be in different sessions i.e. the first query is the last element of the current session and the second query is the first element of the next session.

The main advantages of the Threshold-based algorithm for session identification is, it is simple and easy to implement. The main limitation of this method is that it requires a priori definition of the threshold value which is very difficult to set correctly. If the threshold value is too small, then two related queries may be clustered separately. On the other hand, if the threshold value is too large then the condition for merging two unrelated queries may arrive.

B. Session Segmentation Method using COBWEB

The session level clustering of web log is done using COBWEB algorithm[3]. COBWEB is a hierarchical conceptual clustering algorithm, its input objects are described with categorical attribute -value pairs. It adopts heuristic estimate metrics, and category utility to instruct the tree construction. Category Utility, CU is defined as follows:

$$CU = \frac{\Sigma_n^{k=1} P(C_k)[\Sigma_i \Sigma_j P(A_i{=}V_{ij}|C_k)^2 - \Sigma_i \Sigma_j P(A_i{=}V_{ij})^2]}{n}$$

Here, n means the quantity of current categories and $P(A_i = V_{ij})$ serves as a weight. $P(A_i = V_{ij} \mid C_k)$ is called predictability, it is the probability that an object has the value V_{ij} for feature A_i given that the object belongs to category C_k. The higher the probability, the more likely two objects in a category share the same feature values.

There are mainly two attributes are considered to find the similarity between web log queries namely, Query Time Interval (TI) and Query Likelihood (QL).
Query Time Interval(TI): TI is the time span between current query and the previous query, denoted by :
$TI = T_{QC} - T_{QP}$
Query Likelihood(QL): QL is an attribute to quantize the semantic likelihood between current query and previous query. Queries are divided into several terms and calculate QL according to the formula below:

$$QL = \frac{\sum_{t \in Q_c \cap t \in Q_p} \sharp t_{Q_c} \times \sharp t_{Q_p}}{\sqrt{|Q_c| \times |Q_p|}}$$

Where ♯ t is a variable to define how many times t term appears, Qc represents current query and Qp represents previous query. Here, a binary tree is built to distinguish the session borders with the incremental COBWEB algorithm. Each query is calculated twice by the CU formula on the assumption that it is discriminated into YES class or NO class (ie, either session boundary or Non-session boundary). And the higher one is the last discrimination of the query.
COBWEB Session Discrimination Algorithm
Input: Query Log
Output: Session Sequence
Steps:
1.Read the query records set $(A_1, A_2, ...A_n)$
2.Calculate the attribute value TI and QL for each query
3.Create a tree and initialize each class
4.Input the instance set $(A_1, A_2, ...A_n)$ and the corresponding attribute values
5.Update the root node information
6.Compute Category utility (CU) for each A_n and incorporate $A_i$into that node of the largest CU
7.Continue steps until algorithm ends by outputting two data sets of ”YES” class and ”NO” class

C. Segmentation using Bipartite Graph

D. Beeferman and A. Berger et.al, proposed an agglomerative clustering algorithm[4] for the segmentation of web search log. It is a task-level segmentation method. The first step of this method is to construct a query-page bipartite graph with one set of the nodes corresponding to the set of queries submitted by the user, and the other set of nodes corresponding to the sets of clicked pages. When a user clicks on a page, a link is created between the query and the page on the bipartite graph. After the bipartite graph is obtained, an agglomerative clustering algorithm is used to discover similar queries and similar pages. During the clustering process, the algorithm iteratively combines the two most similar queries into one query node, then the two most similar pages into one page node. This process of combination of queries and pages is repeated until a termination condition is satisfied. The main reason for not clustering all the queries first and then all the pages next are that two queries may seem unrelated prior to page clustering because they link to two different pages but they may become similar to each other if the two pages have a high enough similarity to each other and are merged later. After the bipartite graph is constructed, the agglomerative clustering algorithm is applied to obtain clusters of similar

queries and similar pages.

Agglomerative Clustering Algorithm

Input : A Query-Page Bipartite Graph G

Output : A Clustered Query-Concept Bipartite Graph G^c

Steps:

1. Obtain the similarity scores for all possible pairs of queries in G using the noise-tolerant similarity function given below:

$Sim(x,y) = \{\frac{|N(x)\cap N(y)|}{|N(x)\cup N(y)|} \quad if \quad |\ N(x)\cup N(y)\ |> 0$
$0 \qquad ;Otherwise\}$

2. Merge the pair of queries (q_i, q_j) that has the highest similarity score.

3. Obtain the similarity scores for all possible pairs of concepts in G using the noise-tolerant similarity f unction given in (1).

4: Merge the pair of concepts (c_i, c_j) that has the highest similarity score.

5. Unless termination is reached, repeat steps 1-4.

The advantage of this method is that it is able to detect the interleaving of queries in web search logs. The drawback of this method is that only content based similarity is considered in web log clustering.

D. Weighted Connected Component Method

Claudio Lucchese et.al proposed Query Clustering using Weighted Connected Component (QC-WCC) method[5]. QC-WCC is a graph based algorithm. Based on the given-time gap session ϕ, it builds a complete graph $G_\phi = (V, E, w)$ where set of nodes V are the queries in ϕ. The set of edges E are weighted by the similarity of the corresponding nodes. The weighting function w is a similarity function$w : E \to R \in [0, 1]$ that can be easily instantiated in terms of the distance functions μ_1 or μ_2, described in the next section (i.e.,$w = 1 - \mu_1$ or $w = 1 - \mu_2$). Upon the query similarity function, we build an undirected graph for queries within a session. The vertices of the graph are queries and the edges are similarity scores between queries. After removing the suspicious edges with scores below a threshold, any connected component of the remain graph is identified as a task.

There are mainly two types of similarity measures considered such as Content-based similarity($\mu_{content}$) and Semantic-based similarity($\mu_{semantic}$). Two queries are considered as similar in content if they share some common terms. To capture content distance between queries, a Jaccard index can be used. Semantic-based similarity is the similarity in the meanings of two queries. Inorder to capture the semantic-based similarity, it needs some source of knowledge. Usually, this knowledge comes from large text collections (i.e., corpora) or from semantic resources. Both the Content-based similarity ($\mu_{content}$), and Semantic-based similarity ($\mu_{semantic}$), can be put together by using a distance functionμ_1.

$\mu_1 = \alpha.\mu_{content} + (1-\alpha).\mu_{semantic}$

In addition a conditional distance function is also used based on the following heuristic: If the content-based distance between two queries does not exceed a certain threshold then queries are also task related; otherwise semantic expansion of the queries are considered and the final distance score is computed as the minimum between content-based and semantic-based distance values.

$\mu_2 = \{\mu_{content} \quad if \mu_{content} < t$
$min(\mu_{content}, b.\mu_{semantic}) \qquad ;Otherwise\}$

QC-WCC Algorithm

Input : An undirected graph $G_\phi = (V, E, w)$

Output : An undirected graph $G'_\phi = (V, E, w)$

Steps:

1) All the edges $e \in E$ whose weight is smaller than a given threshold are removed (ie,$W(e) < \eta$). Thus obtaining a pruned graphG'_ϕ

2) The connected components of G'_ϕ are extracted

3) Such connected components identify the clusters of related queries

The main advantages of this method is that it can deal with interleaved and overlapped tasks and also it can capture both lexical and semantic relationship between queries in web logs. The main drawbacks are high Computational complexity and high Time complexity.

E. Query Clustering Using Head-Tail Component (QC-HTC)

Claudio Lucchese et.al proposed QC-HTC algorithm[5] for task-level web log segmentation. QC-HTC is a variation of the QC-WCC algorithm, which does not need to compute the full similarity graph. Since queries are submitted one after the other by the user, the QC-HTC algorithms takes advantage of this sequentiality to reduce the number of similarity computations needed by QC-WCC. The algorithm works in two phases as follows :

The first step aims at creating an approximate fine-grained clustering of the given time-gap session $\phi =< q_1, q_2...q_m >$. Every single web-mediated task generates a sequence of queries and each web-mediated task is observed as a set of fragments, i.e. smaller sets of consecutive queries, and fragments of different tasks are interleaved in the query log because of multi-tasking. The algorithm exploits the sequentiality of user queries, and tries to detect the above fragments, by partitioning the given time-gap session into sequential clusters, where a sequential cluster denoted with $C^{\sim i}$, must contain only queries that occur in a row within the query log, and such that each query is sufficiently similar to the chronologically following one. Now, the similarity between one query and the next in original data

is to be computed.

The second step of the algorithm merges together the set of fragments when they are related, trying to overcome the interleaving of different tasks. Here, the assumption that reduces the computational cost of the algorithm is that a cluster of queries can be described well by just the chronologically first and last of its queries, respectively denoted with $head(C^{\sim i})$ and $tail(C^{\sim i})$. Therefore, the similarity, between two clusters $C^{\sim i}, C^{\sim j}$is computed as:

$$S(C^{\sim i}, C^{\sim j}) = min_{p,q \in \{head(C^{\sim i}), tail(C^{\sim i})\}} w(e(q,p))$$

where w weights the edge $e(q,p)$ linking the queries p and q on the basis of their similarity. The final clustering is produced as follows: The first cluster C^1 is initialized with the oldest sequential cluster $C^{\sim 1}$, which is removed from the set of sequential clusters. Then, C^1 is compared with any other sequential cluster $C^{\sim 1}$(ordered chronologically) by computing the similarity as above. Given a threshold η, if S($C^1, C^{\sim i} < \eta$, then $C^{\sim i}$ is merged into C^1, the head and tail queries of C^1, are updated consequently, and $C^{\sim i}$ is removed from the set of sequential clusters. The algorithm continues comparing the new cluster C^1 with the remaining sequential clusters. When all the sequential clusters have been considered, the oldest sequential cluster available is used to build a new cluster C^2. The algorithm iterates this procedure until no more sequential clusters are left.

The advantages of this method is that Time complexity is less and both content-based and semantic relationships between queries are considered. The drawback is that it violates the task interleaving observation found in search logs.

F. Query Task Clustering Algorithm (QTC)

Zhen Liao and Yang Song et.al proposed Query Task Clustering algorithm[6] for task-level clustering of web search log. QTC Algorithm is based on the observation that consecutive query pairs are more likely belonging to same task than non-consecutive ones. QTC prefers to first compute the similarities for consecutive query pairs by time stamps.

For example, given a sequence of 4 queries q_1, q_2, q_3, q_4, QC-WCC needs 6 times of pair-wise relevance computations. For QC-SP, if q_1 is similar to q_2 and q_2 is similar to q_3, there is no need to compute the relevance between q_1 and q_3 any more. If q_1 is similar to q_2 but q_2 is not similar to q_3, QC-SP still has to compute the relevance between q_1 and q_3 to avoid the task interleaving. For sessions having multiple tasks, if some tasks have more than two consecutive queries, the time cost can still be reduced for the same reason. In the worst case that all tasks are short and interleaved with each other, QC-SP has the same time complexity as QC-WCC.

QTC Algorithm

Input: Queries Q, cut-off threshold b
Output: A set of tasks Θ
Steps:
1) Set of tasks Θ and Query to task table L, are initialized as null
2) For each consecutive queries, check whether they are from same task
3) If two queries are not in the same task, compute the similarity score between them using the formula,
$Sim(Q_x, Q_y) = \Sigma_t min(P(t \mid Q_x), P(t \mid Q_y))$
4) If the similarity score $\geq$ threshold b, merge those queries as a single task
5) Modify L
6) Continue the steps until all queries are clustered

The main advantage of this method is that its time complexity is less. The drawback of this method is that its space complexity is more.

G. Bounded Spread Query Task Clustering (QC-BSP)

Zhen Liao and Yang Song et.al proposed Query Task Clustering algorithm[7] for task-level clustering of web search log . QC-BSP Algorithm is a variation of QTC Algorithm. Here, the idea is that two queries far away from each other are not likely from the same task. By setting a length bound bl, the time complexity of QC-BSP is reduced to O(k bl N). Considering the case where users repeat queries after a while that possibly exceeds the length bound, same queries within a session are identified first; then, QC-BSP will only examine their own neighbours within the length bound separately. In the end, tasks of same queries are merged into a single one.

QC-BSP Algorithm

Input : Query set Q, Cut-off threshold b, Bounded length bl;
Output : A set of tasks Θ
Steps :
1) Initialize same queries into one task
2) If two queries are not in the same task,
3) Set bound length bl
4) Compute similarity between queries using query similarity function
5) If it is greater than certain threshold group queries into same task
6) Repeat until all queries are clustered
7) Return Θ

The main advantage of this method is that it can capture the multi-tasking behaviour of user. The main drawback of this method is that, the semantic attributes are not considered here.

III. PERFORMANCE ANALYSIS

A comparative study of various web log segmentation methods are presented here. The segmentation type, similarity attributes and capturing of multi-tasking behavior are taken as the parameters for comparison. The performance comparison is summarized in Table 1.

Algorithms	Segmentation Type	Similarity Attributes	Capture Multi-tasking Behavior
Threshold-based Algorithm	Session-level	Temporal only	No
COBWEB	Session-level	Temporal and content based	No
Bipartite graph-based Algorithm	Task-level	Content-based only	Yes
QC-WCC	Task-level	Content-based and semantic	Yes
QC-HTC	Task-level	Content-based and semantic	No
QTC	Task-level	Content-based only	Yes
QC-BSP	Task-level	Content-based only	Yes

Table I
COMPARISON OF VARIOUS WEB LOG SEGMENTATION METHODS

IV. CONCLUSION

The web log can be analysed to determine the web user search behaviour. The web log can be analysed at different levels by following Query trails, Task-trails and session-trail. If query trail is followed in web logs to determine user search behaviour, the semantic associations between adjacent query trails will be lost. So it is not an efficient method. Since the session strictly follows the chronological order of user behaviors in search logs, the entire search logs of one user may be segmented into a sequence of disjoint sessions along the time dimension. So, among query trail and session trail, task trail is more precise to determine user search behaviour.

REFERENCES

[1] Ryen W. White, Jeff Huang, "Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs", *ACM 978-1-60558-896-4/10/07,* July 2010.

[2] Nirmala Huidrom, Neha Bagoria., "Clustering Techniques for the Identification of Web User Session", *International Journal of Scientific and Research Publications,* Volume 3, Issue 1, January 2013.

[3] Hou, Zhenshan, Mingliang Cui, Ping Li, L. Wei, Wenhao Ying, Wanli Zuo., "Session Segmentation Method Based on COBWEB", *IEEE 2nd International Conference on Cloud Computing and Intelligence Systems* ,vol., no., pp.327,333, 2012

[4] D. Beeferman, A. Berger., "Agglomerative Clustering of a Search Engine Query Log", *Proc. ACM SIGKDD,* ,2000.

[5] Lucchese, Claudio, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, Gabriele Tolomei., "Identifying Task-based Sessions in Search Engine Query Logs" *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* ,vol. 7, no. 8, pp. 85-97, February 2011.

[6] Z. Liao, Yang Song, Li-Wei He, and Yalou Huang., "Evaluating the Effectiveness of Search Task Trails", *Proceedings of the 21st International Conference on World Wide Web* , vol. 23, no. 6, 2012.

[7] Z. Liao, Y. Song, Y. Huang, L. He and Q. He, "Task Trail: An Effective Segmentation of User Search Behavior", *IEEE Transactions on Knowledge and Data Engineering,* , vol. 26, no. 12, pp. 3090-3102, 2014.

Facial-Expression Database From Movies

Vishnu Venugopal
Eighth Semester, 2011 Admission
Department of Computer Science and Engineering,
Sreepathy Institute of Management & Technology, Vavannoor, Palakkad, India-Pin 679533
E-mail: vishnuvenugopal15@gmail.com

***Abstract*— Collecting richly annotated, large datasets representing real-world conditions is a challenging task. With the progress in computer vision research, researchers have developed robust human facial-expression analysis solutions, but largely only for tightly controlled environments. Facial expressions are the visible facial changes in response to a persons internal affective state, intention, or social communication. Automatic facial-expression analysis is an active field of research for many years with applications in affective computing, intelligent environments, lie detection, psychiatry, emotion and paralinguistic communication, and multimodal human computer interface (HCI). In automatic facial expression analysis domain, realistic data plays an important role. But obtaining such a realistic database is difficult. Several popular facial-expression databases exist, but the majority of them have been recorded in tightly controlled laboratory environments, where the subjects were asked to generate certain expressions. These lab scenarios are in no way a true representation of the real world. Ideally, what we want is a dataset of spontaneous facial expressions in challenging real-world environments. To solve this problem, collect two new facial-expression databases derived from movies via a semiautomatic recommender- based method. Extracting a database for facial expressions from scenes in movies, is helpful as environments in them are more closely resembling to the real world than that of those previous datasets.**

I. Introduction

A facial expression is a visible manifestation of the affective state, cognitive activity, intention, personality, and psychopathology of a person. It plays a communicative role in interpersonal relations. Facial expressions, and other gestures, convey non-verbal communication cues in face-to-face interactions. These cues may also complement speech by helping the listener to elicit the intended meaning of spoken words. It is reported that facial expressions have a considerable effect on a listening interlocutor; the facial expression of a speaker accounts for about 55 percent of the effect, 38 percent of the latter is conveyed by voice intonation and 7 percent by the spoken words.

As a consequence of the information that they carry, facial expressions can play an important role wherever humans interact with machines. Automatic recognition of facial expressions may act as a component of natural human-machine interfaces. Such interfaces would enable the automated provision of services that require a good appreciation of the emotional state of the service user, as would be the case in transactions that involve negotiation, for example. Some robots can also benefit from the ability to recognise expressions. It is an active field of research for many years with various other applications in affective computing, intelligent environments, lie detection, psychiatry, emotion and paralinguistic communication.

Despite the task duality that exists between facial expression recognition and face recognition, it can be observed that similar architectures and processing techniques are often used for both recognition tasks. The duality arises from the following considerations. In addition to conveying expressions, faces also carry other information such as the identity of a person. By definition, the expression of a face is the focal element in facial expression recognition. Hence, personal identity information conveyed by a face is an unwanted source of variability in expression recognition. Conversely, variability arising from facial expression is unwanted in face recognition, where the uniqueness of a face is the central recognition criterion.

Facial expression analysis methods can also be classified into image based and video based. Human facial expressions are dynamic in nature and, therefore, video based methods are more robust since they encode the facial dynamics, which are not available in static, image-based methods. Studies have also proven the effectiveness of video based methods over the static ones. However, there are scenarios where temporal data is not available and image based facial expression analysis methods come into picture [9].

A. Architecture of Automatic Facial Expression Recognition System

Automatic systems for facial expression recognition usually take the form of a sequential configuration of processing blocks. The main blocks are [4]:

- image acquisition
- pre-processing
- feature extraction
- classification
- post-processing.

1) Image Acquisition: Images used for facial expression recognition are static images or image sequences. An image sequence contains potentially more information than a still image, because the former also depicts the temporal characteristics of an expression. Monochrome (grey-scale) facial image sequences was the most popular type of pictures used for automatic expression recognition. However, colour images became prevalent, owing to the increasing availability of low-cost colour image acquisition equipment, and the ability of colour images to convey emotional cues such as blushing.

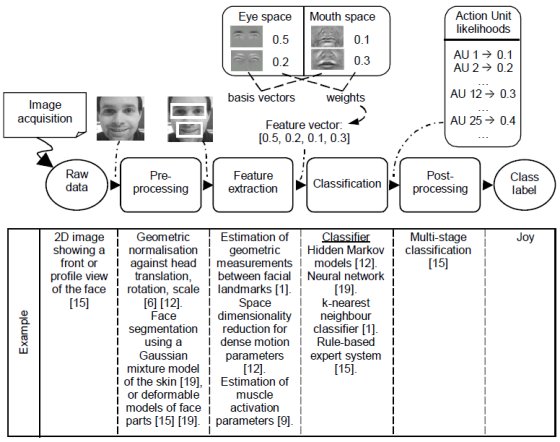


Fig. 1. Automatic Facial Expression Recognition

2) Pre-processing: Image pre-processing often takes the form of signal conditioning (such as noise removal, and normalisation against the variation of pixel position or brightness), together with segmentation, location, or tracking of the face or its parts. Expression representation can be sensitive to translation, scaling, and rotation of the head in an image. To combat the effect of these unwanted transformations, the facial image may be geometrically standardised prior to classification. This normalisation is usually based on references provided by the eyes or nostrils.

Segmentation is concerned with the demarcation of image portions conveying relevant facial information. Face segmentation is often anchored on the shape, motion, colour, texture, and spatial configuration of the face or its components. The face location process yields the position and spatial extent of faces in an image; it is typically based on segmentation results. A variety of face detection techniques have been developed. However, robust detection of faces or their constituents is difficult to attain in many real-world settings. Tracking is often implemented as location, of the face or its parts, within an image sequence, whereby previously determined location is typically used for estimating location in subsequent image frames.

3) Feature Extraction: Feature extraction converts pixel data into a higher-level representation of shape, motion, colour, texture, and spatial configuration of the face or its components. The extracted representation is used for subsequent expression categorisation. Feature extraction generally reduces the dimensionality of the input space. The reduction procedure should (ideally) retain essential information possessing high discrimination power and high stability.

4) Classification: Expression categorisation is performed by a classifier, which often consists of models of pattern distribution, coupled to a decision procedure. A wide range of classifiers, covering parametric as well as non-parametric techniques, has been applied to the automatic expression recognition problem. The two main types of classes used in facial expression recognition are action units (AUs), and the prototypic facial expressions defined by Ekman.

The 6 prototypic expressions relate to the emotional states of happiness, sadness, surprise, anger, fear, and disgust. However, it has been noted that the variation in complexity and meaning of expressions covers far more than these six expression categories. An AU is one of 46 atomic elements of visible facial movement or its associated deformation; an expression typically results from the agglomeration of several AUs. AUs are described in the Facial Action Coding System (FACS). Sometimes, AU and prototypic expression classes are both used in a hierarchical recognition system - for example, categorisation into AUs can be used as a low-level of expression classification, followed by a high-level classification of AU combinations into basic expression prototypes

5) Post-processing: Post-processing aims to improve recognition accuracy, by exploiting domain knowledge to correct classification errors, or by coupling together several levels of a classification hierarchy.

II. Motivation

Although humans recognise facial expressions virtually without effort or delay, reliable expression recognition by machine is still a challenge. The problems that have haunted the pattern recognition community at large still require attention. A key challenge is achieving optimal preprocessing, feature extraction or selection, and classification, particularly under conditions of input data variability [3].

(i) View or pose of the head. Although constraints are often imposed on the position and orientation of the head relative to the camera, and the setting of camera zoom, it should be noted that some processing techniques have been developed, which have good insensitivity to translation, scaling, and inplane rotation of the head.

(ii) Environment clutter and illumination. Complex image background pattern, occlusion, and uncontrolled lighting have a potentially negative effect on recognition. These factors would typically make image segmentation more difficult to perform reliably. Hence, they may potentially cause the contamination of feature extraction by information not related to facial expression. Consequently, many researchers use uncluttered backgrounds and controlled illumination, although such conditions do not match the operational environment of some potential applications of expression recognition.

(iii) Miscellaneous sources of facial variability. Facial characteristics display a high degree of variability due to a number of factors, such as: differences across people (arising from age, illness, gender, or race, for example), growth or shaving of beards or facial hair, make-up, blending of several expressions, and superposition of speech-related (articulatory) facial deformation.

A. Related Work

One of the earliest databases published is the widely used Cohn-Kanade database, which contains 97 subjects who posed

in a lab situation for the six universal and neutral expressions. Its extension CK+ contains 123 subjects, but the new videos were shot in a similar environment. The Multi-PIE database is another popular database that contains both temporal and static samples recorded in the lab over five sessions. It contains 337 subjects covering different pose and illumination scenes. Each of these databases were constructed manually, with the subjects posing in sequential scenes. The MMI database is a searchable temporal database with 75 subjects. All of these are posed, lab-controlled environment databases. The subjects display various acted (not spontaneous) expressions. The recording environment is nowhere near real-world conditions [7].

The RU-FACS (Rutgers and University of California, San Diego, Facial Action Coding System [FACS]) database is a FACS-coded temporal database containing spontaneous facial expressions, but it is proprietary and unavailable to other researchers. The Belfast database consists of a combination of studio recordings and TV program grabs labeled with particular expressions. The number of TV clips in this database is sparse.

III. Theoretical Background

A. Emotion classification

Emotion classification, the means by which one emotion is distinguished from another, is a hotly contested issue in emotion research and affective science. The classification of emotions has mainly been researched from two fundamental viewpoints. The first viewpoint is that emotions are discrete and fundamentally different constructs while the second viewpoint asserts that emotions can be characterized on a dimensional basis in groupings.

1) Emotions as discrete categories: As per the Discrete Emotion Theory, all humans are thought to have an innate set of basic emotions that are cross-culturally recognizable. These basic emotions are described as discrete because they are believed to be distinguishable by an individuals facial expression and biological processes.Various theorists have conducted studies in attempts to determine which are the basic emotions. A popular example of one is Paul Ekman and his colleagues' cross-cultural study of 1972, in which they concluded that the six basic emotions are anger, disgust, fear, happiness, sadness, and surprise. Ekman explains that there are particular characteristics attached to each of these emotions, allowing them to be expressed in varying degrees. Each emotion therefore acts as a discrete category rather than an individual emotional state.

Semantically Distinct Emotions

Eugene Bann proposed a theory that people transmit, their understanding of emotions through the language they use that surrounds mentioned emotion keywords. He posits that the more distinct language is used to express a certain emotion, then the more distinct the perception (including proprioception) of that emotion is, and thus more basic. This allows us to select the dimensions best representing the entire spectrum of emotion. Coincidentally, it was found that Ekmans (1972) basic emotion set, arguably the most frequently used for classifying emotions, is the most semantically distinct.

2) Dimensional models of emotion: For both theoretical and practical reasons some researchers define emotions according to one or more dimensions. Wilhelm Max Wundt, the father of modern psychology, proposed in 1897 that emotions can be described by three dimensions: "pleasurable versus unpleasurable", "arousing or subduing" and "strain or relaxation". In 1954 Harold Schlosberg named three dimensions of emotion: "pleasantness-unpleasantness", "attention-rejection" and "level of activation".

Dimensional models of emotion attempt to conceptualize human emotions by defining where they lie in two or three dimensions. Almost all dimensional models incorporate valence and arousal or intensity dimensions.

Several dimensional models of emotion have been developed, though there are just a few that remain as the dominant models currently accepted by most. The two-dimensional models that are most prominent are the circumplex model, the vector model, and the Positive Activation Negative Activation (PANA) model.

Circumplex model

The circumplex model of emotion was first developed by James Russell. This model suggests that emotions are distributed in a two-dimensional circular space, containing arousal and valence dimensions. Arousal represents the vertical axis and valence represents the horizontal axis, while the center of the circle represents a neutral valence and a medium level of arousal. In this model, emotional states can be represented at any level of valence and arousal, or at a neutral level of one or both of these factors. Circumplex models have been used most commonly to test stimuli of emotion words, emotional facial expressions, and affective states.

Vector model The vector model of emotion first appeared in 1992. This two-dimensional model consists of vectors that point in two directions, representing a "boomerang" shape. The model assumes that there is always an underlying arousal dimension, and that valence determines the direction in which a particular emotion lies. For example, a positive valence would shift the emotion up the top vector and a negative valence would shift the emotion down the bottom vector. In this model, high arousal states are differentiated by their valence, whereas low arousal states are more neutral and are represented near the meeting point of the vectors. Vector models have been most widely used in the testing of word and picture stimuli.

Positive activation negative activation (PANA) model The positive activation negative activation (PANA) or "consensual" model of emotion, originally created by Watson and Tellegan in 1985, suggests that positive affect and negative affect are two separate systems. Similar to the vector model, states of higher arousal tend to be defined by their valence, and states of lower arousal tend to be more neutral in terms of valence. In the PANA model, the vertical axis represents low to high positive affect and the horizontal axis represents low to high negative affect. The dimensions of valence and

arousal lay at a 45-degree rotation over these axes.

PAD emotional state model

The PAD emotional state model is a psychological model developed by Albert Mehrabian and James A. Russell to describe and measure emotional states. PAD uses three numerical dimensions to represent all emotions. The PAD dimensions are Pleasure, Arousal and Dominance.

The Pleasure-Displeasure Scale measures how pleasant an emotion may be. For instance both anger and fear are unpleasant emotions, and score high on the displeasure scale. However joy is a pleasant emotion.

The Arousal-Nonarousal Scale measures the intensity of the emotion. For instance while both anger and rage are unpleasant emotions, rage has a higher intensity or a higher arousal state. However boredom, which is also an unpleasant state, has a low arousal value.

The Dominance-Submissiveness Scale represents the controlling and dominant nature of the emotion. For instance while both fear and anger are unpleasant emotions, anger is a dominant emotion, while fear is a submissive emotion.

Lovheim cube of emotion

In 2011, Lovheim proposed a direct relation between specific combinations of the levels of the signal substances dopamine, noradrenaline and serotonin and eight basic emotions, as shown in figure 2. A three-dimensional model, the Lvheim cube of emotion, was presented where the signal substances forms the axes of a coordinate system, and the eight basic emotions according to Silvan Tomkins are placed in the eight corners. Anger is, according to the model, for example produced by the combination of low serotonin, high dopamine and high noradrenaline. Lvheim wrote that as neither the serotonin nor the dopamine axis is identical to the ”pleasantness” (i.e. valence) dimension in earlier theories, the cube seems somewhat rotated when compared to these models.

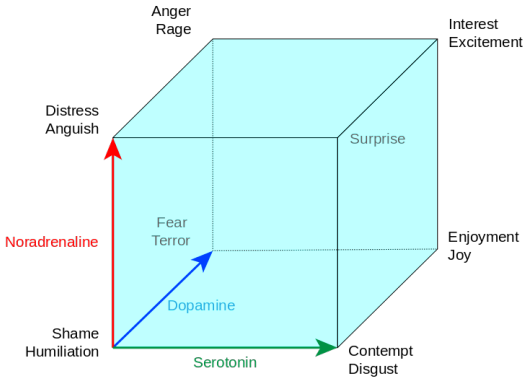


Fig. 2. Lovheim cube of emotion

Plutchik’s model

Robert Plutchik offers a three-dimensional model. Refer Figure 3. It arranges emotions in concentric circles where inner circles are more basic and outer circles more complex. Notably, outer circles are also formed by blending the inner circle emotions. In computer science, Plutchik’s model is often used, in different forms or versions, for tasks such as affective human-computer interaction or sentiment analysis.

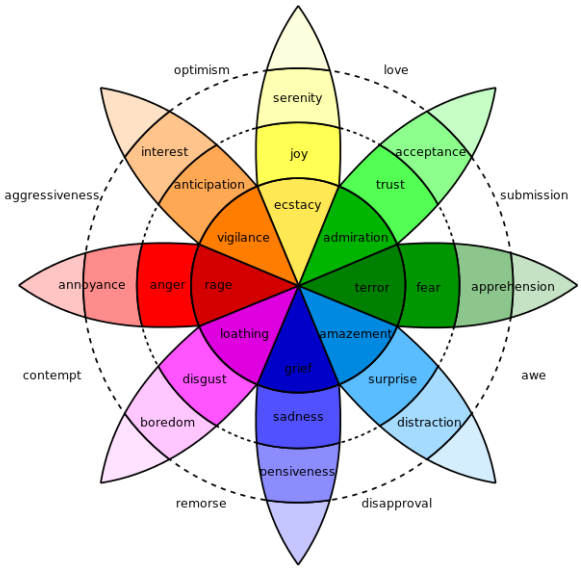


Fig. 3. Plutchik’s model

B. Facial Action Coding System

Facial Action Coding System (FACS) is a system to taxonomize human facial movements by their appearance on the face, based on a system originally developed by a Swedish anatomist named Carl-Herman Hjortsj. It was later adopted by Paul Ekman and Wallace V. Friesen, and published in 1978. Movements of individual facial muscles are encoded by FACS from slight different instant changes in facial appearance. It is a common standard to systematically categorize the physical expression of emotions, and it has proven useful to psychologists and to animators.

Using FACS, human coders can manually code nearly any anatomically possible facial expression, deconstructing it into the specific Action Units (AU) and their temporal segments that produced the expression. As AUs are independent of any interpretation, they can be used for any higher order decision making process including recognition of basic emotions, or pre-programmed commands for an ambient intelligent environment.

FACS defines AUs, which are a contraction or relaxation of one or more muscles. For example, FACS can be used to distinguish two types of smiles as follows:

- Insincere and voluntary Pan-Am smile: contraction of zygomatic major alone

- Sincere and involuntary Duchenne smile: contraction of zygomatic major and inferior part of orbicularis oculi.

Although the labeling of expressions currently requires trained experts, researchers have had some success in using computers to automatically identify FACS codes, and thus quickly identify emotions. Computer graphical face models, such as CANDIDE or Artnatomy, allow expressions to be artificially posed by setting the desired action units.

The use of FACS has been proposed for use in the analysis of depression, and the measurement of pain in patients unable to express themselves verbally. FACS is designed to be self-instructional. People can learn the technique from a number of sources including manuals and workshops, and obtain certification through testing. The original FACS has been modified to analyze facial movements in several non-human primates [13].

1) Codes for Action Units: **Action Units (AUs)** are the fundamental actions of individual muscles or groups of muscles. There are 46 AU main codes.

Action Descriptors (ADs) are unitary movements that may involve the actions of several muscle groups (e.g., a forward thrusting movement of the jaw). The muscular basis for these actions hasn't been specified and specific behaviors haven't been distinguished as precisely as for the AUs.

Intensities of FACS are annotated by appending letters AE (for minimal-maximal intensity) to the Action Unit number (e.g. AU 1A is the weakest trace of AU 1 and AU 1E is the maximum intensity possible for the individual person).

- A Trace
- B Slight
- C Marked or Pronounced
- D Severe or Extreme
- E Maximum

Some example of action units (AUs) and action descriptors (ADs) is given in table II

EMFACS (Emotional Facial Action Coding System) and FACSAID (Facial Action Coding System Affect Interpretation Dictionary) consider only emotion-related facial actions. Examples of these are given in table I:

Emotion	Action Units
Happiness	6+12
Sadness	1+4+15
Fear	1+2+5B+26
Anger	4+5+7+23
Disgust	9+15+16

TABLE I
EMOTIONS WITH ACTION UNITS

C. Face Feature Extraction

The extraction of facial feature points, (eyes, nose, mouth) plays an important role in many applications, such as face recognition, face detection, model based image coding , expression recognition, facial animation and head pose determination. It is important to note that because the systems use spatial geometry of distinguishing facial features, they do not use hairstyle, facial hair, or other similar factors [6] [8].

AU Number	FACS Name
0	Neutral Face
1	inner Brow Raiser
2	Outer Brow Raiser
4	Brow Lowerer
5	Upper Lid Raiser
7	Lid Tightener
8	Lips Towards Each Other
9	Nose
10	Upper Lip Raiser

TABLE II
SAMPLE ACTION UNITS AND ACTION DESCRIPTORS

1) Geometry-based Techniques: The features are extracted by using relative positions and sizes of the important components of face. First, detecting edges, directions of important components or region images contain important components, then building feature vectors from these edges and directions. Using filters such as Canny filter to detect eyes or mouth region of face image, or the gradient analysis method which is usually applied in this direction.

Second, methods are based on the grayscales difference of important components and unimportant components, by using feature blocks, set of Haar-like feature block in Adaboost method[6] to change the grayscales distribution into the feature. In LBP method, it divides up the face image to regions (blocks) and each region corresponds with each central pixel. Then it examine its pixel neighbors, based on the grayscales value of central pixel to change its neighbor to 0 or 1. Therefore, every pixel will be represented in a binary string. Since then, we build histograms for every region. Then these histograms are combined to a feature vector for the face image.

2) Color Segmentation Based Techniques: This approach makes use of skin color to isolate the face. Any non-skin color region within the face is viewed as a candidate for eyes or mouth.

Color based feature extraction

By use Color models such as RGB , YCbCr or HSV with certain range of color pixels, skin region is detected. After getting the skin region, facial features viz. Eyes and Mouth are extracted. The image obtained after applying skin color statistics is subjected to binarization. It is transformed to grayscale image and then to a binary image by applying suitable threshold. This is done to eliminate the hue and saturation values and consider only the luminance part. This luminance part is then transformed to binary image with some threshold because the features for face are darker than the background colors. After thresholding, opening and closing operations are performed to remove noise. These are the morphological operations, which are used to remove holes. Then eyes, ears, nose can be extracted from the binary image by considering the threshold for areas which are darker in the mouth than a

given threshold. So triangle can be drawn with the two eyes and a mouth as the three points in case of a frontal face. And it is easy to get an isosceles triangle (i j k) in which the Euclidean distance between two eyes is about 90-110% of the Euclidean distance between the centre of the right/left eye and the mouth. After getting the triangle, it is easy to get the coordinates of the four corner points that form the potential facial region. Since the real facial region should cover the eyebrows, two eyes, mouth and some area below the mouth, this coordinates can be calculated. The performance of such techniques on facial image databases is rather limited, due to the diversity of ethnical backgrounds.

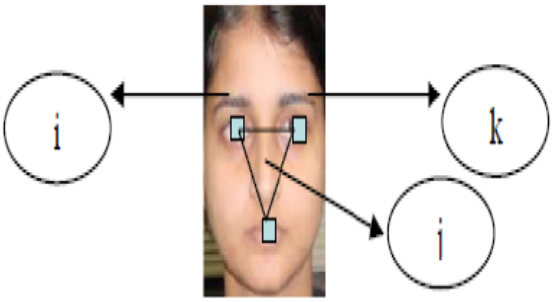


Fig. 4. Color based feature extraction

3) Template Based Techniques: This method group will extract feature of face such as eyes, mouth, etc. based on template function and appropriate energy function. An image region is the best appropriateness with template for eye, mouth or nose, which will minimize the energy. The methods have been proposed such as deformable template and genetic algorithms. In the deformable template method, the feature of interest, an eye, for example, is described by a parameterized template. An energy function is defined to links edges, peaks, and valleys in the image intensity with corresponding properties of the template. Then the template matching is done with the image, by altering its parameter values to minimize the energy function, thereby deforming itself to find the best fit. The final parameter values can be used as descriptors for the features.

IV. System Architecture

Several popular facial-expression databases exist, the majority have been recorded in tightly controlled laboratory environments, where the subjects were asked to generate certain expressions. These lab scenarios are in no way a true representation of the real world. Ideally, we want a dataset of spontaneous facial expressions in challenging real-world environments. To address this problem, we have collected two new facial-expression databases derived from movies via a semiautomatic recommender- based method. We extracted a database of temporal and static facial expressions from scenes in movies, environments that more closely resemble the real world than those of previous datasets. The database contains videos showing natural head poses and movements, close-to-real-world illumination, multiple subjects in the same frame, occlusions, and searchable metadata. The datasets also cover a large age range, including toddler, child, and teenager subjects, which are missing in other currently available temporal facial-expression databases.

Inspired by the Labeled Faces in the Wild (LFW) database, This temporal database is named **Acted Facial Expressions in the Wild (AFEW)** and its static subset **Static Facial Expressions in the Wild (SFEW)**. In this context, in the wild refers to the challenging conditions in which the facial expressions occur rather than spontaneous facial expressions.

A. Labeled Faces in the Wild (LFW) database

Face recognition is the problem of identifying a specific individual, rather than merely detecting the presence of a human face, which is often called face detection. The general term face recognition can refer to a number of different problems including, but not limited to, the following [2].

1) Given two pictures, each of which contains a face, decide whether the two people pictured represent the same individual.
2) Given a picture of a persons face, decide whether it is an example of a particular individual. This may be done by comparing the face to a model for that individual or to other pictures of the individual.
3) Given a picture of a face, decide which person from among a set of people the picture represents, if any.

Labeled Faces in the Wild (LFW), is designed to address the first of these problems, although it can be used to address the others if desired.

The main motivation for the database, is to provide a large set of relatively unconstrained face images. By unconstrained, we mean faces that show a large range of the variation seen in everyday life. This includes variation in pose, lighting, expression, background, race, ethnicity, age, gender, clothing, hairstyles, camera quality, color saturation, focus, and other parameters. The reason we are interested in natural variation is that we are interested in the problem of pair matching given a pair of pre-existing face images, i.e., images whose composition we had no control over.

1) Summary statistics and properties of the LFW database:

- The database contains 13,233 target face images. Some images contain more than one face, but it is the face that contains the central pixel of the image which is considered the defining face for the image. Faces other than the target face should be ignored as background.
- The name of the person pictured in the center of the image is given. Each person is given a unique name (George W Bush is the current U.S. president while George HW Bush is the previous U.S. president), so no name should correspond to more than one person, and each individual should appear under no more than one name (unless there are unknown errors in the database).

- The database contains images of 5749 different individuals. Of these, 1680 people have two or more images in the database. The remaining 4069 people have just a single image in the database.
- The images are available as 250 by 250 pixel JPEG images. Most images are in color, although a few are grayscale only.
- All of the images are the result of detections by the Viola-Jones face detector, but have been rescaled and cropped to a fixed size. After running the Viola-Jones detector on a large database of images, false positive face detections were manually eliminated, along with images for whom the name of the individual could not be identified.
- There are two Views of the database, one for algorithm development, and one for performance reporting.

2) Construction and composition details: The process of building the database can be broken into the following steps [11]:

1) Gathering raw images - The raw images from the Faces in the Wild database collected by Tamara Berg at Berkeley
2) Running a face detector and manually eliminating false positives - A version of the Viola-Jones face detector was run on each image. For each positive detection (if any), the following procedure was performed:
   a) If the highlighted region was determined by the operator to be a non-face, it was omitted from the database.
   b) If the name of the person of a detected face from the previous step could not be identified, either from general knowledge or by inferring the name from the associated caption, then the face was omitted from the database.
   c) If the same picture of the same face was already included in the database, the face was omitted from the database. More details are given below about eliminating duplicates from the database.
   d) Finally, if all of these criteria were met, the face was recropped and rescaled (as described below) and saved as a separate JPEG file.
3) Eliminating duplicate images - Before removing duplicates, it is necessary to define exactly what they are. While the simplest definition, that two pictures are duplicates if and only if the images are numerically equivalent at each pixel. A good deal of effort was expended in removing duplicates from the database.
4) Labeling (naming) the detected people - Each person in the database was named using a manual procedure that used the caption associated with a photograph as an aid in naming the person. It is possible that certain people have been given incorrect names, especially if the original news caption was incorrect.
5) Cropping and rescaling the detected faces - For each labeled face, the final image to place in the database was created using the following procedure. The region returned by the face detector for the given face was expanded by 2.2 in each dimension.
6) Forming sets and pairs for View 1 and View 2 was done using the following process. First, each specific person in the database was randomly assigned to a set. In the case of View 1, each person had a 0.7 probability of being placed into the training set, and in the case of View 2, each person had a uniform probability of being placed into each set.

B. Optical Characater Reader (OCR)

OCR research and development can be traced back to the early 1950s, when scientists tried to capture the images of characters and texts.

1) First generation OCR system: The

rst generation machines are characterized by the constrained letter shapes which the OCRs can read. These symbols were specially designed for machine reading, and they did not even look natural. The

rst commercialized OCR of this generation was IBM 1418, which was designed to read a special IBM font 407. The recognition method was template matching, which compares the character image with a library of prototype images for each character of each font.

2) Second generation OCR system: Next generation machines were able to recognize regular machine-printed and hand- printed characters. The character set was limited to numerals and a few letters and symbols. Such machines appeared in the middle of 1960s to early 1970s. The methods were based on the structural analysis approach.

3) Third generation OCR system: For the third generation of OCR systems, the challenges were documents of poor quality and large printed and hand-written character sets. Low cost and high per- formance were also important objectives. Commercial OCR systems with such ca- pabilities appeared during the decade 1975 to 1985.

4) Fourth generation OCR system: The fourth generation can be characterized by the OCR of complex documents intermixing with text, graphics, tables and mathematical symbols, unconstrained handwritten characters, color documents, low-quality noisy documents, etc.

5) Components of OCR: The principal task of OCR is to develop computer algorithms to identify the char- acters in the text. All OCR implementations consist of a number of preprocessing steps followed by the actual recognition of text as shown in Figure 5.

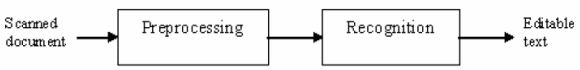


Fig. 5. OCR process

- **Pre-processing** - OCR software often "pre-processes" images to improve the chances of successful recognition. Techniques include :-
  - De-skew If the document was not aligned properly when scanned, it may need to be tilted a few degrees clockwise or counterclockwise in order to make lines of text perfectly horizontal or vertical.
  - Despeckle remove positive and negative spots, smoothing edges.
  - Binarization Convert an image from color or greyscale to black-and-white (called a "binary image" because there are two colours). In some cases, this is necessary for the character recognition algorithm; in other cases, the algorithm performs better on the original image and so this step is skipped.
  - Line removal Cleans up non-glyph boxes and lines
  - Layout analysis or "zoning" Identifies columns, paragraphs, captions, etc. as distinct blocks. Especially important in multi-column layouts and tables.
  - Line and word detection Establishes baseline for word and character shapes, separates words if necessary.
  - Script recognition In multilingual documents, the script may change at the level of the words and hence, identification of the script is necessary, before the right OCR can be invoked to handle the specific script.
  - Character isolation or "segmentation" For per-character OCR, multiple characters that are connected due to image artifacts must be separated; single characters that are broken into multiple pieces due to artifacts must be connected.
  - Normalize aspect ratio and scale.
- **Character Recognition -** There are two basic types of core OCR algorithm, which may produce a ranked list of candidate characters.
  Matrix matching involves comparing an image to a stored glyph on a pixel-by-pixel basis; it is also known as "pattern matching", "pattern recognition", or "image correlation". This relies on the input glyph being correctly isolated from the rest of the image, and on the stored glyph being in a similar font and at the same scale. This technique works best with typewritten text and does not work well when new fonts are encountered. This is the technique the early physical photocell-based OCR implemented, rather directly.
  Feature extraction decomposes glyphs into "features" like lines, closed loops, line direction, and line intersections. These are compared with an abstract vector-like representation of a character, which might reduce to one or more glyph prototypes. General techniques of feature detection in computer vision are applicable to this type of OCR, which is commonly seen in "intelligent" handwriting recognition and indeed most modern OCR software. Nearest neighbour classifiers such as the k-nearest neighbors algorithm are used to compare image features with stored glyph features and choose the nearest match.
- **Post-processing -** OCR accuracy can be increased if the output is constrained by a lexicon a list of words that are allowed to occur in a document. This might be, for example, all the words in the English language, or a more technical lexicon for a specific field. This technique can be problematic if the document contains words not in the lexicon, like proper nouns. Tesseract uses its dictionary to influence the character segmentation step, for improved accuracy.
  The output stream may be a plain text stream or file of characters, but more sophisticated OCR systems can preserve the original layout of the page and produce, for example, an annotated PDF that includes both the original image of the page and a searchable textual representation.
  "Near-neighbor analysis" can make use of co-occurrence frequencies to correct errors, by noting that certain words are often seen together. For example, "Washington, D.C." is generally far more common in English than "Washington DOC".
  Knowledge of the grammar of the language being scanned can also help determine if a word is likely to be a verb or a noun, for example, allowing greater accuracy.

C. Subtitle (Captioning)

Subtitles are derived from either a transcript or screenplay of the dialog or commentary in films, television programs, video games, and the like, usually displayed at the bottom of the screen. They can either be a form of written translation of a dialog in a foreign language, or a written rendering of the dialog in the same language, with or without added information to help viewers who are deaf and hard of hearing to follow the dialog, or people who cannot understand the spoken dialogue or who have accent recognition problems. The encoded method can either be pre-rendered with the video or separate as either a graphic or text to be rendered and overlaid by the receiver. The separate subtitles are used for DVD, Blu-ray and television teletext/DVB subtitling or EIA-608 captioning, which are hidden unless requested by the viewer from a menu or remote controller key or by selecting the relevant page or service (e.g., p. 888 or CC1), always carry additional sound representations for deaf and hard of hearing viewers. Teletext subtitle language follows the original audio, except in multi-lingual countries where the broadcaster may provide subtitles in additional languages on other teletext pages. EIA-608 captions are similar, except North American Spanish stations may provide captioning in Spanish on CC3. DVD and Blu-ray only differ in using run-length encoded graphics instead of text, as well as some HD DVB broadcasts.

1) Creation, delivery and display of subtitles: Today professional subtitlers usually work with specialized computer software and hardware where the video is digitally stored on a hard disk, making each individual frame instantly accessible. Besides creating the subtitles, the subtitler usually also tells

the computer software the exact positions where each subtitle should appear and disappear. For cinema film, this task is traditionally done by separate technicians. The end result is a subtitle file containing the actual subtitles as well as position markers indicating where each subtitle should appear and disappear. These markers are usually based on timecode if it is a work for electronic media (e.g., TV, video, DVD), or on film length (measured in feet and frames) if the subtitles are to be used for traditional cinema film.

The finished subtitle file is used to add the subtitles to the picture, either :

- Directly into the picture (open subtitles);
- Embedded in the vertical interval and later superimposed on the picture by the end user with the help of an external decoder or a decoder built into the TV (closed subtitles on TV or video);
- Or converted (rendered) to tiff or bmp graphics that are later superimposed on the picture by the end user's equipment (closed subtitles on DVD or as part of a DVB broadcast).

Subtitles can also be created by individuals using freely available subtitle-creation software like Subtitle Workshop for Windows, MovieCaptioner for the Mac and Subtitle Composer for Linux, and then hardcode them onto a video file with programs such as VirtualDub in combination with VSFilter which could also be used to show subtitles as softsubs in many software video players.

2) Same-language captions: Same-language captions, i.e., without translation, were primarily intended as an aid for people who are deaf or hard of hearing. Internationally, there are several major studies which demonstrate that same-language captioning can have a major impact on literacy and reading growth across a broad range of reading abilities. This method of subtitling is used by national television broadcasters in China and in India such as Doordarshan. This idea was struck upon by Brij Kothari, who believed that SLS makes reading practice an incidental, automatic, and subconscious part of popular TV entertainment, at a low per-person cost to shore up literacy rates in India.

Closed Captions

Closed captioning is the American term for closed subtitles specifically intended for people who are deaf and hard of hearing. These are a transcription rather than a translation, and usually contain descriptions of important non-dialog audio as well such as "(sighs)" or "(door creaks)" and lyrics. From the expression "closed captions" the word "caption" has in recent years come to mean a subtitle intended for the hard of hearing, be it "open" or "closed". In British English "subtitles" usually refers to subtitles for the hard of hearing (HoH); however, the term "HoH subtitles" is sometimes used when there is a need to make a distinction between the two.

SDH

"SDH" is an American term the DVD industry introduced. It is an initialism for "Subtitles for the deaf or hard-of-hearing", and refers to regular subtitles in the original language where important non-dialog information has been added, as well as speaker identification, useful when the viewer cannot otherwise visually tell who is saying what.

The only significant difference for the user between "SDH" subtitles and "closed captions" is their appearance: SDH subtitles usually are displayed with the same proportional font used for the translation subtitles on the DVD; however, closed captions are displayed as white text on a black band, which blocks a large portion of the view. Closed captioning is falling out of favor as many users have no difficulty reading SDH subtitles, which are text with contrast outline. In addition, DVD subtitles can specify many colors, on the same character: primary, outline, shadow, and background. This allows subtitlers to display subtitles on a usually translucent band for easier reading; however, this is rare, since most subtitles use an outline and shadow instead, in order to block a smaller portion of the picture. Closed captions may still supersede DVD subtitles, since many SDH subtitles present all of the text centered, while closed captions usually specify position on the screen: centered, left align, right align, top, etc. This is very helpful for speaker identification and overlapping conversation. Some SDH subtitles (such as the subtitles of newer Universal Studios DVDs/Blu-ray Discs) do have positioning, but it is not as common.

DVDs for the U.S. market now sometimes have three forms of English subtitles: SDH subtitles; English subtitles, helpful for viewers who may not be hearing impaired but whose first language may not be English (although they are usually an exact transcript and not simplified); and closed caption data that is decoded by the end-user's closed caption decoder. Most anime releases in the U.S. only include as subtitles translations of the original material; therefore, SDH subtitles of English dubs ("dubtitles") are uncommon.

High-definition disc media (HD DVD, Blu-ray Disc) uses SDH subtitles as the sole method because technical specifications do not require HD to support line 21 closed captions. Some Blu-ray Discs, however, are said to carry a closed caption stream that only displays through standard-definition connections. Many HDTVs allow the enduser to customize the captions, including the ability to remove the black band.

D. XML

Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable. It is defined by the W3C's XML 1.0 Specification and by several other related specifications, all of which are free open standards.

The design goals of XML emphasize simplicity, generality and usability across the Internet. It is a textual data format with strong support via Unicode for different human languages. Although the design of XML focuses on documents, it is widely used for the representation of arbitrary data structures such as those used in web services.

Several schema systems exist to aid in the definition of XML-based languages, while many application programming

interfaces (APIs) have been developed to aid the processing of XML data.

1) Applications of XML: As of 2009, hundreds of document formats using XML syntax have been developed, including RSS, Atom, SOAP, and XHTML. XML-based formats have become the default for many office-productivity tools, including Microsoft Office (Office Open XML), OpenOffice.org and LibreOffice (OpenDocument), and Apple's iWork. XML has also been employed as the base language for communication protocols, such as XMPP. Applications for the Microsoft .NET Framework use XML files for configuration. Apple has an implementation of a registry based on XML.

XML has come into common use for the interchange of data over the Internet. IETF RFC 7303 gives rules for the construction of Internet Media Types for use when sending XML. It also defines the media types application/xml and text/xml, which say only that the data are in XML, and nothing about its semantics. The use of text/xml has been criticized as a potential source of encoding problems and it has been suggested that it should be deprecated.

E. Acted Facial Expressions in the Wild (AFEW)

Until now, researchers have manually collected all facial-expression databases, which is time consuming and error prone. To address this limitation, a video clip recommender system based on subtitle parsing. Rather than manually scan a full movie, our labelers reviewed only the video clips suggested by the recommender system, which searched for clips with a high probability of a subject showing a meaningful expression. This method lets us collect and annotate large amounts of data quickly. Based on the availability of detailed information regarding the movies and their content on the Web, the labelers then annotated the video clips with dense information about the subjects. An XML-based representation for the database metadata, which makes it searchable and easily accessible using any conventional programming language [1].

For the AFEW dataset, video clips where labelled with one of six basic expressions: anger, disgust, fear, happiness, sadness, surprise, or neutral. The database captures facial expressions, natural head pose movements, occlusions, subjects races, gender, diverse ages, and multiple subjects in a scene.

Although movies are often shot in somewhat controlled environments, they are significantly closer to real-world environments than current lab-recorded datasets. AFEW is not a spontaneous facial-expression database. However, method actors attempt to mimic real-world human behavior to give audiences the illusion that they are behaving spontaneously, not posing, in movies.

The AFEW dataset, in particular, addresses the issue of temporal facial expressions in difficult conditions that are approximating realworld conditions, which provides for a much The AFEW dataset, in particular, addresses the issue of temporal facial expressions in difficult conditions that are approximating realworld conditions, which provides for a much

1) Database Creation: To construct the database, a semi-automatic approach was followed and divided the process into two parts. First, the subtitles are extracted and parsed in the recommender system. Second, a human labeler annotates the recommended clips based on information available on the Internet.

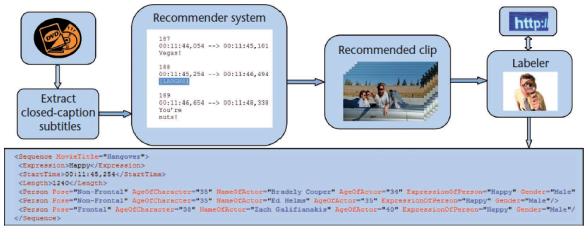


Fig. 6. Database creation process

Subtitle Extraction

Different movie DVDs was analyzed. Subtitles for the deaf and hearing impaired (SDH) and closed caption (CC) subtitles where extracted from the DVDs because they contain information about the audio and nonaudio context such as emotions and information about the actors and scene (for example, [CHEERING], [SHOUTS], and [SURPRISED]). The subtitles from the movies can be exracted using the Vob- Sub Rip (VSRip) tool. The extracted subtitle images were parsed using optical character recognition (OCR) and converted into the .srt subtitle format using the Subtitle Edit tool. The .srt format contains the start time, end time, and textual content with millisecond accuracy.

Video Recommender System

Once the subtitles have been extracted, we parse the subtitles and search for expressionrelated keywordsfor example, happy, sad, surprised, shouts, cries, groans, cheers, laughs, sobs, silence, angry, weeping, sorrow, disappoint, and amazed. If found, the system recommends video clips to the labeler. The clips start and end time is extracted from the subtitle information. The system plays the video clips sequentially, and the labeler enters information about the clip and its characters and actors from the Web.

If clips contain multiple actors, the labeling sequence is based on two criteria. For actors appearing in the same frame, the order of annotation is left to right. If the actors appear at different timestamps, then it is in the order of appearance. The dominating expression in the video is labeled as the theme expression. The labeling is then stored in an XML metadata schema. Finally, the labeler enters the characters age or his or her estimated age if this information is unavailable.

In total, the subtitles from the 54 DVDs contained 77,666 individual subtitles. Out of these, the recommender system suggested 10,327 clips corresponding to subtitles containing expressive keywords. The labelers chose 1,426 clips from these on the basis of criteria such as the visible presence of subjects,

at least some part of the face being visible, and the display of meaningful expressions.

Because subtitles are manually created by humans, they can contain errors. This might lead to a situation where the recommender system suggests an erroneous clip. However, the labelers can reject a recommendation. When annotating the clips, the labelers use the clips video, audio, and subtitle information to make informed decisions. The proposed recommender system can easily add more clips to the database and scale it up in the future.

Database Annotations

Database contains metadata about the video clips in an XML-based schema, which enables efficient data handling and updating. The human labelers densely annotated the video clips with the expression and subject information. The subject information contains various attributes describing the actor and/or character in the scene:

- Pose. This denotes the head pose based on the labelers observation. In the current version, we manually classify the head pose as frontal or nonfrontal.
- Character age. Frequently, only the age of the lead actors characters are available on the Web. The labeler estimated any other ages.
- Actor name. Here we provide the actors real name.
- Actor age. The labelers extracted the actors real ages from www.imdb.com. In a few cases, the age information was missing, so the labeler estimated it.
- Expression of person. This denotes the expression class of the character as labeled by the human observer. This could differ from the higher-level expression tag because there might be multiple people in the frame showing different expressions with respect to each other and the scene/theme.
- Gender. Provides the actors gender.

Expression tag specifies the theme expression conveyed by the scene. The expressions were divided into the six expression classes, plus neutral. The default value is based on the search keyword found in the subtitle textfor example, we use happiness for smile and cheer. Human observer can change it based on their observation of the audio and scene in the clip.

XML-based metadata schema has two major advantages. First, it is easy to use and search using any standard programming language on any platform that supports XML. Second, the structure makes it simple to add new attributes about the video clips in the future, such as the pose of the person in degrees and scene information, while keeping the existing data and ensuring that pre-existing tools can exploit this information with minimal changes.

Extracting features and classifying

To extract a face, we computed the Viola-Jones detector. Then we compute feature descriptors on the cropped faces from all the databases. The cropped faces were divided into 4 X4 blocks for local binary pattern (LBP), local phase quantization (LPQ), and pyramid histogram of gradients (PHOG). For LBP and LPQ, set the neighborhood size to eight. Then apply principal component analysis (PCA) on the extracted features. For classification, a support vector machine (SVM) learned model is used.

2) Database Contributions: The AFEW and SFEW databases offer several novel contributions to the state of the art. AFEW is a dynamic, temporal facial-expression data corpus consisting of short video clips of facial expressions in close-to-real-world environments. SFEW is also the only static, tough conditions database covering the seven facial-expression classes [12].

Subjects ranged from 1 to 70 years old, which makes the resulting datasets generic in terms of age, unlike other facial-expression databases. The databases have many clips depicting children and teenagers, which can be used to study facial expressions in younger subjects. The datasets can also be used for both static and temporal facial age research.

AFEW is currently the only facial-expression database with multiple labeled subjects in the same frame.

The databases also exhibit close-to-real illumination conditions. The clips include scenes with indoor, nighttime, and outdoor natural illumination. Although movie studios use controlled illumination conditions, even in outdoor settings, these are closer to natural conditions than lab-controlled environments and, therefore, are valuable for facial-expression research. The diverse nature of the illumination conditions in the dataset makes it useful for not just facial-expression analysis but potentially also for facial recognition, facial alignment, age analysis, and action recognition.

The movies where chosen over a large set of actors. Many actors appear in multiple movies in the dataset, which will enable researchers to study how their expressions have evolved over time, whether they differ for different genres, and so forth.

The design of the database schema is based on XML. This enables further information about the data and its subjects to be added easily at any stage without changing the video clips. This means that detailed annotations with attributes about the subjects and the scene are possible.

V. Conclusion

Facial-expression analysis is a well-researched field. However, progress in the field has been hampered due to the unavailability of databases depicting real-world conditions.This is due to a lack of robust in the wild face-alignment methods and efficient temporal descriptors. Expression analysis in close-to-real-world situations is a nontrivial task and requires more sophisticated methods at all stages of the approach, such as robust face localization and tracking, illumination, and pose invariance.

References

[1] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, *Collecting Large, Richly Annotated Facial-Expression Databases from Movies*, IEEE Computer Society, 2012

[2] G.B. Huang et al., *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*, Univ. of Massachusetts, Amherst, 2007.

[3] A. Dhall et al., *Static Facial Expression Analysis in Tough Conditions: Data, Evaluation Protocol and Benchmark*, IEEE Int'l Conf. Computer Vision Workshop (BeFIT), IEEE Press, 2011
[4] R. Gross et al., *Multi-PIE*, 8th IEEE Intl Conf. Automatic Face and Gesture Recognition (FG), 2008
[5] M. Pantic et al., *Web-Based Database for Facial Expression Analysis*, Proc. IEEE Intl Conf. Multimedia and Expo (ICME), IEEE CS Press, 2005
[6] M.S. Bartlett et al., *Automatic Recognition of Facial Actions in Spontaneous Expressions*, J. Multimedia, vol. 1, no. 6, 2006
[7] E. Douglas-Cowie, R. Cowie, and M. Schroder, *A New Emotion Database: Considerations, Sources and Scope*, Proc. ISCA ITRW on Speech and Emotion, 2000
[8] M.J. Lyons et al., *Coding Facial Expressions with Gabor Wavelets*, Proc. IEEE Intl Conf. Automatic Face Gesture Recognition and Workshops (FG), IEEE CS Press, 1998
[9] F. Wallhoff, *Facial Expressions and Emotion Database*, 2006, www.mmk.ei.tum.de/ waf/fgnet/ feedtum.html.
[10] I. Laptev et al., *Learning Realistic Human Actions from Movies*, Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), IEEE CS Press, 2008
[11] P.A. Viola and M.J. Jones, *Rapid Object Detection Using a Boosted Cascade of Simple Features*, Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), IEEE CS Press, 2001
[12] D. Huang et al., *Local Binary Patterns and its Application to Facial Image Analysis: A Survey*, IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 41, no. 6, 2011
[13] M. S.Bartlett, J. R. Movellan, P. Ekman, G. Donato, J. C. Hager, T. J. Sejnowski, *Image representations for facial expression coding*

Analyzing Digital Personas in Cybercrime Investigations

Vinitha Mathew
Eighth Semester, 2011 Admission
Department of Computer Science and Engineering,
Sreepathy Institute of Management & Technology, Vavannoor, Palakkad, India-Pin 679533
E-mail: vinithamathew34@gmail.com

***Abstract*— Online cybercrime activities often involve criminals hiding behind multiple identities (so-called digital personas).The Isis toolkit offers the sophisticated capabilities required to analyze digital personas and provide investigators with clues to the identity of the individual or group hiding behind one or more personas .Digital communities not only bring people closer together but also, inadvertently, provide criminals with new ways to access potential victims online. Digital personas play a key role in criminal tactics in online social media. One criminal might hide behind multiple digital personas or a group of criminals might share a single persona when engaging with potential victims .It develop a natural language analysis techniques to help Identify child sex offenders from chat logs and monitoring mechanisms that can be noninvasively attached to file sharing systems for identifying the distributors of child abuse media.**

I. Introduction

Digital communities not only bring people closer together but also, inadvertently, provide criminals with new ways to access potential victims online. Digital personas play a key role in criminal tactics in online social media. One criminal may hide behind multiple digital personas or a single persona may be shared by a criminal group when engaging with potential victims. Furthermore, the fluid nature of identity on online social media means that criminals can disguise themselves with relative ease to gain the trust of potential victims. Examples of such criminal exploitation of digital personas include:

- Child sex offenders masquerading as young persons to gain the trust of their victims. An offender may use multiple personas over the course of an interaction (introducing himself/herself as a young person and then introducing another persona, e.g., that of an older relative). Alternatively, a single persona may be shared by an offender group so that a victim is groomed by multiple people over a period of time [1].
- Romance scam operators using digital personas with appropriate age and gender to engage with multiple victims in online dating sites, gaining their trust and exploiting them for financial gain [2].
- Radicalization of youth in online forums through persuasive messaging [3]. Multiple digital personas are used as a tactic at times. For instance, one persona is used to vigorously support a radical cause, followed by silence for a few days and then a different persona is used to claim that the original protagonist has left to fight for the cause.

Effective policing of such environments is, however,extremely challenginga vast amount of information is communicated within online social media, making manual analysis difficult or even impossible. Consequently, law enforcement agencies face huge online communication data analysis backlogs during cybercrime investigations, with backlogs of six to nine months being commonplace. Even though a range of commercial tools such as EnCase (www.guidancesoftware.com/encase-forensic.htm) and Internet Evidence Finder (www.magnetforensics.com/products/internet-evidence-finder) can assist in such investigations, they mainly focus on data extraction. Any analysis of the data is left to the investigator, who has access only to simple techniques such as keyword-based searches or phrase detection based on user-defined lists.
Such techniques do not scale, and they do not include models of deceptive behavior or sharing of online personas. It is not uncommon for investigators to extract data from hard disks or mobile phones using a tool such as EnCase and then manually read it to identify when an offense might have occurred and make a value judgment about whether one or more digital personas were used as part of the offenders tactics. Given the large amounts of text and number of online participants during such investigations, it is virtually impossible for the investigator to analyze all digital personas involvedthe cognitive load is immense.

Given the vast amount of information that is communicated within online social networks, new monitoring and analysis technologies need to be developed in order to tackle the growing problem of child grooming and the distribution of child abuse media. The development of such technologies faces three significant research challenges:

- How to identify active child sex offenders across online communities?
  Paedophiles and other child sex offenders often masquerade as children in order to establish contact with potential victims and gain their trust. Distinguishing the innocent

interaction amongst children or amongst children and adults from such predatory advances is a non-trivial task yet effective, early and accurate identification of sexual offenders is vital for the protection of children. At the same time such offenders may use multiple online identities and known child sex offenders may move to other online social networks upon detection in one network. It is, therefore, vital that once a suspected child sex offender is detected in one network, s/he can be successfully detected in other networks which s/he may attempt to employ for grooming children.

- How to identify the core distributors of child abuse media?
  The key research challenge is to accurately identify child abuse media from the plethora of perfectly legal material that exists within file sharing systems. The problem is compounded by the fact that offenders often use specialized vocabulary to describe their shared mediaa vocabulary that evolves and changes over timeand operate over different file sharing networks. Any monitoring framework must be noninvasively attachable to existing file sharing systems given the wealth of such systems and clients available today. In addition to identifying child abuse media within such systems, any monitoring tools must be able to distinguish core distributors of such media from mere users. This is essential for child protection as this would help law enforcement agencies in tackling the problem at its roots.

- How to ensure that such developments maintain ethical practices?
  The development of such monitoring and analysis techniques raises a number of ethical challenges pertaining, on the one hand, to utilizing the framework and tools in a beneficial way for child protection and, on the other hand, the need to protect innocent users of online social networks from the potential of falsely being identified as child sex offenders and safeguarding their privacy.

Isis is aiming to tackle the above three challenges by developing novel chat log analysis and non-invasive file sharing monitoring techniques based on natural language processing and aspect-oriented programming [4] practices respectively. The resulting framework and tools will assist law enforcement agencies while ensuring that they fall within current ethical boundsnote our goal is not automation but to provide support for detecting potential sexual offences through analysis of large amounts of data which cannot manually be analyzed in an efficient manner.

II. Motivation

Recently, I came across a news paper article that, a 15 years old girl Keerthana Ragesh was sexually abused through online social media. I was shocked to know that my friend was also sharing about the same brutal thing happened to an another gir.Painfully, I got into the topic and could find numerous of such pain exploitations faced by the poor teenage girls,even in our highly literated kerala. Likewise,almost all days we used to see atleast one fake message that we won 1 million dolors or BMW car etc.Unfortunately most of the youngsters are running behind these illusions and provides personal information such as bank account numbers,address,phonenumber,personal passwords etc and and they become money loosers.From these, I got some inspiring sparks and a thirst to prevent such exploitation and protect the blossming buds of of our nation,India. This seminar is an attempt to analyze the prevalence of girl child abuse in India. News papers of three national dailies and Crime Records Bureau sites of Kerala (SCRB) and India (NCRB) were referred for the purpose of obtaining relevant data. It was found that there is an alarming rate of increase in the sexual abuse of girl children in Kerala. Based on the available information from the above media the study also tried to examine the details of the victim, the abuser/offender and news paper reports. A great majority of children who are exposed to sexual abuse are done so by someone they may know or not thorw various online social medias. The interim report found that 2,409 children and young people were confirmed victims between August 2013 and October 2014. A further 16,500 children were at 'high risk'. of sexual exploitation between April 2013 and March 2014. A societal fear of sex offenders and their presence online has received much attention in sociology research over the past ten years. Generally, there is concern that sex offenders utilize the Internet to gain access to young victims, lurking in online locations typically accessed by children or young people. There is significant concern that the information divulged on social networking sites is being used by sex offenders to identify potential victims.
Research in sociology and psychology has addressed social aspects of these technology-facilitated crimes through the study of the vulnerabilities of children and youth to the threat of online sexual solicitation.

A. Related Work

Relevant research on mining and analysis of information from online social media has mainly focused on extracting key messages prevalent in such media. Davulcu et al. [4] focus on detecting sentiment markers that indicate radicalisation and counter messages in online forums. Diesner and Carley [5] have shown how common word use across actors can be used to derive knowledge about the structure of covert social networks and their weak points. Other recent work has shown that clustering of individuals in online communities is not driven by homophily [6] and that it is possible to gain deeper insights through analysis of latent structures in online conver- sations [7]. In recent years, techniques from the fields of corpus-based natural language processing and text mining have been applied to these problems. Corpus analysis, particularly at the semantic level, can provide a way of describing the key features in extremist discourse [8] and

authorship attribution enables automatic identification of a given writer or speaker [9]. Analysis of digital per- sonas and the deception tactics inherent therein have not been considered to date. In this article, we present the Isis toolkit3 which addresses this particular challenge by enabling efficient and so- phisticated analysis of digital personas in large-scale online textual communications. Our approach complements recent research such as that by Afroz et al. [10], which highlights the difficulty in identifying authorship when language is in- tentionally obfuscated and that of Narayanan [11] which shows that it is viable to automatically predict text authorship on a large-scale. Our work shows that it is possible to predict key attributes of a persona (i.e. age and gender) with acceptable accuracy regardless of whether the author is obfuscating the language.

B. Internet Evidance Finder(IEF)Tool

Internet Evidence Finder is an easy-to-use software application which enables you to detect and remove old or unnecessary files, in order to declutter your computer and recover some free space. It comes packed with advanced, yet intuitive options that should meet the requirements of all user levels, whether they have previous experience with this type of software or not. Setting up the application is done quickly and effortlessly. Its interface is represented by a regular window with a wizard-like layout, where you can perform either a quick or full scan on your computer. Internet Evidence Finder handles chat messages in various social networking and instant messaging services, such as Facebook (chat messages and wall posts), Yahoo!, GoogleTalk, AIM and MySpace, Twitter statuses, and others. During the scanning job you can view total megabytes searched and remaining, along with elapsed and remaining time. The search summary shows the item types to be evaluated, together with the source and output folder (for creating the log file with recorded activity). Furthermore, you can load the Internet Evidence Finder report viewer.

The application runs on low CPU and RAM, so it does not affect the overall performance of the computer. Since it can be minimized to the system tray area, it does not interrupt normal user activity either. It has a good response time and may take a while to complete a scan job. However, no error dialogs were shown in our tests, and the utility did not hang or crash. Too bad that Internet Evidence Finder has not been updated for a while[23].

C. Encase

EnCase Forensic is the global standard in digital investiga- tion technology for forensic practitioners who need to conduct efficient, forensically-sound data collection and investigations using a repeatable and defensible process.It discover how an Internet Crimes Against Children (ICAC) investigator uses EnCase Forensic v7 to investigate computer crimes. Keith Vincent, a member of Chesterfield Police Department's Special Victim's Unit (SVU) discusses his workflow and investigative process in handling cases involving images and requiring a variety of search techniques to uncover the most crucial potential evidence. Keith specifically discusses an arson case where the evidence identified through EnCase Forensic v7 on suspect's computer led to a conviction[27].

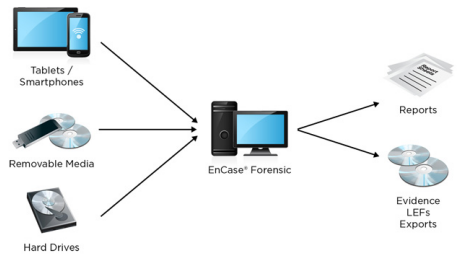


Fig. 1. EnCase Forensic model

III. Theoretical Background

A. Role of the Internet

Officials believe that the Internet is facilitating the growth in the number of sex abusers of children (Brennan, 2013). This may be one of the reasons which explain why police believe that the number of female sex abusers of children is increasing. The Internet is therefore a key component to exploring increases in female perpetrators of child sexual abuse and will be explored in detail in the latter part of the research project. The issue of whether technology is driving cases in female paedophilia will be discussed or whether it is the case that paedophiles would offend regardless of access to technology[27]. Having set the issue of female sex offenders within its wider context the following chapter will provide an explanation of the methodology used to carry out the study.

B. Sex offending and the Internet

It has been argued that the Internet has facilitated the formation of a new category of sex offender, the Internet sex offender (Sheldon and Howitt, 2011). There are many ways that the Internet can be utilised for sexual means. It is important to state that the Internet can be used for non-deviant sexual means. For example, it may allow those from the gay community to communicate with one another (Cooper and Sportolari, 2013). Chat rooms are one of the modes of the Internet used by individuals to contact others from all over the world who possess common interests. Despite the fact that the Internet will not solely be used for negative sexual purposes, there is the possibility for deviant interests to be validated and encouraged if like minded users associate via this mode of communication[24].

C. Child Pornography

One of the most documented ways sex offenders misuse the Internet is to access and download child pornography. Accessing child pornography is argued to be a key factor in the development of sexual offending, with 40 percent of those arrested for accessing pornographic images of children also found to have sexually abused children (Wolak et al, 2005 cited in Jewkes, 2007). However, there is a controversial, conflicting argument put forward that online child pornography may actually prevent offenders carrying out contact offences on children. Holmes et al stated that they expected to see the emergence of offenders for whom there has been no direct contact with children, and whose crimes are exclusively related to the Internet (ibid, 2008/ cited in Calder, 2014).

Due to the many arguments surrounding this area and whether it has a direct causation in relation to contact offences, it is important to discuss the role of the Internet in facilitating access to these images. This chapter will then look at other ways in which sex offenders may become physical perpetrators of CSA as a result of using the Internet. Here the Vanessa George case will also be explored, looking at the potential reasons why the Internet may have led her to contact offending[23].

1) Definitions: Compared to the pre-Internet era it is argued that the Internet has caused more people to access and view such images (Carr, 2004). The term child pornography itself is problematic in terms of that what it may encompass due to its cultural specificity[13]. For example, in Denmark child pornography was decriminalised in 2009. Therefore its definition would not have included the illegality of such material as it would have in other countries during that time. It was recriminalised in the 2011s in Denmark however it has been suggested that many of the images which are in circulation on the Internet are in fact images which were developed during this period of decriminalisation[20].

The term Internet sex offending can include a variety of behaviours all involving child pornography, from trading to producing such images (Jewkes and Yar, 2009). With the technological development of the Internet some argue that this led to the increase in the quantity of this material[16]. However others believe that the quantity may not have increased but that it is now more widely available and that many images are copied (ODonnell and Milner, 2007). The vast amount of images in circulation makes policing this issue very complex as many sites have restricted access or have specific codes and passwords which must be entered before visiting a site or downloading an image. However, one way in which offenders are tracked down is through their credit card details as often offenders will have used these in order to access illegal sites. Some users may believe that in downloading images of children they are not actually committing CSA as they are not physically abusing a child. However, one must consider the victims in these images which Internet sex offenders view[17].

D. How important is technology?

On the other hand, others may argue that technology is irrelevant in the occurrence of sex offending and that if someone is going to sexually offend against children it will happen with or without the involvement of the Internet. This argument however does not correspond to the case of Vanessa George where, without use of the Internet the crimes could not have occurred in the manner in which they did. Colin Blanchard would not have encountered the women who he did not know and who were from all areas of the country that he then encouraged to sexually offend against children. Apart from one of the four women involved, the first time he and the other women came into physical contact was in the court when their trial took place. This begs the question as to whether the Internet may in fact be driving contact offences of CSA, even for those who do not display a prior sexual interest in children (Carr, 2004).

1) Characteristics of the Internet: This section will discuss how the specific characteristics of the Internet may lead some females to engage in CSA. Three key and unique aspects of the Internet are that it is both accessible and affordable whilst also having the crucial component of allowing users to be comparatively anonymous (Quayle and Taylor, 2005). Demetriou and Silke (2003) illustrate how the anonymous element of the Internet, along with the availability of deviant content, has the potential to facilitate a high level of illegal online activity. They set up a game site advertising free games but also a fake link to hardcore porn and they found that more people clicked on the hardcore porn link than the free games (ibid, 2003 cited in ODonnell and Milner, 2007). The degree of anonymity afforded to the Internet may evoke a false sense of security within individuals causing them to say and do things in the virtual world that they would not do in reality. This is known as the Mardi Gras phenomenon (Fowler, 2007 cited in Gillespie 2000), a term often utilised by psychologists. In this case it describes how Internet users feel that they are wearing a mask thus allowing them to act incognito. Users are able to assume a variety of online personalities including deviant ones, as they do not have to fearthe same reprisal as they may do should they act in such a way in daily life. In addition, online they will be connected to other users with similar interests. However, the case of Colin Blanchard questions this phenomenon as the women did not solely indulge his deviant requests online, through chat room conversations, but they became perpetrators of CSA in reality in their offline worlds. Many may find the reasons for the womens compliance incomprehensible however it is important to consider how the Internet may have driven this extreme form of deviance.

IV. Deadly combination of factors leads to childrens vulnerability

Child sex offenders are grooming children over the internet for the sole purpose of online sexual abuse in an alarming new trend, highlighted by the Child Exploitation and Online Protection (CEOP) Centre. A deadly combination of factors leads to some children being particularly at risk from online grooming. Parental or carer involvement in a childs online life can make the crucial difference between a child being protected CEOP and University of Birmingham research says. Offenders may target hundreds of children at a time in order to satisfy their sexual fantasies and once initial contact is made this often rapidly escalates into threats and intimidation. Children who are groomed into performing sexual activity online can feel ashamed, that they lost control, desperate or even suicidal. There were 1,145 public reports in 2012 relating to incidents of online grooming. 7A cluster of grooming reports from the public in 2010 led to a CEOP-led international investigation called Operation Hattie, spanning 20 months and 12 countries. This led to the arrest and conviction in December 2012 of two brothers in Kuwait who had targeted 110 children worldwide, including 78 in the UK, and forced them into performing sexual acts online. There was no evidence of an offline meeting with victims ever being a motivation. Factors which make children vulnerable to contact abuse also make them more vulnerable to online abuse when combined with frequent internet access:

- Personal issues; low self-esteem, confusion about their sexuality and loneliness.
- Social isolation; perhaps through problems/dissatisfaction at school with limited support from their peer group or family.
- Lack of parental monitoring or involvement in online activities; coupled with factors such family problems.

Risk-taking by young people is the key factor in their vulnerability to grooming and potential contact with child sex offender. However, children whose internet activities are monitored and who have an open dialogue with their parents/carers about what they do or see online are better protected from grooming and more resilient to the techniques used by offenders.

Adolescents who take risks online by having sexualised chats or exchanging sexual images are particularly prone to the increasingly sophisticated, coercive and sinister tactics of online predators[24].

Smart phone ownership has increased by 21% among 12-15 year olds in just a year and six out of ten (62%) now have one. With built-in cameras, these devices and a new generation of apps are giving children the ability to easily communicate with strangers online and share images on the move. The Centre also knows that instant messaging on smart phones and other devices is a popular method of communicating and is used by groomers to approach potential victims. Instant messaging was used by offenders to make contact with children in around third of public reports of grooming in 2012/13.

Over two thirds (69%) of parents of 12-15 year-olds with a phone that can be used to go online do not have mobile phone parental controls or filters. This compares to the one in two parents (49%) who have technical controls in place for their childs PC, laptop, or netbook at home.

On Safer Internet Day, CEOP is urging parents/carers to open up a dialogue with their children about their online lives and visit www.thinkuknow.co.uk/parents which has a newly revised parents and carers area, supported by Visa Europe. Parents can find advice on talking about difficult and sensitive subjects, a number of short films on specific dangers faced by young people online and practical easy steps for them to protect their loved ones[21].

CEOP also encourages parents to download a new free CEOP app for Windows 8 users that has been developed with Microsoft. This makes it easier than ever for users to access CEOPs online safety advice pages, or make a report about suspicious or inappropriate contact online. Available from the Windows Store the app allows parents and children to quickly explore CEOPs award winning educational videos, see the latest campaigns or follow CEOPs Facebook page or Twitter feed updates. Peter Davies, Chief Executive at CEOP, said: On a daily basis we see the devastation caused to young peoples lives by online grooming. What we are seeing is that for a growing proportion of grooming cases reported to the Centre, online abuse is an end in itself. UK children can be targeted from anywhere and offenders will cast their net widely to target large numbers of children. Things can quickly spiral out of control for victims[19].

Children may be targeted because of their vulnerability but any child can be a victim. What is apparent is that parents and carers can make that vital difference in whether or not a child becomes a victim of these ruthless predators online. Claire Lilly, safer internet lead at the NSPCC, said:

The internet is part and parcel of young lives and most cant remember a world before it existed. We cannot put the genie back in the bottle, but we can talk to young people and educate them on staying safe online just as we do about stranger danger or drugs. We are seeing a sharp rise in young people contacting ChildLine about being approached online, sending images to strangers or being exposed to online pornography. And a new generation of smart phone apps are presenting yet more problems. CEOP are doing a great job in tracking down ever more sophisticated offenders and technology companies are starting to improve their safeguards but this problem will not go away until everyone ISPs, mobile phone companies, parents, schools and young people themselves play their part in tackling it.

A. Related Stories

Two brothers, who worked together to use the internet to sexually abuse 110 children around the world, including 78 in the UK, have been jailed in Kuwait. Mohammed Khalaf

Al Ali Alhamadi, 35, and Yousef Al Ali Alhamadi, 27, targeted victims aged 12-16. The pair were jailed at a court following an investigation led by Britain's Child Exploitation and Online Protection Centre (Ceop). They were convicted of blackmail relating to sex abuse offences.
The pair often pretended to be someone the children already knew on social networking and instant messaging applications, They would trick victims into giving them their online passwords using a link, before threatening them into "engaging in sexual activities via webcam," Ceop said.
Following the arrest of the abusers, Ceop worked with the Kuwaiti authorities, international police forces and child protection agencies in Australia, Canada, Cyprus, Denmark, Iceland, Ireland, Jersey, the Netherlands, New Zealand, Portugal, Sweden and the United States.

B. Partners

The Isis project team brings together experts from the areas of online social network monitoring,natural language analysis,child protection,ethics and ethnographic/user studies.

- Lancaster University
- Middlesex University
- Swansea University
- Child Exploitation and Online Protection center

In addition,local schools,child safety experts and law enforcement agencies from across the UK have helped guided and test the technologies developed within the project.

For more information about the Isis project contact:
Awais Rashid
Computing Department
InfoLab 21
Lancaster University
Lancaster
LA1 4WA
UK
Email:isis_contactcomp.lancs.ac.uk
Tel:+44(0)1524510316

V. System Architecture

A. THE ISIS TOOLKIT

As Figure 1 shows, the Isis toolkit combines statistical methods from corpus-based natural language processing with authorship attribution tools. Analysis techniques from corpus linguistics and natural language processing, such as keyword profiling, offer the capability to compare word frequencies. Previous work extended this approach to extract key grammatical categories (equating to features of style) and key semantic fields (showing key concepts)[12]. These techniques use large representative samples of writing or transcribed speech for training and reference comparison, have high accuracy, and are designed to be robust across various types of text. Using tools and methods from the authorship attribution field makes it possible to narrow the focus from language varieties down to the individual writer to identify the authors stylistic fingerprint. In the past, authorship attribution techniques were mainly applied to determine the authorship of historical texts. Recently, more robust evaluation techniques have been developed, and authorship attribution methods have been applied to known problems with standard benchmark data[9]. The specific challenges that we faced in implementing the Isis toolkit included integrating the statistically sophisticated but knowledge-poor techniques from authorship attribution with linguistically informed methods from corpus-based natural language analysis, combining the macro level (models of language varieties) with the micro level (models of individuals use of language). Additionally, these methods must operate on small quantities of noisy language data observed in online social networks and deal with masquerading or similarly deceptive behavior that an individual might assume in an attempt to hide his or her identity. The novel investigative features of the Isis toolkit include the following:

- Establish a stylistic language fingerprint of potential suspects or victims. These fingerprints can then be overlaid on each other and compared to study whether one person might be hiding behind a single persona or if multiple people are sharing a single persona.
- Establish the age and gender of the person behind a digital persona. Isis achieves this by synthesizing the stylistic fingerprint and extracting additional markers using a natural-language-analysis engine. Furthermore, the toolkit can detect masquerading tactics with a high degree of accuracyfor example, detecting when an adult is masquerading as a child.
- Establish online interaction patterns of particular digital personas. Isis analyzes both the conversation structure and the language used therein to determine a specific personas key characteristics such as signature moves when signing off from a conversation or frequently used words and phrases. The toolkit also can analyze a personas behaviorfor example, identifying when a participant is typically activenot only within an average 24-hour period but also in terms of day of the week. It also can determine whether a persona becomes increasingly sexual or aggressive over a period of days or weeks.

These techniques are equally applicable either for building up a profile of potential suspects or victim identification. Investigators can use them to gain a better understanding of the digital personas involved, and their use also potentially provides clues to the identity of an individual or group in the physical world.
The Isis toolkit can compare the metric scores produced for two or more text collections to indicate how likely it is that the sources of the texts overlap or are written by people of a similar age and gender.

1) Stylistic language fingerprint: The Isis toolkit can observe and scrutinize a wide range of subtle language

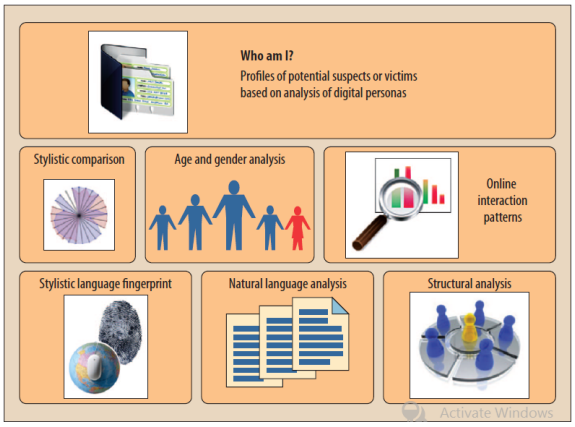


Fig. 2. The Isis toolkit.

traits to assist in authorship analysis. Examples include the proportions of punctuation characters, the use of emoticons, and vocabulary measures. The toolkit uses these language traits to build a stylistic fingerprint that it can, in turn, use to represent the language of a particular user, a set of users, or a collection of texts. Metrics used to construct the stylistic fingerprint range from simple counts, such as the number of exclamation marks present, to more complicated measures, such as the type token ratio, a vocabulary indicator. The calculation of each metric takes into account text length so that a mixture of sources can be combined where appropriate for example, email texts are generally longer than chat room texts.

Through this process, Isis can assign a list of metric scores to a single text or collection of texts. Examples include the collated messages from a single chat room user or a sample of texts chosen to represent the language of adult female chat room users. The toolkit can then compare the metric scores produced for two or more text collections to indicate how likely it is that the sources of the texts overlap or are written by people of a similar age and gender.

The Isis toolkit uses the metric scores in two ways:

- To assist with automatic age and gender analysis.
- To providea visual impression of how close two text sources are with regard to their linguistic style.

2) Age and gender analysis: Isis performs this analysis in four steps. The first three steps utilize a natural-language-analysis engine, while the fourth combines the knowledge thus extracted with the metric scores from the stylistic fingerprint:

- Step 1: Tokenize an incoming text sample and tag each word with a part-of-speech (POS) labelnoun, verb, adverb, adjective, and so on.
- Step 2: Assign each word or phrase within the text to one semantic field using general conceptual labels such as finance, warfare, government, sports, and so on. These first two steps rely on a set of hybrid techniques to select the most likely tag in each context.
- Step 3: Count features such as the language styles used at the word, POS, and semantic field levels.
- Step 4: Compare each level to standard reference datasets that have previously been processed through the same pipeline.

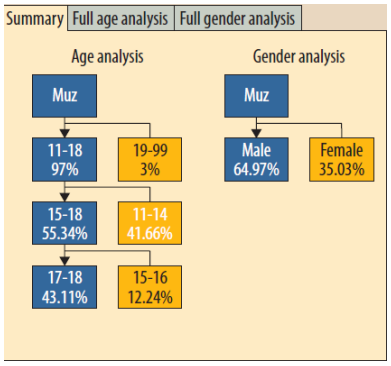


Fig. 3. Age and gender decision trees.

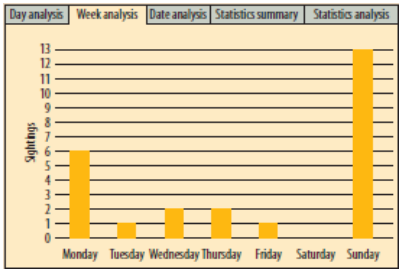


Fig. 4. online-offline time analysis.

In the case of gender, we prepare two reference datasets, one for males and one for females. A distance metric then calculates the similarity between the incoming text sample and each of the two reference corpora for each of the three levels. Isis produces metric scores from the stylistic fingerprint for each reference dataset such as different gender groups and uses them as features for training a text classifier. Various machine learning algorithms and methods for feature extraction are used for a range of text classification purposes. Isis uses logistic regression with the metric scores as features to classify a given text into gender groups. Probabilities are produced that indicate the likelihood that the given text should be classified as each gender group. These are then combined with the word, POS, and semantic field analyses to derive weighted combined scores. The system then assigns a value for how likely it is that the incoming text is written by a male or female author. Similarly, Isis can prepare reference datasets by age range and compare them in the same manner. It is possible to focus on

smaller age ranges by preparing specific reference datasets. This allows the toolkit to present an overview of the likelihood that a text is written by an adult or a child, and then drill down to results for more precise age ranges. As Figure 2 shows, the Isis toolkit provides this information as a decision tree that a law enforcement officer can consult and interact with.

3) Comparing stylistic fingerprints: While the automatic prediction of age and gender is useful in many cases, visualizing language differences and similarities also can be helpful to an investigator. The metric scores offer the ability to plot stylistic differences on a graph. While more than two lists can be compared on the same plot, here we discuss only the comparison of two lists. Given two lists of metric scores, each

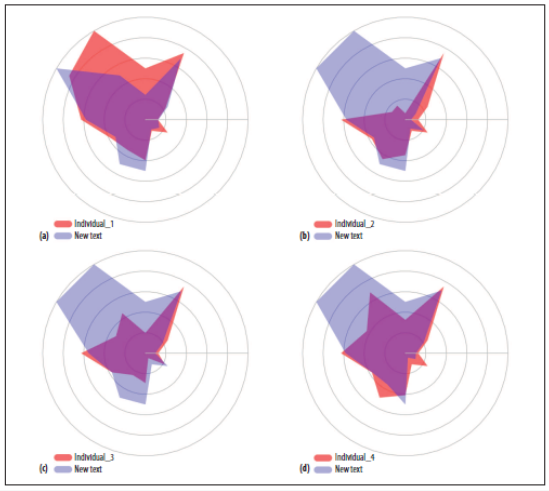


Fig. 5. Comparison of language style fingerprints for a new text against four individuals fingerprints..

score is divided by the maximum of the two scores for that metric. Hence, one adjusted score for each metric is now 1, and the other is a fraction of that (between 0 and 1). The adjusted scores are then multiplied by metric weights derived through machine learning, which can be specialized for the text comparisons being performedfor example, comparing a users text against age group datasets. Radar plots of the adjusted and weighted metric scores can then be used to visually represent language style fingerprints. When the two plots are overlaid, the similarity or difference between the two text sources represented is evident, with substantial overlap indicating that the language style is similar and little overlap indicating contrasting language styles. In addition to displaying how close a users text is to a given age and gender dataset, the fingerprinting method also can be used to compare the text from two personas to establish whether they are actually the same individual, or to compare texts from one persona at different times to explore whether multiple individuals share the persona. To describe this process and demonstrate the fingerprint comparison technique, Figure 3 shows the language style fingerprint comparison of a previously unseen text (the messages of a single user in one chat session) against the collated texts of four individuals (for each individual, the messages are taken from six chat sessions). A larger overlap (shaded purple in the figure) of the fingerprints indicates that the new texts language style is similar to that of previous texts for an individual, hence the new text is more likely to be from that user. In Figure 3, the overlap is most marked for Individual 1 (top left), so a judgment could be made that the new collection of chat room messages is likely to be from that user. In this case, that judgment would be correct: the fingerprints are from real chat sessions conducted in a simulated cybercrime scenario.

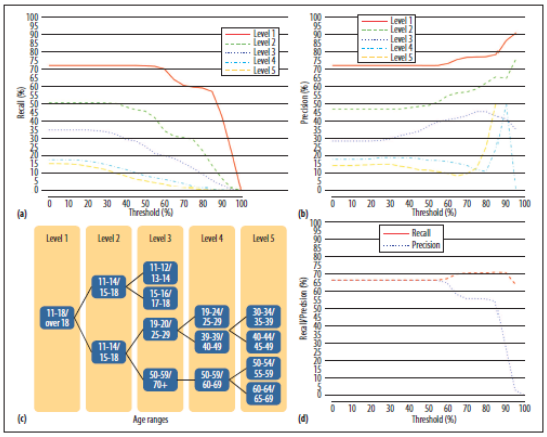


Fig. 6. Age and gender classification: (a) recall and (b) precision for age classification at (c) different specificity levels, and (d) gender classification.l

4) Online interaction patterns: The Isis toolkit also supports identification of patterns typical to a personas online presence and its interaction with other participants. This is achieved through structural analysis of the text, which extracts details such as the usernames of those participating in the chat or date and time information that can be used to model the conversation flow to identify patterns and trends over time. All conversation logs entered into the toolkit are converted into a generic format. A key aspect of this is breaking down and modeling the log in terms of the participants and their respective activity, such as posting messages, sharing links, leaving or joining the conversation, and so on. Once it has built this model, the toolkit can quickly analyze and present intelligence about a particular participant. This can include an analysis of language usefor example, frequently used stylistic characteristics such as keywords, names, topics, and so onor identifying patterns of online and offline times. Semantic categorization allows classifying parts of a conversation based on their meaningfor example, whether it is sexual or aggressive in nature. By applying these techniques to the model of a participants conversation, it is possible to view any trends that might occur over the duration of the conversation, for example, to help determine whether a conversation is becoming increasingly sexualized. As Figure 2b shows, the analysis of

online-offline windows becomes particularly relevant in online social media where many participants are active. By cross-referencing different participant models, the toolkit can show when participants are online together as well as the content of their conversation at those times. This information can be used to make inferences about who tends to communicate with whom and about what. It also can help determine if a suspect is switching between multiple user accounts a trend that is frequently seen when online personas are exploited for criminal purposes.

VI. Profiling cybercriminals and victims

The various analysis techniques combine to form a key feature of the toolkitthe ability to generate identity profiles of specific digital personas. Isis can automatically create profiles for a specified digital persona, drawing upon the conversations in which it has participated to produce an overall analysis of its online activity, language, and identity characteristics. The generated profile is built from several elements, including:

- Language usage:
  This element provides a model of the personas language use within conversations and highlights characteristics such as people or place names, dates and times, frequently used words and phrases, aggressive or sexual content, or email addresses and URLs. It also includes nondictionary words about which an investigator might or might not be aware that could indicate an attempt at disguising what is being discussed or represent unique jargon used within that domain.
- Age and gender analysis:
  Investigators can use a decision tree to provide an inferred estimation of the age and gender of the person behind the persona. By default, this provides a summary view that presents the strongest path through the tree, but they also can view the full tree, allowing them to examine the decisions the toolkit made at all points if the certainty of the decision is not clear-cut.
- Online activity:
  An analysis of overall online activity can highlight when the persona has appeared online within relevant conversations. This analysis can take many forms, including indicating when a persona is most likely to be online over a 24-hour window and on which days during the week. These profiles can provide investigators with additional intelligence about trends and characteristics not immediately apparent to the human eye.

A. Differentiating Between Genuine Personas And Deceptive Behaviour

We have used the Isis toolkit on reference datasets and in live environments to test its effectiveness in correctly detecting the attributes of an individual behind a persona. In the two test sets presented here, no deception is intended in the first, while in the second, an individual is using deceptive tactics[21].

1) Classifying age and gender of genuine personas: For this test we used the British National Corpus (BNC), a reference dataset with 100 million words of written and spoken language that represents a wide cross-section of British English. We utilized the portion of BNC (1,684 people, which constitutes 10 percent of the entire collection) where metadata about an individual, including age and gender, was available. We used leave-one-out crossvalidation to train our system using the texts from all 1,684 individuals except the text from the individual being used as a test subjectthe person whose age and gender was being classified. We repeated this classification for all 1,684 individuals as test subjects. For each classification, the Isis toolkit provides probabilities that the individual belongs to a specific age band; for example, an individual might be predicted to be age 11 to 18 with a probability of 74 percent, and over age 18 with a probability of 26 percent. The prediction then moves down a level, that is, to between ages 11 to 14 with a probability of 49 percent, and between ages 15 to 18 with a probability of 25 percent, and so on, with gender probabilities also calculated[22]. As the age and gender classification decision trees in Figure 2a show, the age group or gender with the highest probability is taken as the prediction at each decision point. A probability threshold is also used to decide whether a prediction is used. If the highest probability is below this threshold, no prediction is made, and the age or gender is marked as unknown. Recall and precision are used to measure the algorithms ability to correctly classify the age and gender of each individual in the test set. Recall is the proportion of individuals tested for which the correct prediction is made; precision is the proportion of predictions madethat is, not unknownthat are correct according to the metadata. Figures 4a and 4b give the recall and precision values obtained for age classification at different specificity levels, which are outlined in Figure 4c. Recall is 72.15 percent, and precision is 72.24 percent at Level 1, that is, distinguishing between children (ages 1118) and adults (over age 18). This is based on a probability threshold of 50 percent. By increasing the threshold, greater precision can be achieved at the cost of fewer classification decisions being madethat is, more unknowns are returned. With a higher threshold of 80 percent, the precision of adult and child classification increases slightly to 77.35 percent, but recall drops to 59.20 percent. Naturally, the precision and recall drop at higher levels of the age decision tree where the age ranges are more specific. As Figure 4d shows, for gender classification with a threshold of 50 percent, recall is 66.74 percent, and precision is 66.86 percent. Again, a higher threshold can be set; increasing the threshold to 80 percent improves precision to 71.07 percent, but recall drops to 56.08 percent.

Our analysis performs better when deception tactics are being used, thus demonstrating the effectiveness of digital persona analysis as a valuable tool in the investigators workbench.

2) Classifying deceptive personas: We tested our toolkit on the BNC data to determine the accuracy of our algorithms when individuals are not being deceptive. However, given our

focus on detecting misuse of digital personas, we tested the toolkit on detecting masquerading behaviorwhen an individual hides behind a false persona, for example, pretending to be a child. We set up a live environment, similar to a Turing test, in two schools. Subjects ages 11 to 18 years chatted online with 10 individuals behind the scenes in sessions divided by age group. In each session, one-half of the individuals behind the scenes were children or young people of the same age as the chat participant, while the other half were masquerading behind personas purporting to be of that age. We then employed a similar evaluation process as for the BNC dataset to test the effectiveness of our toolkit in classifying whether the people behind the scenes were children or young people or masquerading as being in those age groups[21].
For deciding whether an individual is a child or an adult masquerading as a child, the age classification algorithm achieves precision and recall of 84.29 percent, with a probability threshold of 50 percent. The precision can be increased with a higher threshold; at an 80 percent threshold, precision increases to 93.18 percent, but recall drops to 58.57 percent, with fewer predictions being made. These results obtained using our toolkit are in stark contrast to the accuracy of the childrens responses, with only 18 percent of children across the year groups able to correctly identify whether they were chatting with an adult or a child. For gender, precision is 80.6 percent with a 50 percent threshold, while recall is 77.14 percent. Increasing the threshold to 80 percent improves precision to 84.09 percent, but recall drops to 52.86 percent. These results are in contrast to the children correctly identifying the gender of the person with whom they were chatting in 58.8 percent of the cases.

B. Challenges Tackled by the Isis Approach

Existing work on policing online social networks has focused primarily on the monitoring of chat and file sharing systems. Chat policing software for home use such as Spector Pro2, Crisp3 and SpyAgent4 allow the logging of online conversations, but are restricted in that they need to be installed on the actual PC that is partaking in the activity. Less obtrusive chat policing systems, as used by policing organisations, typically use a network-level tracing methodology [15] to identify and log chat traffic at the network-level for later analysis. However, police surveillance tactics deployed at network-level present real challenges to law enforcement in terms of detecting edge-based criminal activity and achieving effective online guardianship [14]. Three significant shortcomings can be observed. Firstly, too much data is produced to make pro-active analysis practical. Secondly, child sex offenders often masquerade as children in order to make contact, making detection difficult. Thirdly, systems tend to be developed to monitor a predator stereotype (adult male) which does not reflect patterns of internet based sexual predation of children and young people [7]. For example, Finkelhor [8] found that young people themselves make aggressive sexual solicitations in almost half of all cases and that of those known to be adults (25%), the majority are aged between 18-25. In 27% of cases in this study (conducted in the US) the age of predators was unknown and could well include adults masquerading as young people. A key question yet to be addressed is how to distinguish both between adults predating as young people and between normal youth sexual behaviour on the internet and youth predation. Due to these challenges, policing organisations focus primarily upon reactive policing, wherein known culprits are identified and tracked and children are provided with mechanisms to report suspicious behaviour. Unfortunately, this approach is incapable of tackling many cases where children do not report incidents (in [18] only 3unknown to the authorities. Moreover, these policing tactics do little to advance a preventive approach to the problem of online grooming and predation within social networks by enabling effective guardianship and the potential for law enforcement intervention in pre-criminal situations, e.g., at the point of an early friendly online encounter between a prospective offender and a child the significance of offender search, precriminal situations, opportunity and other contextual factors in the prevention of Internet crimes against children)[26].
In terms of language monitoring capabilities the existing chat policing software tools rely on human monitoring of logs or simple-minded keyword or phrase detection based on userdefined lists. Such techniques do not scale. Nor do they enable identification of adults masquerading as children or support pro-active policing. Techniques do exist which make use of statistical methods from computational linguistics and corpus-based natural language processing to explore differences in language vocabulary and style related to age of the speaker or writer. The existing methodologies, such as key word profiling [19], draw on large bodies of naturally occurring language data known as corpora (sing. corpus). These techniques already have high accuracy and are robust across a number of domains (topics) and registers (spoken and written language) but have not been applied until now to uncover deliberate deception. The second relevant set of techniques is that of authorship attribution. The current methods [16] would allow a narrowing in focus from the text to the individual writer in order to generate a stylistic fingerprint for authors.
For policing file sharing systems two significant tools exist, Peer Precision5 and LogP2P6. Both systems also use a network-level tracing methodology in conjunction with a honey-pot approach, wherein the policing peer offers an illegal file to the network and when an offender attempts to download this file, client-side software captures the offenders IP address at the packet-level. This approach suffers from two significant shortcomings. Firstly, it is unable to differentiate between those who download and share a single file, and those who are the core distributors of child abuse media (e.g. distributing many thousands of files, producing and distributing child abuse imagery or uploading newly-produced child abuse material for the first time). This is a significant problem for frequently backlogged child protection agencies

with limited resources. Secondly, as these systems work at the network-level, they can potentially be thwarted by encryption at the application-level. This is of particular significance as recent research has shown that users are migrating to more anonymous and secure file sharing systems [13]. Finally, and perhaps most critically, the honey-pot approachrelies upon the use of well-known files. Hence, it is incapable of identifying those offenders

who may be sharing recently-produced material. The incorporation of monitoring functionality in file sharing systems requires significantly altering multiple components to ensure that monitoring takes place at the right points in the system. However, such invasive changes are expensive and hard to maintain and evolve across various releases of a system. The recent rise of aspect-oriented software development techniques [4] has facilitated noninvasive composition of such systemic concerns as monitoring, which makes in-step evolution of such functionality with changes in the rest of the system more modular and manageable. Though aspect-oriented techniques have been used in individual systems (e.g., the widely used mysql database system) for logging purposes, to date, they have neither been applied for monitoring online social networks nor on a scale spanning multiple systems and various releases of such systems. A particular issue in file sharing systems is that filenames reflect specialised vocabulary which changes over time [17].

Taking an ethical perspective on this research is not only important in respect of the abuse of children [12], and the protection of the researchers who conduct this type of research, but also because of the issues surrounding the use of monitoring technologies that have an impact on user privacy [30]. Researchers in the field of computer ethics have noted that values are embedded within technology design, e.g., [24,25], and as a result, there have been numerous calls for the integration of ethical assessment, evaluation and stakeholder impact analysis within the design and development of computer systems to mitigate adverse effects, [26,27]. In advocating this approach there is a recognition that not only are the potential risks associated with the software development reduced [18], but also that the awareness of the development team to the ethical aspects inherent in these systems is raised thus creating a body of ethically aware information professionals [6]. To date there has been a lack of suitable case studies in the computer ethics literature and appropriate guidance for technology developers to incorporate ethical considerations within the development cycle. This project is developing new understandings of user-centred methods for highly sensitive systems and of effective designs of privacy/awareness interfaces that will benefit other developments and mitigate the effects of adverse outcomes that impact on public acceptability.

VII. Conclusion

Online social media affords connectedness that enables individuals and groups from various geographical, cultural, and socioeconomic backgrounds to interact and share experiences. However, the very nature of identity in online social mediaa uid and dynamic concept that can be created, adapted, and discarded with easemakes such identities prone to misuse. Exploitation of digital personas has be- come an integral part of the tactics that cybercriminals use. This new digital world and these sophisticated criminal tactics call for new tools to aid investigators of online crime. My experience with the Isis toolkit demonstrates that it is possible to detect key characteristics of individuals or groups behind digital personas with a high degree of accuracy by combining techniques from corpus-based natural language analysis with those from authorship attribution. In fact, our analysis performs better when deception tactics are being used, thus demonstrating the effectiveness of digital persona anal- ysis as a valuable tool in the investigators workbench. Naturally, such linguistic analysis cannot provide 100 percent accuracy because of the intricacies of human language and its use. In addition, My experience in ongoing trials of the toolkit in UK law enforcement agencies shows that expert investigator knowledge is indispensable to the investigative process. The toolkit is, therefore, intended as a means to support the work of investigators rather than offering full automation. Only by combining such sophisticated tools with the expert knowledge of investigators can we hope to understand and nullify the online tactics that criminals deploy.

VIII. Dedication

Bestowed to the memory of children who impaired their dreams in the lurch of social networking media.

Fig. 7. Dedication

References

[1] A. Rashid et al., *Technological Solutions to Offending*, Willan CyberPsychology, Behavior, and Social Networking, vol. 10, no. 1, pp. 228-243, 2012.

[2] M.T. Whitty and T. Buchanan, *The Online Dating Romance Scam: A Serious Crime*, CyberPsychology, Behavior, and Social Networking, vol. 15, no. 3, pp. 181-183,2009.

[3] G. Weimann and K. von Knop, *Applying the Notion of Noise to Countering Online Terrorism*, Studies in Conflict and Terrorism, vol. 31, no. 10, pp. 883-902, 2008.

[4] H. Davulcu et al., *Analyzing Sentiment Markers Describing Radical and Counter-Radical Elements in Online News*, Proc. 2nd Int'l Conf. Privacy, Security, Risk and Trust (PASSAT 10), IEEE, pp. 335-340, 2010.

[5] J. Diesner and K. Carley, *Using Network Text Analysis to Detect the Organizational Structure of Covert Networks*, Proc. Conf. Computational Analysis of Social and Organizational Systems (CASOS 04), Nat'l Assoc. Computational Social and Organizational Science, 2004.

[6] H. Bisgin et al., *A Study of Homophily on Social Media*, World Wide Web, vol. 15, no. 2, pp. 213-232, 2012.
[7] P. Greenwood et al., *Udesignit: Towards Social Media for Community-Driven Design*, Proc. Int'l Conf. Software Engineering (ICSE 12), IEEE, pp. 1321-1324, 2012.
[8] S. Prentice et al., *The Language of Islamic Extremism: Towards an Automated Identification of Beliefs, Motivations and Justifications*, Int'l J. Corpus Linguistics, vol. 17, no. 2, pp. 259-286, 2012.
[9] E. Stamatatos, *A Survey of Modern Authorship Attribution Methods*, J. Am. Soc. Information Science and Technology, vol. 60, no. 3, pp. 538-556, 2009.
[10] S. Afroz et al., *Detecting Hoaxes, Frauds, and Deception in Writing Style Online*, Proc. IEEE Symp. Security and Privacy (S&P 12), IEEE, 2012, pp. 300-314. Privacy (S&P 12), IEEE, pp. 461-475,2012.
[11] P. Rayson, *From Key Words to Key Semantic Domains*, Int'l J. Corpus Linguistics, vol. 13, no. 4, 2008, pp. 519-549.
[12] H. Davulcu et al., *Analyzing Sentiment Markers Describing Radical and Counter-Radical Elements in Online News*, Proc. 2nd Int'l Conf. Privacy, Security, Risk and Trust (PASSAT 10), IEEE, pp. 335-340,2010.
[13] T. Byron, Safer children in a digital world: the report of the Byron review, http://www.dcsf.gov.uk/byronreview/, 8:40:59Pm,Jan 24,2015.
[14] J. Palfrey, Enhancing child safety & online technologies: final report of the Internet safety technical task force, Berkman Center, Havard University, 2008.
[15] D. Hughes, S. Gibson, J. Walkerdine, G. Coulson, Is deviant behaviour the norm on P2P file sharing networks? IEEE Distributed Systems Online 7(2), 2006.
[16] R. Filman, T. Elrad, S. Clarke, M. Aksit (eds.), Aspect-Oriented Software Development, Addison-Wesley, 2001.
[17] D. Hughes, J. Walkerdine, K. Lee, Monitoring challenges and approaches for P2P file sharing systems, Proc. 1st International Conference on Internet Surveillance and Protection (ICISP06), 2006.
[18] M. Taylor, E. Quayle. The Internet and Abuse Images of Children: Search, Pre-criminal Situations and Opportunity Situational Prevention of Child Sexual Abuse, R. Wortley, S. Smallbone (eds.), Criminal Justice Press, pp. 169-195,2006.
[19] S. Dombrowski, K. Gischlar, T. Durst, Safeguarding young people from cyber pornography and cyber sexual predation: a major dilemma of the Internet. Child Abuse Review 16, pp. 153-170, 2007.
[20] D. Finkelhor, K. Mitchell, J. Wolak, Online Victimization: A Report on the Nations Youth, National Center for Missing and Exploited Children, Alexandria, VA, 2000.
[21] P. Rayson (2008). From key words to key semantic domains. International Journal of Corpus Linguistics. 13:4 pp. 519-549,2008.
[22] P. Juola, J. Sofko, P. Brennan, A prototype for authorship attribution studies, Literary and Linguistic Computing 21, pp. 169-178, 2007.
[23] D. Hughes, P. Rayson, J. Walkerdine, K. Lee, P. Greenwood, A. Rashid, C. May-Chahal, C., M. Brennan (2008) Supporting law enforcement in digital communities through natural language analysis. In proceedings of the 2nd International Workshop on Computational Forensics (IWCF 2008), Washington DC, USA, August 7-8, 2008. Lecture Notes in Computer Science 5158, pp. 122-134,2008.
[24] M. Eneman, The new face of child pornography, in Human Rights in the Digital Age, Cavendish Publishing, 2005.
[25] D. J. Cook, S. K. Das, How smart are our environments? An updated look at the state of the art, Pervasive and Mobile Computing 3(2), pp.53-73, 2007. [14] H. Nissenbaum, Values in the design of computer systems, Computers and Society, March, 1998.
[26] J. van den Hoven, ICT and value sensitive design, in The Information Society: Innovation, Legitimacy, Ethics and Democracy, P. Duquenoy, P. Goujon, K. Kimppa, S. Lavelle (eds.), Springer, 2007.
[27] P. Duquenoy, O. Burmeister. Exploring ethical aspects of Pervasive Computing in Risk Assessment and Management in Pervasive Computing: Operational, Legal, Ethical and Financial Perspectives, Varuna Godara (Ed.), IGI Global,pp.264-284, 2008.
[28] D. H. Gleason, A software development solution, Proc. Ethicomp: Systems of the Information Society, Poland, 2001.
[29] D. Gotterbarn, Reducing software failures: Addressing the ethical risks of the software development lifecycle, Australian Journal of Information Systems, 1999.
[30] P. Duquenoy, D. Whitehouse, A 21st century ethical debate: Pursuing perspectives on Ambient Intelligence, Proc. Landscapes of ICT and Social Accountability, Finland,2004