IP protection: Detecting Email based breaches of confidence

Neil Cooke
*University of Surrey
and CESG, GCHQ
n.cooke@surrey.ac.uk
neil.cooke@cesg.gsi.gov.uk*

Lee Gillam
*University of Surrey
Guildford, UK
l.gillam@surrey.ac.uk*

Ahmet Kondoz
*University of Surrey
Guildford, UK
a.kondoz@surrey.ac.uk*

Abstract

In this paper we discuss the ease with which email can be used to breach confidence by the propagation of corporate secrets and intelligence, and propose an intelligent filtering system for outgoing emails aimed at preventing disclosures. We report on a number of experiments undertaken with a corpus of over half a million Enron emails and the use of a variety of techniques from the field of Corpus Linguistics for reducing the number of false alarms produced by naïve keyword filtering systems, and discuss the results in detail. We also give due consideration to the danger of missing messages that should have been prevented from propagation.

Keywords: Intellectual property; breach of confidence; text analysis; outgoing email filters

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 emerges in the form of copyright, designs, patents and trade marks. 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, for example the copyrighted article, or the patent application.

Early knowledge management literature focused on knowledge as processes, on the ability to convert between “tacit” and “explicit” forms of the known, and on storing knowledge within, and extracting knowledge from, corporate databases [1], [2], [3]. 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, it appears, was aimed at managing all of what a company “knows”, from client lists to customer relationships to business processes to trade secrets. The law of confidentiality applies to ensuring that these high-value collections remain known only to the organization, and are not disclosed to others in ways that would cause harm to the organization or benefit to its competitors. Breach of confidence tends to make headlines when a disaffected employee, or ex-employee, purposely discloses such corporate property to the public at large or to competitor organizations.

In this paper we consider the potential for breaches of confidence to occur rapidly and on a large scale, and the difficulty of preventing such disclosures of, in some cases, corporate intellectual property, by employees using email systems. If employees are easily able to distribute the company’s secrets around the world in a few seconds by email, or perhaps by other insecure electronic means[1], 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 also be capable of ensuring that outgoing emails are unlikely to contain information that would otherwise be detrimental to the organization, and perhaps of ensuring that corporate

[1] USB memory sticks, as the US military discovered, provide yet another information security issue: http://tinyurl.com/22fayq

policies preventing the personal use of email are being correctly adhered to.

We discuss a number of initial 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 released into the public domain by the Federal Energy Regulatory Commission. We explore the use of a number 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. 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. Little appears to have been published, outside of corporate pamphlets and legal advice[2] on this subject and available techniques and their accuracy, and we've found no direct consideration of the problem of false positives raised due to confidentiality banners.

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 patented[3]. The ready accessibility of spam filtering systems means that companies are implementing them at the same time that spammers are using them to create emails that successfully pass through the filters, and variations of words that include misspellings and the incorporation of "foreign" characters or numbers can be used that remain generally readable, e.g. *vïagara*. Keyword-based approaches to spam filtering are defeated, also, by the incorporation of text into images [4]. Collaborative filtering [5], where a group of users effectively "vote out" emails as spam by adding these emails to a central database, have proven variously successful. Such techniques, combined with white-lists and black-lists, Bayesian filtering [6], [7], [8], 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 making an assumption that users are, more often than not, only involved in unintentional disclosure. Arguably, therefore a keyword-based approach should be effective, and there are many commercial offerings which provide security features for outgoing emails, and the majority of these are incoming mail guards used in a different orientation. However, while a simple keyword filtering approach may be helpful on a small scale, the keyword "confidential" used as a filter will result in 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" keyword-only blocking 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 a list 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 its 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] http://tinyurl.com/2spz4n

[3] USPTO 6,609,196: http://tinyurl.com/26koxg

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 website[4]. 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 [9]. 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 financial trading [10], [11]. 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 Enron email dataset[5] consists of the email folders of 158 Enron employees, providing a total of 619,446 emails [12]. 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. 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.

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. Constructing a test dataset by "eyeballing" a small number of confidentiality banners and identifications of confidentiality in the Enron corpus
2. Identifying an initial set of similarities that enable a skilled human to make a binary decision.
3. Using the System Quirk software to determine whether the similarities have any statistical significance, using word frequency, word weirdness and word frequency/proximity statistical analysis on a training set
4. Evaluating the approach against the full Enron corpus.
5. Classifying emails as containing confidentiality indicators in (a) unseen banners; (b) body text; (c) both.
6. Constructing a confidentiality banner database for further evaluation.
7. Assessing the email corpus for further features, e.g. personal vs. business emails as may be discernible by register variation.

For the purpose of this paper, we are concerned with steps 1-5. The "obvious" human choices for keywords and similarities (steps 1-2) are not necessarily the best, and proper statistical analysis can reveal easier and better patterns to exploit, a point well made elsewhere [13].

4. Experiments

A 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. 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 1 shows the top 10 keywords discovered for each: there are some

[4] Available at: http://www.computing.surrey.ac.uk/SystemQ/

[5] Available at: http://www.cs.cmu.edu/~enron/

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.

Table 1: Top 10 keywords discovered in body and in banner paragraphs

Keywords: Body			Keywords: 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

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 2. Further details about the statistical significance of these values can be found in [14]

Table 2: Frequencies of the word "privileged" within a window of 5 words of "confidential"

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

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 3). 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 results 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.

Table 3: Banner/raw corpus sample

Proximity raw Corpus frequency			**Banner By Frequency**		**Banner By weirdness**	
1	**privileged**	8122	information (3)	68	**email** (10)	inf!
2	**contain**	4902	mail (a)	68	dissemination	2793
3	information	4722	**intended** (8)	66	prohibited (c)	2700
4	material	2818	message (11)	61	**attachments** (22)	2123
5	affiliate	2318	**recipient** (19)	46	sender (b)	1581
6	relevant	2305	please	45	disclosure (i)	1523
7	legally	1837	**email** (10)	32	**recipient** (19)	1494
8	**intended**	1594	**privileged** (1)	32	notify (f)	1077
9	proprietary	1340	sender (b)	30	delete (e)	1178
10	**email**	1185	received	30	mail (a)	969
11	message	1078	prohibited (c)	29	copying (g)	945
12	exempt	1075	error (d)	28	**privileged** (1)	798
13	otherwise	952	delete (e)	27	addressee	340
14	subject	947	immediately	27	**intended** (8)	288
15	enron.com	750	notify (f)	27	error (d)	245
16	contains	726	copying (g)	22	solely (18)	229
17	communication	684	other	21	strictly	197
18	solely	622	distribution	20	contained	98
19	**recipient**	612	**contain** (2)	19	**contain** (2)	95
20	protected	606	**attachments** (22)	19	copy	77
21	e-mail	592	communication (17)	19	contains (16)	73
22	**attachments**	589	disclosure (i)	18	named	68

In table 3 we noted that six words **(in bold)** were common to all columns and felt that these 6 words would be a logical choice for our first keywords. 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[6], we expanded our window of consideration to 20, without consideration for sentence boundaries, and weighted each word inversely proportional to distance. We considered only 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 treated. 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 external annotations, each banner instance was replaced with "zzzzzzzzzzial" (3223 in total) and each body instance with "xxxxxxxxxxxial" (2663 in total), effectively tagging each.

We computed individual weights for all "confidential" key word instances in both banner and body. The resulting graph, figure (1) shows the error % (1-precision) against trigger weight for body and banner. 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.

[6] http://www.plainenglish.co.uk/medicalguide.pdf

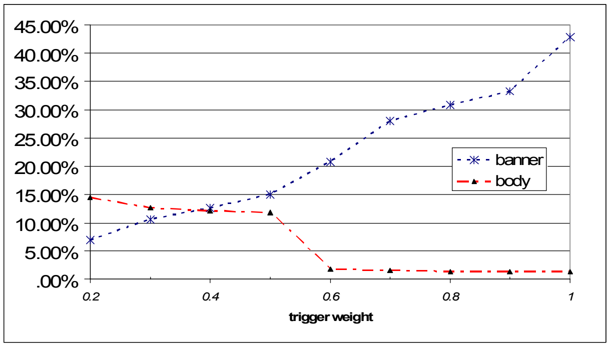

Figure 1 Error % against trigger weight

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 (2). 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 (3). 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.

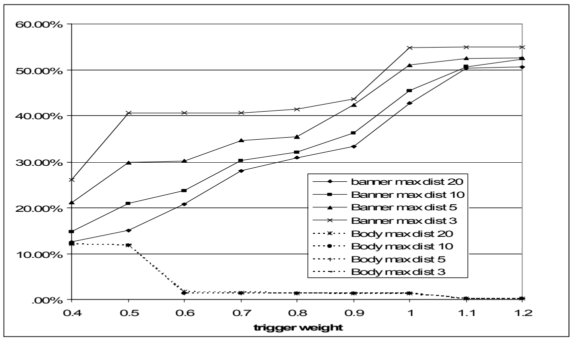

Figure 2 Error % against maximum distance

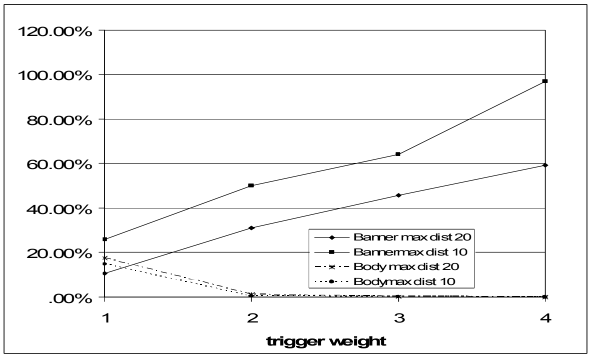

Figure 3 Error % against maximum distance

(iii) We ran the experiment using the three different keyword sets of table 4 - see figure (4).

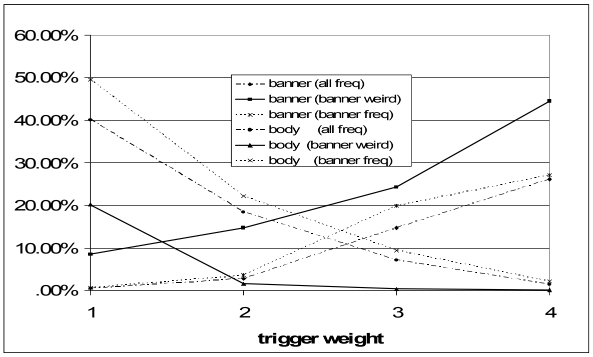

Figure 4 Error % against different word sets

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 (0.37%) would be mis-categorized and not presented to a human for inspection and 2752 banners (85.4%) would be correctly filtered. For others to give good results for Body categorization required a trigger weight of 4, resulting in significantly worse banner discrimination characteristics. Following this result we reexamined the statistics from the training corpus and produced a table in banner weirdness order against body frequency (Table 4). This demonstrated that the weirdest words were 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 [9], [10], [11] in a different orientation.

Table 4: Body Freq/Banner Weirdness

Body	Word	Banner	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

5. Related Work

Work on the Enron corpus elsewhere has investigated automatic classification of emails as "Business" or "Personal" based on inter-annotator agreement [15]. 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 probabalistic 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, and 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 and 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 work aimed at further improvements over these results is already in progress and will be reported when fully verified.

7. References

[1] Leonard, D., *Wellsprings of Knowledge*, Harvard Business School Press, Boston MA, 1998.

[2] Davenport, T.H., Prusak, L., *Working Knowledge*, Harvard Business School Press, Boston MA, 1998.

[3] Nonaka, I., H. Takeuchi *The Knowledge Creating Company*, Oxford University Press, New York, 1995.

[4] G. Fumera, I. Pillai, and F. Roli, "Spam Filtering Based On The Analysis Of Text Information Embedded Into Images", *Machine Learning Research* 6: 2699-2720, 2006

[5] E. Damiani, S.D.C. di Vimercati, S. Paraboschi and P. Samarati, "An Open Digest-based Technique for Spam Detection". *Proc. of ISCA PDCS* 2004: 559-564, 2004.

[6] J. Dong, H. Cao, P. Liu, and L. Ren, "Bayesian Chinese Spam Filter Based on Crossed N-gram", *Proc. of ISDA 2006* Volume 3, pp:103 – 108, October 2006.

[7] I. Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G., and Spyropoulos, C., An evaluation of naive Bayesian anti-spam filtering, *Proc. of ECML 2000*, Barcelona, Spain, 9—17, 2000.

[8] 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.

[9] L. Gillam, Systems of concepts and their extraction from text, Unpublished PhD thesis, University of Surrey, 2004.

[10] L. Gillam, M. Tariq, and K. Ahmad, Terminology and the construction of ontology, *Application-Driven Terminology Engineering*, John Benjamins, 2007

[11] L. Gillam and K. Ahmad, Pattern mining across domain-specific text collections, *LNAI 3587*, 570-579, 2005

[12] B. Klimt and Y. Yang, The Enron Corpus: A New Dataset for Email Classification Research, Language Technologies Institute, Carnegie Mellon University, 2004

[13] B. Pang, L. Lee, V. Vaithyanthan, Thumbs up? Sentiment Classification using Machine Learning Techniques, *Proc. of EMNLP 2002*.

[14] F. Smadja, Retrieving collocations from text: Xtract, Computational Linguistics 19(1), Oxford University Press, 2003.

[15] S. Jabbari, B. Allison, D. Guthrie, L. Guthrie, Towards the Orwellian Nightmare: Separation of Business and Personal Emails. *Proc. of ACL 2006.*