Internet Safety Technical Task Force
Technology Submission Template

Chatsafe

David Crystal & Ian Saunders
Crystal Reference Systems

Keywords
Child safety, chatrooms, Chatsafe, protection, paedophiles

Abstract
A technique for identifying ongoing paedophile activity in chatrooms is described. Known as Chatsafe, it makes a linguistic analysis of the vocabulary used in the conversational turns between adult and child, and employs a cumulative index to identify the point at which a conversation is becoming dangerous.

Functional Goals
Please indicate the functional goals of the submitted technology by checking the relevant box(es):
☒ Limit harmful contact between adults and minors
☐ Limit harmful contact between minors
☐ Limit/prevent minors from accessing inappropriate content on the Internet
☐ Limit/prevent minors from creating inappropriate content on the Internet
☐ Limit the availability of illegal content on the Internet
☐ Prevent minors from accessing particular sites without parental consent
☐ Prevent harassment, unwanted solicitation, and bullying of minors on the Internet
☐ Other – please specify

PROBLEM INTRODUCTION

Chatsafe is the first of several possible applications to the field of public safety of a 'sense engine' general strategy, referred to as the Global Data Model (GDM, patented UK & USA). The core of the strategy is a broad-based system of knowledge categories, each of which has been operationalised with reference to clusters of keywords. The breadth of the knowledge categories stems from their origin in several encyclopedia projects of the 1990s, notably *The Cambridge Encyclopedia*. The keyword clusters were derived using lexicographical methods and avoided the oversimple statistical algorithms (e.g. looking for 'most frequent words') which are widely employed elsewhere.

Insofar as individuals who pose a threat to public safety have to use language in order to coordinate their activities, the texts they communicate can be analysed using the same methods that we have used in identifying and discriminating other knowledge domains. In most instances individuals use their normal language, not realising that linguistic techniques are available to identify their subject-matter and intentions. This is evidently the case in the context of paedophile activity, where the whole point is to communicate as normally as possible; but it would also apply to many cases of terrorist communication, as well as various other kinds of 'plotting' activity where a large group of people is involved.

PROPOSED SOLUTION

There are two distinct aspects to the analysis of paedophile (P) data: monitoring the incoming messages from P and advising the target (T) how to respond to them. Only the monitoring function is illustrated below.

The aim is to distinguish an innocent from a dangerous conversation, on the basis of the loaded words (LW) they contain. There are three types of innocent conversation, in this respect:

- **Type 1** Conversations which use virtually no LW.
- **Type 2** Conversations which happen to use some LW at the outset, because of the subject-matter, and then quickly and steadily decline.
- **Type 3** Conversations which stay low but have the occasional peak, as the subject-matter changes.

A P conversation, by contrast, will have various characteristics:

- The 'grooming' approach is slow at the outset, so we will expect the number of LW to take time to build up. It is the accumulation of LW over time, conversational turn (CT) by CT, which is critical.
- P will 'test the water' at intervals by using LW, and there will therefore be a pattern of peaks and valleys in the sequence of CT scores.
- As the suggestions become more focused, some individual CTs will achieve very high LW scores. If Chatsafe is used, this stage will not be reached.

We need a system which will identify danger early on, but not so early as to bring up incidental high-scoring usage in innocent conversations. We need to avoid misassigning

cases where there is a high score at the outset because of some chance subject-matter.

Method

A total of 217 keywords and phrases (363, if we include variant forms, e.g. *picture*, *pictures, pic, pics*) were scored on a scale from 1 to 5 in terms of increasing lexical sensitivity. For example:

- **Level 1** words: age, friend, girl, learn, school, thinking
- **Level 2** words: enjoy, legs, mouth, put on, size, watch
- **Level 3** words: alone, cam, explore, on your own, shows, thoughts, trust
- **Level 4** words: bare, bedroom, kinky, photo, picture, tell me, what + like
- **Level 5** words: breasts, meeting, naked, porn, punish, sex, strip, underwear, what + wearing

The full list of terms can be made available to potential partners or clients, under the usual confidentiality terms.

For each conversational sample, a Cumulative Paedophlle Index (CPI) is calculated, CT by CT: the LW score in the second conversational turn CT2 is added to that in CT1, CT3 to that in CT2, and so on. For example, in the following sequence of turns :

if CT1 scores 0	cumulative LW score is	0
if CT2 scores 0		0
if CT3 scores 2		2
if CT4 scores 2		4
if CT5 scores 3		7
etc		

The CPI is obtained by dividing the cumulative LW score by the number of the CT and multiplying by 100. Thus, in the above example:

at CT1 the CPI is 0	(0/1 x 100 = 0)	
at CT2 the CPI is 0	(0/1 x 100 = 0)	
at CT3 the CPI is 66	(2/3 x 100 = 66)	
at CT4 the CPI is 100	(4/4 x 100 = 100)	
at CT5 the CPI is 140	(7/5 x 100 = 140)	etc

A sensitivity level has been set at 100. An innocent conversation will routinely score well below 100. A P conversation, once it has 'taken off', will only rarely dip below 100. The analysis can take place in real time or using a log of a conversation.

In an inital test of one P/T conversation, placed on the web by a child protection agency for illustrative purposes, the technique performed exceptionally well. Comparisons of the paedophile conversation were made with a number of innocent conversations, one of which is shown in the Appendix. However, legal restrictions have so far made it impossible to obtain more real conversations in order to carry out further tests, and we need a partner to take this forward - hence the present submission.

EXPERTISE

Chatsafe was developed by Professor David Crystal, Honorary Professor of Linguistics at the University of Bangor, UK, a world authority of linguistics and the author of over 100 books on the subject of language, including several relating to child language and discourse interaction. The technical development team is led by Dan Wade, a noted developer of information retrieval products and the architect of the software behind the Crystal Semantics technical implementation.

COMPANY OVERVIEW

Crystal Reference Systems is a wholly owned subsidiary of ad pepper media International N.V, a leading online advertising company, with 268 employees in 19 offices in 16 countries. The company is quoted on the German Dax stock exchange. Crystal Reference is a research and development division of ad pepper, responsible for developing semantic solutions for targeting and optimization of advertising, as well as solutions dealing with online safety and security.

BUSINESS MODEL OVERVIEW

We have a flexible range of pricing models varying from a simple per use model to a full technology licensing model. This will make the product accessible to all ranges of user types.

MORE INFORMATION

http://www.isense.net

http://www.davidcrystal.com

CONTACT INFORMATION

Ian Saunders

Managing Director, Crystal Reference Systems

26 Stanley Street, Holyhead, LL65 1PB, UK

T: +44 1407 761550 F: +44 1407 769830

CERTIFICATION

I certify that I have read and agree to the terms of the Internet Safety Technical Task Force Intellectual Property Policy

Ian Saunders

APPENDIX

As an illustration, the table below shows the CPIs for the first 90 conversational turns (CTs) from samples of two conversations, using data taken from available Web sites. The first (A) is a group of nine young Buffy the Vampire Slayer addicts; the second (B) is a paedophile-child interaction.

The contrast is very clear. A's score stays very flat. M's score passes 100 on the third CT and stays there, in a peak+valley pattern. By his 400th CT (not illustrated here), M's score passes 1000 and continues to rise.

CTs	***5***	***10***	***15***	***20***	***25***	***30***
	A					
LW score	0	0	0	0	0	0
CPI	*0*	*0*	*0*	*0*	*0*	*0*
	B					
LW score	0	5	18	36	39	44
CPI	*0*	*50*	*120*	*180*	*156*	*147*

CTs	***35***	***40***	***45***	***50***	***55***	***60***
	A					
LW score	*0*	*0*	*3*	*5*	*5*	*5*
CPI	*0*	*0*	*6*	*10*	*9*	*8*
	B					
LW score	*54*	*56*	*61*	*89*	*97*	*97*
CPI	*154*	*140*	*135*	*178*	*176*	*162*

CTs	***65***	***70***	***75***	***80***	***85***	***90***
	A					
LW score	5	5	6	6	6	9
CPI	*8*	*7*	*8*	*7*	*7*	*10*
	B					
LW score	100	109	109	109	113	118
CPI	*154*	*156*	*145*	*136*	*133*	*131*