

Intellectual property escaped with the email? Press F1 for help

Lee Gillam¹, Neil Cooke²

¹University of Surrey, Computer Science Department,
Guildford
l.gillam@surrey.ac.uk

²University of Surrey, ILab,
Guildford
n.cooke@surrey.ac.uk

Abstract: In this paper we describe an approach to information assurance in which we can prevent breach of confidentiality. Specifically, we examine aspects of the propagation of confidential information via email. Email provides one simple mechanism by which naive users can rapidly breach confidence, and effectively demonstrates the difficulty in enforcing “best practice”. We propose and demonstrate an intelligent filtering system aimed at prevention of disclosure and enforcement of “best practice”. We show how work in corpus linguistics and data mining can contribute significantly to information assurance and report on experiments on a relatively large email collection that demonstrates the value of such work. Our experiments further demonstrate the potential for reducing false positives. We give due consideration to the danger of missed messages that should have been prevented from propagation.

Keywords: Email, confidentiality, linguistics, data mining, Enron Corpus.

1. Introduction

The Oxford English Dictionary describes Intellectual Property as a “general name for property (such as patents, trademarks, and copyright material) which is the product of invention or creativity, and which does not exist in a tangible, physical form”. Legal protection for intellectual property or the expression thereof is in the form of copyright, designs, patents and trademarks. These variously protect literature, music, films, the visual appearance of a product, technical and functional aspects, and signs associated to products, goods and services. Further, lesser-known, forms of IP also exist, and receive protection in some form, including plant varieties. The key to this form of protection is the existence of a trace of the IP in documentary form: the copyrighted article or the patent application.

Early knowledge management literature [1] [2] [3] focused on knowledge as processes, on the ability to convert between “tacit” and “explicit” forms of the known, on storing knowledge within, corporate databases, and on extracting knowledge from, corporate databases. Policies, processes, and indeed software, played various supporting roles in allowing the propagation of “knowledge” around an organization. The intellectual property, perhaps knowledge

assets, of an organization could, if such claims were to be believed, be captured and transformed to the benefit of the business.

Knowledge management variously considered client lists, customer relationships, business processes and trade secrets. The law of confidentiality applies to ensuring that these kinds of information remain known only to the organization, and are not disclosed to others in ways that would cause harm to the organization. Breach of confidence tends to make headlines when a disaffected employee, or ex-employee, discloses such corporate property to the public at large or to competitor organizations: a recent example of this occurred in Formula 1 racing.

In this paper, we consider the potential for breaches of confidence to occur rapidly on a large scale, and the difficulty of preventing disclosures of, in some cases, corporate intellectual property, by employees using email systems. If employees can easily distribute the company’s secrets around the world in a few seconds by email, or by other insecure electronic means, all other mechanisms used to secure this information are immediately rendered redundant. Our goal is an intelligent and adaptive filtering system for outgoing emails that prevents disclosure of information deemed confidential or otherwise expected to have limited distribution. Such a system should ideally be able to ensure that outgoing emails are unlikely to contain information that would otherwise be detrimental to the organization, and that corporate policies preventing the personal use of email are being adhered to. Little appears to have been published, outside of corporate pamphlets and legal advice [4], on this subject and available techniques and their accuracy, and we have found no direct consideration of the problem of false positives raised due to confidentiality banners.

We discuss a number of experiments we have undertaken with the University of Surrey’s System Quirk text analysis software (section 2.1) and the Enron email corpus (Section 2.2), a collection of emails originally released into the public domain by the Federal Energy Regulatory Commission that has received significant attention (for example [5]). We explore the use of analytical techniques from the field of

corpus linguistics for reducing the number of false triggers, with due consideration given to the truly harmful false negatives – messages that should be caught but are not.

Effectively, this is a kind of data cleansing and would usually be considered under the rubric of data mining, wherein the removal of outliers and consideration of key variables and their dependencies and correlates are made. For us, “confidentiality” banners are our outliers and the false triggers sources we wish to remove. Our work makes strong use of relative word frequency differences between knowledge domains (“weirdness”) [6] and collocation patterns [7] to identify the signatures of these banners. We further consider the use of extended collocation patterns to identify text “zones”, following [8].

On the basis of our analysis, we propose that a system capable of capturing and preventing harmful disclosures would best be integrated with email clients to prevent propagation to the email distribution system in the first place. However, we are aware of the risk that this poses: such a system potentially provides an immediate back-door to specific knowledge, or perhaps intelligence, held elsewhere in the organization that the email user would not normally be privileged to.

2. Background

Email filters are normally concerned with ensuring that emails are free from viruses, worms and other forms of system attacks, and with preventing the acceptance or propagation of spam and latterly of phishing attacks. Secure transmission of emails to trusted sites using both encryption and all of the above filters has also been discussed, and even patented¹. The ready accessibility of key word based spam filtering systems means that companies are implementing them at the same time that spammers are using them to create emails that successfully pass the filters. To do this, spammers use surprisingly simple techniques such as variations of words that include misspellings, incorporation of “foreign” characters, and number substitution, where the results remain generally readable, e.g. Vieagra, Viāgra, V1agra. Key word based approaches to spam filtering are also defeated, by more complex approaches such as the incorporation of text into images [9]. Collaborative filtering [10] where a group of users effectively “vote out” emails as spam by adding these emails to a central database, have proven variously successful against these issues. Such techniques, combined with white-lists and black-lists, Bayesian filtering [11] [12] [13], and a host of other predictive and classificatory techniques, produce varying degrees of successes in prevention of incoming email. One can but marvel at the game-playing approach and the continued inventiveness of the spammers.

For outgoing emails, we are assuming that users are, more often than not, only involved in unintentional disclosure or are naïve in their attitudes and understanding of IT and forensic security capabilities. Arguably, a keyword-based approach should be effective, and there are

many commercial offerings which provide security features for outgoing emails: the majority of these are incoming mail guards used in a different orientation. A simple keyword filtering approach may be helpful on a small scale, however a keyword such as “confidential” will produce a large number of false triggers or false positives since the advent of confidentiality banners. These banners also contain other potential triggers – privileged; attorney; intended recipient – and a “whole-text” approach becomes expensive. Email responses containing a full-quote of the original email, including the banner or perhaps several other banners, serve only to increase the frequency catch and compound the difficulty. The human efforts involved in releasing all such emails captured on the basis of keywords alone can be substantial in large organizations. This is before we consider the potential waste of email archive space due to the profligate use of these banners. To properly assess whether these captured emails contain confidential information, those involved in allowing their release would have to have extensive knowledge of, or access to, all of the confidential material. The logical conclusion would be that an all-knowing group of humans would have to know or have access to all of the knowledge and intelligence within an organization, and to read, understand and allow or deny each and every piece of email traffic - a somewhat expensive, and likely error-prone, process and likely to lead to substantial, if not insurmountable, delays in communication. Computers are much faster at such processing, if the processing engine is well formulated and tested, however packaging up all of the organization’s knowledge and intelligence into a system near the edges of the company firewall may not a desirable approach.

We expect our eventual solution to draw together work in a variety of areas, including but not limited to corpus linguistics, and subtopics of sentiment analysis, text segmentation, text classification, text mining, topic identification and analysis of register variation. Consideration will be made, also, of machine learning algorithms, feature selection and binary classification tasks undertaken elsewhere. We are well-placed, also, for making the all-important considerations regarding systems and security.

2.1 Analytical Software: System Quirk

System Quirk is a package of software for tasks such as text analysis, ontology learning, and terminology and text management. A subset of these applications is freely available at the University of Surrey’s website². System Quirk provides software that implements a variety of analytical techniques from the field of corpus linguistic analysis, from simple frequency counts to keyword-in-context (KWIC) to statistical analyses of distance-based co-occurrence and to contrastive analysis with reference corpora producing so-called “weirdness” values [6]. In this paper, we demonstrate results from the use of a variety of these techniques, validated previously across a range of domains from nanotechnology to automotive engineering to

¹ US Patent Office patent number 6,609,196

financial trading [14] [15]. We augment these techniques with others developed in the course of our work and more specific to the task at hand.

2.2 Dataset: The Enron email corpus

The availability of Business email collections on which to base such analysis is somewhat limited; unsurprisingly given the potential for loss of competitive advantage and individual's privacy. The one widely available collection for such an effort is the Enron Corpus.

The history of Enron and its fall from 7th largest company in the US, a highly regulated financial environment, to and "off balance sheet" losses and bankruptcy in 2001 has been well documented (see, for example, [16]). The Enron story demonstrated, at least, that having a code of ethics was one thing, but abiding by it was clearly another. As part of the investigations into Enron, the Federal Energy Regulatory Commission released a collection of 1.5m emails into the public domain, reportedly so that the public would be able to see the evidence forming part of the investigation. The discrepancy in number of emails is down to certain "data cleansing" activities undertaken elsewhere, including the deletion of messages "as part of a redaction effort due to requests from affected employees". The remaining dataset still demonstrates a large range of the social interactions undertaken using email, including as it does messages within the organization, with other organizations, with friends and family, and sometimes containing material that would be unsuited for lower age groups. It is worth remembering, also, that a number of these employees were not complicit in the fraudulent activities of Enron.

The Enron Corpus contains a reflection of the day-to-day business, and sometimes a trace of the personal activities of the employees, for a large corporate. In its original form, 619,446 emails were reportedly available in folders of 158 users. A database comprising 92% of Enron's staff emails is supposedly available at the Federal Energy Regulatory Commission³ along with a vast array of other documents relating to the investigations into Enron. It is not clear on what basis the 92% is calculated. Other researchers have asked questions about the integrity of the datasets, given the removal of some email account folders and some removal of duplicate records. The Enron Corpus most readily available⁴ [Enron-Raw], comprising 517,431 emails (approx 84% on number of emails), would still appear to be a useful collection for such analysis. The size, number of files per directory, duplications, attachments, odd character codes and rawness of the data within this corpus has caused difficulties for others wishing to perform analysis of this corpus.

The Enron Corpus has been used in related work, much of which has been concerned with data cleansing or classification. These researchers use varying numbers of

emails or produce a new number of emails as a result of some cleansing activities, and subsequently these cleansed versions are used in yet other work. A sample of these studies is provided in Table 2-1, with brief details of the number of emails analysed.

Application	Corpus size (# emails)	Further Description
Automatic classification	12,500	Determining whether emails are for "Business" or "Personal" uses [20], University of Sheffield, UK
Data cleansing, Preliminary analysis	200,399	Analysis of email threads and message distribution. Some folders removed. [17], Carnegie Mellon
Annotation; visualisation	1,700	Manual annotation of email categories. http://bailando.sims.berkeley.edu/enron_email.html
	255,636	University of California, Berkeley Visualisation and clustering. Use of database structure separating bodies, headers and other elements. http://bailando.sims.berkeley.edu/enron_email.html
Automatic classification	20,581	University of California, Berkeley / Automatic approach to building email folders
Data de-duplication	250,485	[18], Massachusetts Amherst MD5 Hashes on body text to identify duplicates, resulting in 250,485 emails ⁵ . http://ciir.cs.umass.edu/~corrada/enron/Massachusetts, Amherst
Social Network Analysis.	Not available	Link Discovery for Counter terrorism & Fraud. http://sgi.nu/enron/use.php?s=usc Southern California
Deception Theory	289,695	[19] Queens, Canada.

Table 2-1: Related work on the Enron Corpus

The approach to manually annotated emails of Jabbari et al [20], building on prior work by Marti Hearst at Berkeley, is interesting for us since the 94% inter-annotator agreement suggests a large degree of differentiation is possible. Of the remainder, the objective of the email, as the authors identify, comes into question: business vs. personal travel; purpose of inter-employee meetings, and so forth. As with much of the work on samples of the corpus, the basis for selection of these emails is not known – hence, the extent to which the sample is representative of the corpus is unclear. Furthermore, the automatic classifier appears to have been run on a smaller sample of 5000 emails, and using what appears to be the two extreme classes identified; this may contribute significantly to high performance figures. This work is interesting for us since we may be able to identify breaches of email policies where personal emails are forbidden, or where policies allow, identifying those unintentionally providing confidential information about themselves to wider audiences – at potential loss only to the sender.

Work on deception theory [19], suggests that those intending to deceive using text as the only medium leave a

² Available at: <http://www.computing.surrey.ac.uk/SystemQ/>

³ <http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp>

⁴ <http://www.cs.cmu.edu/~enron/>

⁵ Large numbers of low entropy responses, for example, "yes" "no" "proceed" "thanks for that" and "see attached" may result in duplicate codes, and such work suggests further investigation is needed.

particular linguistic trace. According to this theory, authors attempt to tell a simpler story and to disassociate from the story. For these researchers, such deception is traceable in text by fewer instances of first-person pronouns, fewer low-frequency words and increased frequency of both verbs and negative sentiments. The latter, which is of some interest for our work, would suggest that sentiment analysis and deception theory overlap – perhaps the analysis of negative movie reviews would assist in determining the overlap? The researchers investigate word frequencies with respect to the British National Corpus (BNC⁶) but do not appear to have constructed a set of expected values for such items in general communication, or within given contexts and especially within email corpora. Without such expected values, it is difficult to know whether claimed results would be robust. One question, for us, would be whether it is even possible to differentiate between lexical cohesion due to repetition in well-formed and focused arguments, or whether the increased frequency of certain words is an indicator of deception.

The General release version “Enron-Raw” we are using was issued March 2, 2004⁷, and consists of 517,431 emails, 150 users with 3349 folders, 209,204,013 tokens and has a vocabulary of 282,595 words. Our work as currently formulated is specifically aimed at providing a general model for avoiding the confidentiality banners – or more generally, corporate disclaimers. We are focused on “confidentiality banners” causing false positives for outgoing email filtering systems, with emphasis on protective markings as used, for example, by UK government departments. The distinctions between business and personal emails may inform, and be informed by, these efforts, and finer-grained deception analysis may help such efforts - or with the removal of these objects provide for a better input set for deception analysis.

2.3 Application Domain: Can you keep a secret

Data security is strategically important for the protection of government and military information and personnel, and is of growing importance in combating identity-based crimes such as fraud and cyber-stalking. It is usually the high profile breaches of data security that become newsworthy: the lost government laptop; the US military secrets on a memory sticks for sale at a bazaar in Afghanistan; the high street banks encouraging customers to shred bank statements while leaving un-shredded account details in rubbish bags, and the fraudsters recovering bank details from PCs sent to Africa for recycling. Of course, those volunteering their personal information to a world-wide audience, or to disreputable companies, may find a range of problems also [21].

Emails have become a primary mode for asynchronous communication in modern business life. In the same way that an organization’s website, allied to effective use of search engines, provides a substantial market presence,

emails can represent the organization in other ways. The benefits of effective use of email systems can be in carrying out and gaining trade through discussion, exchange, learning, contacts, contracts etc. Yet there is also a severe risk of loss of reputation, breach of confidence, loss of intellectual property, and loss of tactical and strategic business information. Aspects of human behaviour also present the risk of loss of reputation: organisations need to maintain corporate professionalism within emails leaving their organisation, and a careless or unguarded reply can rapidly bring embarrassment to individual and business alike. Machine-based monitoring of email communications, in a fair and timely manner, can help to avoid such lapses. And yet the technology to support this vital activity is extremely limited. Such a system should check outgoing emails for:

- sensitive subject areas discussed in the body text,
- the wrong kind of sentiment,
- sensitive attachments,
- inappropriate addressees, (competitors, reporters etc.)
- authors out of context,

and cope with all the vagaries of large numbers of unique sparse emails.

A capable system needs to be developed on the basis of a benchmark data collection. For us, this entails a good, freely available, corpus which preferably contains all the vagaries of human behaviour in a business context. Previous research has been undertaken on email corpora, primarily to detect and remove spam. Spam filtering aims at preventing the receipt of propagated emails. Many small corpora, and related publications, exist for such tasks including:

Email corpus	Number of emails	
	SPAM	Non-SPAM
Spam Assassin ⁸	1897	4150
Synthetic (Annexia/Xpert) Corpus [22] ⁹	10,025	22,813
LingSPAM[23] ¹⁰	481	2412
GenSpam anonymised email/SPAM corpora[23] ¹¹	32332	9072
TREC Spam corpus (2006) [24]	110597	52989
TREC Spam corpus (2005) [25]	113129	205353

Table 2-2: Corpora typically used for the detection of Email spam

Our efforts differ from those involved in spam filtering in that the action required for detection of an offending item requires more granular identification, although spam filtering would provide additional data cleansing, given the estimate of 2.5% of the Enron Corpus comprising spam [26]. For the spam detection task, an email is either allowed or blocked based on scoring mechanisms which usually include some form of naïve keyword filtering. For our task, an approach based on naïve keyword filtering could produce

⁸ [http://spamassassin.apache.org/ and http://spamassassin.apache.org/publiccorpus/]

⁹ [http://www.trudgian.net/spamkann/synthetic_corpus.php]

¹⁰ [http://www.iit.demokritos.gr/skel/i-config/downloads/lingspam_public.tar.gz]

¹¹ [http://www.cl.cam.ac.uk/~bwm23]

⁶ British National Corpus may be found at: <http://www.natcorp.ox.ac.uk/>

⁷ The Enron Corpus may be found at: <http://www.cs.cmu.edu/~enron/>

a torrent of false positives: for example, the simple expedient of including the keyword “confidential” would block all email responses with quoted content containing confidentiality banners. This would necessitate extensive human intervention, probably unnecessary from a technical perspective. This one keyword, amongst many, could be strongly indicative of a potential breach of confidence, so such intervention becomes highly necessary from an organisational perspective. Allied, also, to intentional breaches, the significant potential for human lapses of judgment could result in incorrectly propagated, and hence harmful, messages. We are aiming, therefore, at preventing propagation or somehow intervening at point of production rather than at the point of distribution, i.e. before the email reaches a mail server. It is clear that the context of the keyword is vital in automating judgments.

3. Approach

With any approach to (artificially) intelligent processing, the most important factor is the choice of heuristic: it should represent value for information gain, be easy to implement and make effective use of the information elements. The intention of our present efforts is to construct and implement an algorithm that identifies and discounts confidentiality banners. Our initial efforts, therefore, concern determining whether a pattern of such banners can be learnt. Our approach involves:

1. System Quirk software Analysis of the Enron Corpus to confirm experiment suitability re: evidence of banners and consistency of distribution.
2. Construct a test dataset by “eyeballing” a small number of confidentiality banners and identifications of confidentiality in the Enron Corpus
3. Identify an initial set of similarities that enable a skilled human to make a binary decision.
4. System Quirk software analysis to determine whether the similarities have any statistical significance, using word frequency, word weirdness and word frequency/proximity statistical analysis on a training set
5. Evaluate the approach against a larger part of the Enron Corpus.
6. Define the banner context boundaries (these may be different from their physical text boundary.)

For the purpose of this paper, the “obvious” human choices for keywords using similarities (step 3) is not necessarily the best and is provided as a comparative to the proper statistical analysis (step 4) which can reveal easier and better patterns to exploit, a point well made elsewhere [24].

4. Experiments

4.1 Finding the structure

As noted above, the keyword “confidential” could be used to indicate material of a sensitive nature, but is now prevalent in email privacy banners. As such, even following the

removal of email headers there will remain some proportion of content that is not interesting for analytical purposes – beyond, perhaps, understanding the structure of such banners. Besides, the act of removal suggests the need to understand their structure. An example of such a banner is included below. Note the length of the banner could make up a large proportion of the text in brief messages and contribute to corpus pollution. The banners will also act to conceal the behaviour in free running text, of a variety of other contained words.

```
+++++CONFIDENTIALITY NOTICE+++++
The information in this email may be
confidential and/or privileged. This email is
intended to be reviewed by only the individual
or organization named above. If you are not the
intended recipient or an authorized
representative of the intended recipient, you
are hereby notified that any review,
dissemination or copying of this email and its
attachments, if any, or the information
contained herein is prohibited. If you have
received this email in error, please immediately
notify the sender by return email and delete
this mail from your system. Thank You
```

Figure 4-1: Example Banner

Enron-Raw contains 35,621 instances of the word confidential, and collocates analysis suggests that more than fifty percent of these are from banners. Using “confidential” as the nucleate of our collocations, frequencies of collocating words in a 5-word window (L5-R5, with adjacent frequencies at L1 and R1) in Enron-Raw are as show in Table 4-1 (ordering in the Table is based on further analysis with these values, according to the work of [7], but not presented here):

Collocate	Total	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
and	30584	623	300	7512	1961	3310	15262	768	230	372	246
may	15264	307	1144	161	10290	5	74	2815	249	190	29
contain	11004	1613	205	428	7	8630	0	0	38	83	0
the	13886	459	1485	681	160	473	151	551	966	667	8293
for	10129	493	377	149	88	116	105	296	629	7228	648
privileged	19390	68	13	1375	4780	1647	71	6599	2593	1398	846
material	6367	2	0	30	0	0	21	22	5122	1153	17
relevant	4863	4856	7	0	0	0	0	0	0	0	0
information	11143	688	1013	704	271	123	5379	338	1111	715	801
affiliate	5051	185	4855	7	0	0	0	0	0	0	4

Table 4-1: Enron-Raw “confidential” collocations

Collocations with “confidential” appear to suggest a pattern similar to that of the example confidentiality notice. A clearer pattern emerges with the simple removal of the 2000 most frequent words of the BNC (Table 4-2).

Word	Total	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
contain	11004	1613	205	428	7	8630	0	0	38	83	0
privileged	19390	68	13	1375	4780	1647	71	6599	2593	1398	846
affiliate	5051	185	4855	7	0	0	0	0	0	0	4
legally	4724	3	1117	342	71	0	70	139	2475	499	8
intended	3990	69	17	4	0	0	10	2535	516	570	269
exempt	2480	0	0	0	0	0	0	430	218	1832	0
proprietary	3097	147	107	1726	70	7	649	258	129	3	1
unauthorized	1415	0	0	0	0	0	0	0	8	0	1407
solely	1399	0	0	1	0	0	0	9	1275	98	16
email	2864	63	510	932	506	1	5	4	21	5	817

Table 4-2: Enron-Raw “confidential” collocations: BNC top 2000 removed

Significant peaks for “privileged” can be seen at L2 and R2. We cannot yet discount the possibility that there are substantial contributions to these values from body text. The values above appeared to be increasingly indicative of banners.

4.2 Checking consistency

To attempt to discover robust statistics for banner keywords, for about 40% of the raw corpus we obtained collocation statistics for use of the word “confidential”, not considering the 2000 most frequent words of the BNC. In this subset, 14,384 instances of confidential were found. We further split this subset into four, on the basis of folder names alone, and looked at the proportions of collocates with “privileged”.

Count	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	
1	918	4	1	17	132	85	2	539	87	41	10
%	11.3	0.44	0.11	1.85	14.38	9.26	0.22	58.71	9.48	4.47	1.09
2	1737	2	1	116	234	91	6	831	239	92	125
%	21.4	0.12	0.06	6.68	13.47	5.24	0.35	47.84	13.76	5.30	7.20
3	3465	1	1	308	1377	270	13	824	465	107	99
%	42.7	0.03	0.03	8.89	39.74	7.79	0.38	23.78	13.42	3.09	2.86
4	2002	12	0	128	364	125	8	922	212	83	148
%	24.6	0.60	0.00	6.39	18.18	6.24	0.40	46.05	10.59	4.15	7.39
Total	8122	19	3	569	2107	571	29	3116	1003	323	382

Table 4-3: Enron-Raw 4 subset comparison, collocations centred on “confidential”

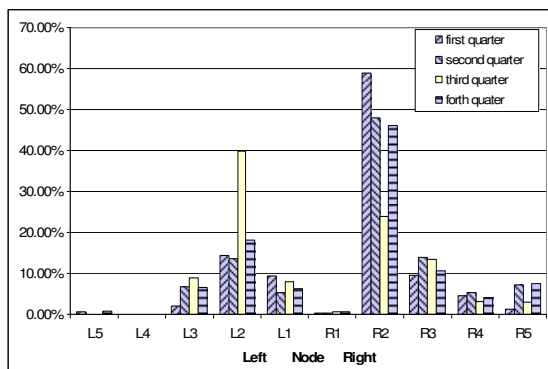


Figure 4-2: Enron-Raw comparison of collocation for “confidential” and “privileged” across four subsets of a proportion of the corpus.

The pattern “confidential X privileged” accounts for 39% of the overall. The four subsets tend towards slightly different patterns: for three of these, this pattern is rather higher, while for the fourth, the pattern “privileged X confidential” shows a peak.

The intention here is to ascertain probabilities for “confidential” and “privileged” co-occurring at location n, and to determine the extent to which this holds across the corpus. Extending this further, to consider the probability of “prohibited” co-occurring with this pattern at location m, would improve confidence in detection. We would be attempting to discover a set of optimum values, or ranges, which can be used for confidence that the item we are dealing with is a banner. Given the variation seen above, we plan to continue this work to make determinations over the whole corpus and in contrast with other corpora.

4.3 Corpus creation

From this base we proceeded to extract from the Enron Corpus produce training and test sets with confidence that samples may be taken from one part of the corpus without any significant concern as to consistency. The training set containing 50 unique banners and 46 body paragraphs (each with at least one instance of the word “confidential”) was created manually by “eyeballing” a number of emails.

The test corpus was developed from a collection of emails in the Enron Corpus that contained “confidential”. A subset of this collection, based on the first 25 email account names in alphabetical order, was selected as the first test corpus collection. This collection was manually evaluated to determine whether the instances of “confidential” were in banners or body. To ensure that these could be treated separately, and in lieu of so-called “stand-off”, or “multidimensional” annotations, each banner instance was replaced with “zzzzzzzzzzial” (3223 in total) and each body instance with “xxxxxxxxxxxxial” (2663 in total), effectively tagging each.

4.4 Choosing the words for the Banner discrimination

Similarities in the use of words such as “privileged” at a short distance from the keyword “confidential” were initially noted. We performed word frequency analysis, with and without stop words, and calculated values for “weirdness” using the British National Corpus (BNC) to identify and contrast prevalent keywords in the “banner” and “body” test sets. Table 4-4 shows the top 10 keywords discovered for each: there are some indications of difference, given the spreads of frequency values in these top 10s, and note that “privileged” is shared between these sets, albeit at a greater frequency in the banners.

Key Words: Body			Key Words: Banners		
Freq	Weirdness	Word	Freq	Weirdness	Word
64	2763	confidential	68	969	mail
22	inf!	enron	66	288	intended
9	456	transportation	51	1925	confidential
8	1022	confidentiality	46	1494	recipient
8	258	agreements	32	inf!	email
8	228	privileged	32	798	privileged
7	inf!	ferc	30	1581	sender
7	7456	ena	29	2700	prohibited
7	677	disclosure	28	245	error
7	20	non	27	1178	delete

Table 4-4: Top 10 Keywords discovered in Body and in Banner paragraphs

Next we calculated frequencies of words within a 5 word window of the keyword “confidential” across the whole Enron Corpus (209,204,013 tokens, according to System Quirk computations) and compared this to the extracted banners. Consider, for example, occurrences of “privileged” within this 5 word window – in the Enron Corpus, “confidential” occurs 35621 times. The word “privileged” occurs 19390 times within 5 words either side of this. Of these 19390 times, it occurs 6599 times at one word separated from confidential (at position 2, e.g. “confidential X privileged”). A further 4780 occurrences are opposite to this (“privileged X confidential”). See Table 4-5. Further

details about the statistical significance of these values can be found in [7]

Position	-5	-4	-3	-2	-1	1	2	3	4	5
Frequency	68	13	1375	4780	1647	71	6599	2593	1398	846

Table 4-5: Frequencies of the word “privileged” within a window of 5 words of “confidential”

The extent to which the 35621 instances of “confidential” denote a banner can be assessed by contrasting the totals of collocating frequencies with the frequency analysis of the eyeballed banners Table 4-6. The top 22 words collocating with “confidential” are indexed by the first column. These indexes are used in brackets after the identical words found in the lists generated by frequency and weirdness calculations. Differences in ranking due to frequency and weirdness calculations can be seen by alphabetic indexes. According to these results, a relatively large proportion of the instances of “confidential” appear to be indicative of banners, though the true extent remains to be assessed.

To confirm that the Enron Corpus was statistically similar across email account names and that the Banner training selection was a representative sample, we performed a proximity (+/-5 words to confidential) frequency analysis across 60 million tokens of the raw corpus and then compared the top 22 words of the whole to the top 22 words from the banner training sample for frequency and weirdness. The impact of stemming and lexical variation remains to be assessed.

Proximity raw Corpus frequency	Banner By Frequency	Banner By weirdness
1 privileged	8122	information (3) 68
2 contain	4902	mail (a) 68
3 information	4722	intended (8) 66
4 material	2818	message (11) 61
5 affiliate	2318	recipient (19) 46
6 relevant	2305	please 45
7 legally	1837	email (10) 32
8 intended	1594	privileged (1) 32
9 proprietary	1340	sender (b) 30
10 email	1185	received 30
11 message	1078	prohibited (c) 29
12 exempt	1075	error (d) 28
13 otherwise	952	delete (e) 27
14 subject	947	immediately 27
15 enron.com	750	notify (f) 27
16 contains	726	copying (g) 22
17 communication	684	other 21
18 solely	622	distribution 20
19 recipient	612	contain (2) 19
20 protected	606	attachments (22) 19
21 e-mail	592	communication (17) 19
22 attachments	589	disclosure (h) 18
		named 68

Table 4-6: Banner/raw corpus sample

In Table 4-6 we noted that six words (in bold) were common to all columns and felt that these 6 words would be a logical choice to for our first keyword instance list. We decided, also, that instances collocating within, approximately, one sentence of our target key word “confidential” could be of interest, but would assign less importance to those at a greater distance. Since 15 to 20 words is a good length for a sentence, we expanded our window of consideration to 20, without consideration for sentence [27] boundaries, and weighted each word inversely

proportional to distance.

4.5 Banner Discrimination computation

We computed individual weights for all “confidential” key word instances in both banner and body. The resulting graph, Figure 4-3 shows the error % (1-precision) against trigger weight for body and banner. The sub sample of the first 25 names within the Enron Corpus was used with the 5886 manually “tagged” instances of confidential. These at a trigger level greater than 0.5, 46 from 2663 instances (1.7%) false negatives would be generated, and 2737 false positives (84.9%) would now be correctly filtered.

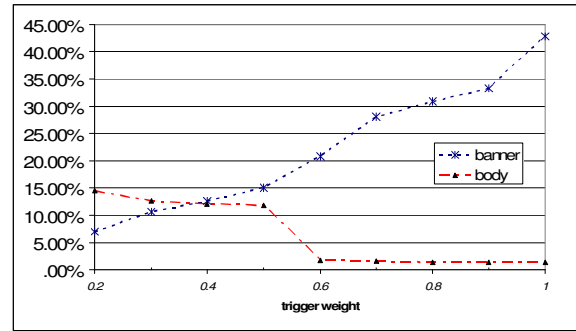


Figure 4-3: Error % against trigger weight

These initial results were encouraging, however we needed a further assessment of the three key assumptions: (i) best distance – whether a 20 word window was a good choice; (ii) impact of weighting on precision; (iii) lexical selection – quality of the chosen word list.

(i) We used max distance at values of 3, 5, 10, & 20 and plotted the effects of max distance on precision see Figure 4-4. For body instances, no significant change in precision resulted; for banners, reducing the max distance caused a reduction in precision. This indicated that the instance word list data in the surrounding area was relatively rare in the body case.

(ii) We removed the discount for distance, and evaluated results at a maximum distance of 10 & 20. Results of the effects of max distance on precision can be seen in Figure 4-5. This showed that the attenuation was actually having a detrimental effect on body precision, and a beneficial effect on banner precision. However with such a small word instance list the granularity may be considered crude.

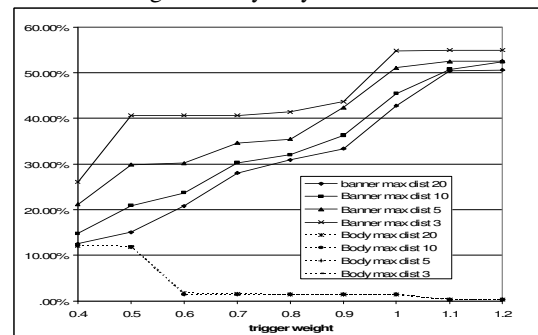


Figure 4-4: Error % against maximum distance

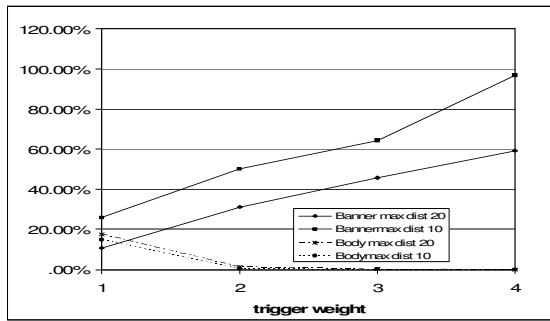


Figure 4-5: Error % against maximum distance

(iii) We looked again at Table 4 6 and ran the experiment using these three difference keyword sets of 22 words each – see Figure 4-6.

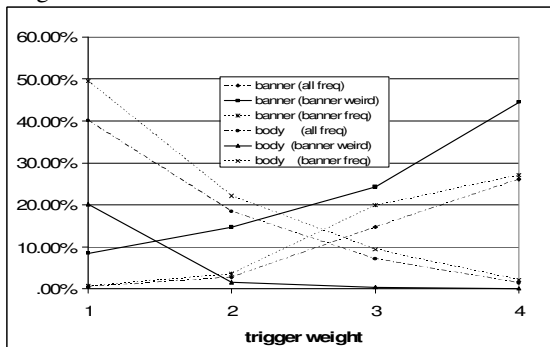


Figure 4-6: Error % against different word sets

Surprisingly, the Proximity raw Corpus frequency set (all freq) out performed (banner freq), showing that there was a significant pattern coming from the banners in the raw Enron Corpus. The most frequent banner set (banner freq) did reasonably well, but not as well as expected. The most significant improvement came from the weird set (Banner weird), with exceedingly good results, with a trigger level set to greater than 2, only 10 body confidential instances or 0.37% would be miss categorized and not presented to a human for inspection and 2752 banners or 85.4% would be correctly filtered. For others to give good results for Body categorization required a trigger weight of 4 and had significantly worse banner discrimination characteristics. Following this result we then re-examined the statistics from the training corpus and produced a table in banner weirdness order against body frequency see Table 4-7. This demonstrated that the weirdest words were on the most part very or exceedingly rare in the body text. So the best way of choosing instance words for the banner filter was to use some function of banner weirdness and body rarity, for example techniques from, [6] [14] [15] in a different orientation.

Body Freq	Word	Banner Freq	Weirdness
1	email	32	inf!
0	dissemination	14	2793
0	prohibited	29	2700
0	attachments	19	2123
0	sender	30	1581
7	disclosure	18	1523
0	recipient	46	1494
0	delete	27	1178
0	notify	27	1077

5	mail	68	969
0	copying	22	945
8	privileged	32	798
0	addressee	15	340
1	intended	66	288
0	error	28	245
0	solely	12	229
0	strictly	17	197
1	contained	15	98
0	contain	19	95
2	copy	11	77
0	contains	11	73
0	named	13	68

Table 4-7: Body Freq/Banner Weirdness

4.6 Banner Context Zoning

Cleanly and efficiently extracting, delimiting, or otherwise removing banners are not a simple mechanistic process. According to our investigations, the banners appear to have some comparable structure, but do not follow a strict format according merely to email system protocols, as would be expected for email headers. Banners may be very different for each originating organisation. One may also have an expectation that email headers appear at the top of emails, and banners appear as footers, however the reality of quoted emails means that we could be searching for multiple instances of both throughout a given email. The challenge, then, is to identify the “zone” of “context” for each banner within an email and to successfully delimit it. This notion of zoning is inspired, in part, by Teufel’s work on attribution of scientific text [8], and may be helpful in dealing with quoted responses.

In the above we demonstrated the results of frequency analysis on 100 manually extracted instances of “confidential”, comprising 50 unique banners and 50 non-banner paragraphs. Results of the analysis were compared to the BNC and to a subset of the Enron Corpus. The manual extraction step demonstrated that banners consist of a large, but relatively limited set, of words, and in some instances account for a large proportion of the email body. Using a fixed-distance window and simple summation produced good discrimination for banner and body: 85.4 percent of banner instances were correctly identified and 0.37 percent of body instances were incorrectly identified as banners. In the application domain, body instances should be presented to a human for inspection, while banners incorrectly presented are of less importance than body instances not being presented (missed messages).

We expand on this work, analysing 3226 manually extracted confidentiality banners. We considered an expansion to distances of 120 words either side of our selected keyword as a means to detect the extent of the banner. This contrasts with traditional analysis of collocations, presented above, although we are using the same keyword set. On the basis of this analysis we can identify a clustering effect, with certain words dominant in particular positions, and suggest that banners are, on average, around 80 words in length (Figure 4-7, Table 4-8). We can see two interesting peaks closely centred on “confidential”:

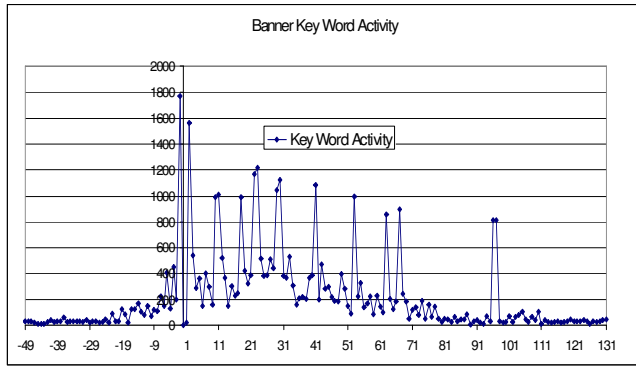


Figure 4-7: Key word activity surrounding “confidential”

Word	Distance
contain	-1
privileged	+2
intended	+10, +29, +67, +96
recipient	+11; +30, +97
disclosure	+18
strictly	+22
prohibited	+23
sender	+41
delete	+53
attachments	+63

Table 4-8: Key word activity peaks

There are two potential conclusions from the peaks identified above: (i) there are a lot of identical or very similar banners within the corpus; (ii) banners are large constructs with a predictable structure.

Analysis of the same keywords as above, centred on “confidential”, for body text produces a substantially different result Figure 4-8. The results show one peak, and further investigations have shown that the source of this is email correspondence with lawyers involved with litigation actions. The peak does not coincide with the banner instances for “privileged”, and may partly explain the results in subset 3 of Table 4-3 at position L2.

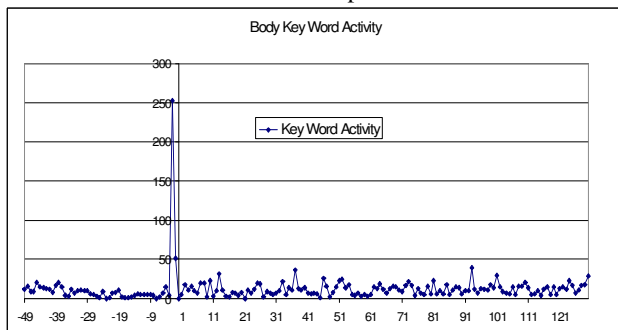


Figure 4-8: Key word activity surrounding “confidential” in ordinary body text

4.7 Finding the banner context zone boundaries

There is a difference between the context zone boundary and the actual physical textural boundary of the context object. The discrimination process has so far identified the presence of the object and only knows it’s bounds in terms of the outermost key words of the context constellation and this will err on the small side.

In our domain of application key words outside the context zone boundary are deemed body and hence presented to a human for final decision making. This implies that if the context boundary heuristic errs on the small side there could be false positive presentations to the human for key words other than the banner context constellation key words. An assessment of how much banner residual data lies beyond the outermost keywords was made on the 50 training banners. The table below shows the top 15 residual banner data key words in frequency order after removal of the top 2000 BNC words.

Match	Frequency	Match	Frequency
transmitted	4	notice	3
author	3	attachments	2
confidentiality	3	deleting	2
contents	3	destroy	2
contract	3	disclose	2
copies	3	endorsed	2
disclaimer	3	estoppel	2
electronic	3	etol	2

Table 4-9: Top 15 “residual data” key words (less BNC top 2000)

The constellation of key words used for banner detection uses the top 22 most weird words, it has not been optimised nor tuned to minimise the residual data, yet the results are promising. There are of course other methods for determining the physical boundary which may produce less residual data, such as finding the paragraph or sentence start end using grammar, capitalisation and punctuation, or by using the known context object (in this case banners) as a nearest fit example for the boundary. This method and the other above mentioned methods will be investigated later and compared for accuracy and precision across the whole Enron Corpus.

4.8 Improved discriminator

On the basis of this evidence, we have developed and evaluated an algorithm for discriminating banner and body text use of “confidential”. The original algorithm used in the initial analysis has been modified for a window of -25 to +115 words, and scores according to the sum of the keyword positional evidence given by collocations of both positive evidence and negative evidence. Figure 4-9 shows the positional evidence by collocation for “privileged” with “confidential” as the node, using the manually “tagged”

5886 instances from the first 25 names.

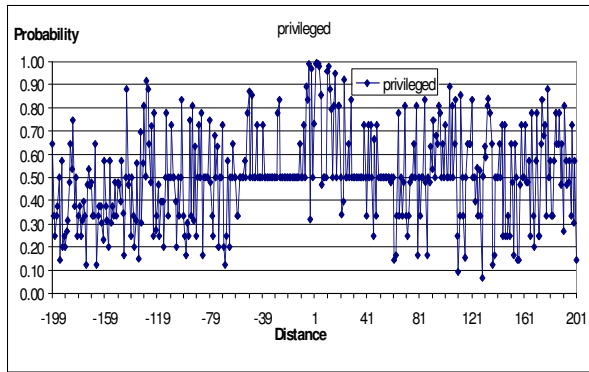


Figure 4-9: Privileged probability by collocation position

No evidence equals 0.5, strong positive evidence tends to 1 and strong negative evidence tends to zero. Hence the non banner instances at L2 (Figure 4-2 & table 4-2) results in a low weighting.

With the trigger point set to 2.75, this resulted in 91.7percent of banner instances correctly identified and only 0.3percent of body instances incorrectly identified (Figure 4-10), with minimal improvements at higher values.

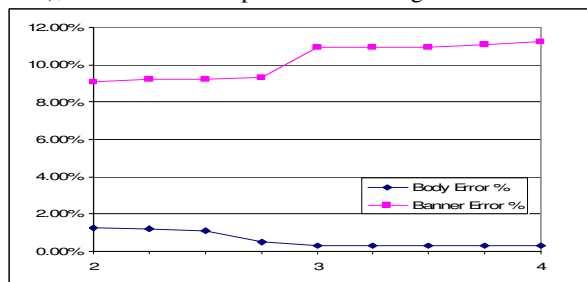


Figure 4-10: Misidentification percent against trigger point (total weight)

Assuming that our discoveries for collocation between “confidential” and “privileged” hold across the corpus, which we will be investigating, we should be able to remove a very high proportion (91.7 percent) of the 20,000 estimated banners from the corpus. Further evaluation, both manual and automatic, is planned, and the results will be published in due course.

5 Related Work

Work on the Enron Corpus elsewhere has investigated automatic classification of emails as “Business” or “Personal” based on inter-annotator agreement [20]. The authors suggest that around 17% of a sample of around 12,500 emails were identified as personal correspondence, based on 94% agreement between 4 annotators, and a probabilistic classifier reportedly achieves good performance against a subset of these documents. This work is directly related to Step 7 of our approach, and it will be interesting to measure the extent to which banners might act as useful classifiers for business emails.

6 Conclusions

In this paper we discussed the ease with which email can be used for breaches of confidence and the potential for harm to organizations as a result. We identified a lack of literature regarding the problem of correctly identifying such potential breaches. We have proposed an intelligent filtering system for outgoing emails aimed at preventing such disclosures. We have demonstrated through a number of relatively straightforward, yet encouragingly effective, experiments how the use of a few techniques from the field of corpus linguistics could be used to reduce the number of false alarms – false positives - produced by keyword filtering. We have also considered the proportion of harmful false negatives. These experiments were undertaken on the publicly accessible Enron email corpus. These early results are highly promising, and future work aimed at improving on these initial results is already in progress and will be reported when fully verified.

Our attempts to identify confidentiality banners, deal with email headers, and subsequently to deal with other vagaries of email systems are steps towards this. Correct delimitation of the zone or zones occupied by banners within emails will help to ensure that we are dealing, more or less, with email content. Manual verification at various stages of the automation will be required, but with the intention of moving towards greater levels of automation. Work to date has demonstrated that automatic identification of banners in the “full” Enron Corpus is highly possible, but will have to be provably accurate for use in mission-critical enterprises.

References

- [1] D. Leonard-Barton, *Wellsprings of Knowledge*, Harvard Business School Press, ISBN 0-87584-859-1 Boston MA, 1998.
- [2] T.H. Davenport, L. Prusak, *Working Knowledge*, Harvard Business School Press, ISBN 1-57851-301-4 Boston MA, 1998.
- [3] I. Nonaka, H. Takeuchi, *The Knowledge Creating Company*, Oxford University Press, ISBN 0-19-509269-4 New York, 1995.
- [4] J. Forder, P. Quirk, *Email policies considered*, Faculty of Law, Law papers, Bond University, 1998
- [5] K.M. Carley and D. Skillicorn, *Special Issue on Analyzing Large Scale Networks: The Enron Corpus*, Computational & Mathematical Organization Theory 11(3), Kluwer Academic Publishers, 2005
- [6] L. Gillam, *Systems of concepts and their extraction from text*. Unpublished PhD thesis, University of Surrey, 2004
- [7] F. Smadja, “Retrieving collocations from text”, *Xtract Computational Linguistics Oxford university Press*, 19(1) March 1993, ISSN:0891-2017, 1993.
- [8] S. Teufel, M. Moens, “What's yours and what's mine: Determining Intellectual Attribution in Scientific Text”, *Proc. of 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong, 2000

- [9] G. Fumera, I. Pillai, F. Roli, "Spam Filtering Based On The Analysis Of Text Information Embedded Into Images", *Machine Learning Research* 6: 2699-2720, 2006
- [10] E. Damiani, S.D.C. di Vimercati, S. Paraboschi, P. Samarati, "An Open Digest-based Technique for Spam Detection". *Proc. of ISCA PDCS 2004*: 559-564, 2004.
- [11] J. Dong, H. Cao, P. Liu, L. Ren, "Bayesian Chinese Spam Filter Based on Crossed N-gram", *Proc. of ISDA 2006* Volume 3, pp:103 – 108, October 2006.
- [12] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras, C.D. Spyropoulos, "An evaluation of naive Bayesian anti-spam filtering". *Proc. of Workshop on Machine Learning in the New Information Age, ECML 2000*. Barcelona, Spain, 9—17, 2000.
- [13] K.M. Schneider, "A comparison of event models for Naive Bayes anti-spam e-mail filtering", *Proc. of ACL 2003*, Budapest, Hungary, April 12-17, 2003.
- [14] L. Gillam, M. Tariq, K. Ahmad, "Terminology and the construction of ontology". *Application-Driven Terminology Engineering*, Ibekwe-SanJuan, Fidelia, Anne Condamines and M. Teresa Cabré Castellví (eds.), 49–73, John Benjamins Publishing Company, 2007.
- [15] L. Gillam, K. Ahmad, "Pattern mining across domain-specific text collections". *LNAI 3587*, pp 570-579, 2005
- [16] S. Chatterjee, "Enron's Incremental Descent into Bankruptcy: A Strategic and Organisational Analysis". *Long Range Planning* 36(2), pp. 133-149(17). Elsevier, 2003
- [17] B. Klimt, Y. Yang, "The Enron Corpus: A New Dataset for Email Classification Research"; *ECML 2004*, Language Technologies Institute, Carnegie Mellon University, 2004
- [18] R. Bekkerman, A. McCallum, G. Huang, "Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora", *UMass CIIR Technical Report IR-418*, (Massachusetts), 2004
- [19] P.S. Keila, D.B. Skillicorn, "Detecting Unusual and Deceptive Communication in Email". *External Technical Report* ISSN-0836-0227-2005-498. Queen's University, CA 411, 2005
- [20] S. Jabbari, B. Allison, D. Guthrie, L. Guthrie, "Towards the Orwellian Nightmare: Separation of Business and Personal Emails". *Proc. of COLINGACL 2006 on main conference poster*. 2006
- [21] Y. Jewkes Editor *Dot.cons: Crime, Deviance and Identity on the Internet*, *Cullompton: Willan Press*, 256 pp., ISBN 1843920018, 2003
- [22] D. Trudgian, Z R Yang, "Spam Classification Using Nearest Neighbour Techniques". *Proceedings of Fifth International Conference on Intelligent Data Engineering and Automated Learning IDEAL04*, Exeter, UK, Lecture Notes In Computer Science, vol. 3177, pp. 578-585. 2004
- [23] B. Medlock, "An Adaptive, Semi-Structured Language Model Approach to Spam Filtering on a New Corpus", *Cambridge University Computer Laboratory; CEAS 2006 Third Conference on Email and Anti-Spam*, Mountain View, California USA. 2006
- [24] G.V. Cormack, "TREC 2006 Spam Track Overview", *Proceedings of TREC 2006: The Fifteenth Text Retrieval Conference*, November 2006.
- [25] G.V. Cormack and T.R. Lynam., "TREC 2005 Spam Track Overview". *Proc. TREC 2005 - the Fourteenth Text REtrieval Conference*, Gaithersburg, 2005.
- [26] G.V. Cormack and T.R. Lynam, "Spam Corpus Creation for TREC", *Proc. CEAS 2005 - The Second Conference on Email and Anti-Spam*, Palo Alto, July 2005.
- [27] Plain English Campaign, How to write medical information in plain English, *Plain English Campaign*, 2001

Author Biographies

Lee Gillam PhD in Artificial Intelligence (Surrey, 2004); BSc in Mathematics and Computer Science (Surrey, 1995). Member of the British Computer Society (MBCS). Currently a Lecturer in the Department of Computing at the University of Surrey. Previous publications and research in the areas of Ontology Learning, Metadata and Grid Computing Systems.

Neil Cooke BSc(2.1 hon) Computer & Control Systems Lanchester Polytechnic Coventry UK 1981 (now University of Coventry), Chartered Engineer 1987, Fellow of the Institute of Engineering Technology 2005. The author is part time studying for a PhD, the research subject area is context sensitive filtering of emails. The author has 17 years of experience in the field of information assurance engineering for the UK government's National Technical Authority for Information Assurance. Prior to this the author was an avionics and marine systems engineer.