COZO - CONTENT ZONING FOR SPAM EMAILS

Claudine Brucks, Cynthia Wagner, Michael Hilker and Ralph Weires
University of Luxembourg, Campus Kirchberg, CSC, 6, Rue Richard Coudenhove-Kalergi, L-1359 Luxembourg

Keywords: Content Zoning, Spam Email Detection, Text Analysis and Statistics, Human-Computer Interaction.

Abstract: Spam is an increasing problem when using email as communication medium. Spam is detected and removed using spam filters. Furthermore, the spammers use more and more intelligent and complex techniques so that novel approaches are required to enhance existing spam filters. One promising technique is the Argumentative Zoning that classifies a text in different parts where each part has a meaning. In this paper, we want to use this technique in order to divide an email into different zones and evaluate whether the email is spam or not. We introduce the technique Content Zoning, the way we use it, the implementation, and our results.

1 INTRODUCTION

In the 21st century with the growing importance of electronic messaging, the problem of spam emails, also called junk emails, arises. Many solutions already exist and improvements for avoiding this kind of messages are common. The most used techniques are spam filters (Androutsopoulos et al., 2000; Wittel and Wu, 2004; Rigoutsos and Huynh, 2004) and IDS (Intrusion Detection Systems) (Roesch, 1999; Hilker and Schommer, 2006). IDS check each network packet whether they contain a certain pattern and if a packet contains such a pattern it is removed in order to secure a network against intrusions. In contrast, spam filters are specialised to detect spam. They use different techniques: white and black lists describing allowed and denied email addresses. Another approach is to define patterns used by spammers or to analyse the header and content of the email in order to evaluate whether it is spam or not. In addition, language filters remove all email written in a foreign language and user-defined rules that remove spam emails, too.

Unfortunately, the spammers continuously introduce novel techniques in order to fail spam filters and to hide spam emails. For example, spam emails can contain text hidden between complex HTML tags or replace characters, e.g. 0 through o, so that it is hard for normal spam filters to detect it. Another example is the recent trend in spam emails for using scrambled words: this means words recognizable by the human brain but not by normal spam filters, e.g. “online” becomes “onilne”. In the permanent fight against spam emails, the need in finding new ways for spam email detection grows permanently.

In this article, we want to analyse whether Content Zoning can increase the performance of spam filters. In Content Zoning, an email is divided into different zones and each zone has a meaning, e.g. a price-zone describes the price with currency of an offered product and the offer-zone describes the product. We will introduce the Content Zoning, implement a program for Content Zoning, and perform tests on two types of spam emails: financial and pharmaceutical spam. Thereafter, we conclude the paper with some of our next steps.

2 RELATED WORK

Major inspirations for this work can be found in (Teufel, 1999; Feltrim et al., 2005), which describe a technique called “Argumentative Zoning”. In this work, zones are used to describe scientific documents. The aim of this new approach is to provide an intelli-

Brucks C., Wagner C., Hilker M. and Weires R. (2007).
COZO - CONTENT ZONING FOR SPAM EMAILS.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Internet Technology*, pages 23-27
DOI: 10.5220/0001263400230027

gent library search tool with generation of summaries on scientific articles and the detection of intellectual ownership of documents, e.g. for a better document management or for plagiarism detection.

We apply similar principles to the analysis of spam emails, as means to separate spam from non-spam mails. A part of this work can also be seen in (Brucks and Wagner, 2006), which deals with the manual analysis of up-to-date spam emails to check whether they contain significant patterns in the subject and content.

In this paper, we focus on a tool for analysis of emails by their content structure to find out about common patterns in spam emails. By doing this, we want to gain information that can be used to realise automatic recognition of zones for further work. Automatically derived information from spam emails can then be used to help classifying spam emails and separating them from normal emails.

The aim of the Content Zoning technique is to provide a useful function or add-on to normal spam filters for detecting patterns in the content of spam emails and for analyzing if spam emails have characteristic content structures. By that way, spam filters can perform better results in spam detection, because the technical aspect is extended by content analysis.

3 CONTENT ZONING

Before going into detail about applying Content Zoning to spam emails, we first give some general information in this section about what Content Zoning actually is.

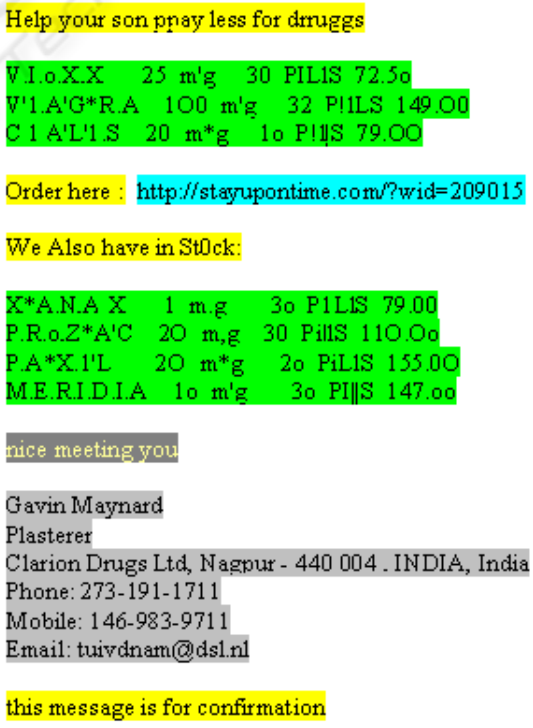
Generally, Content Zoning means to divide a given document into different parts. The separation is done according to the given structure of the documents (e.g. different paragraphs) and especially because of the semantic meaning of the document parts. In the end, the zones are supposed to represent the various semantic units of the document.

Besides the plain division of a document in zones, additional analysis can also be performed to gain further information about the separate zones. Rather simple to extract information could be statistics about the size and layout of zones but a more sophisticated analysis of their text content is also possible. The latter can lead to an extraction of the semantic content and purpose of a zone. Such kind of information can be used for various purposes, such as comparing documents to each other (regarding their analaysed zone structure and content). In the following, we refer to such information about the zones as zone variables.

4 APPLYING CONTENT ZONING TO SPAM EMAIL

By using Content Zoning, the email text can be divided into different regions, and it can be seen if there exists redundancy in the structure and in the content of spam emails. Email analyses have given reason to the assumption that spam emails contain the same structure or zone ordering, so they are in general more similar than other emails with a completely different structure.

The test set of spam emails that has been applied to the process of “zoning” mainly belongs to two different categories: spam emails of pharmaceutical domain (i.e. drug offerings), and spam emails from the financial domain (more precisely, stock spam emails). Financial emails typically seem to be defined as emails with large content, where the offer and company have a very detailed description. Pharmaceutical emails on the other hand can be described as emails with short content, direct offers and only a brief description. By analysing the characteristics of financial and pharmaceutical spam emails, it became obvious that emails of pharmaceutical domain generally had a lower amount of zones than financial spam emails (see also (Brucks and Wagner, 2006)).


Help your son ppay less for drruggs

V.I.o.X.X 25 m'g 30 PIL1S 72.5o
V'1.A'G*R.A 1O0 m'g 32 P!1LS 149.O0
C 1 A'L'1.S 20 m*g 1o P!‖S 79.OO

Order here : http://stayupontime.com/?wid=209015

We Also have in St0ck:

X*A.N.A X 1 m.g 3o P1L1S 79.00
P.R.o.Z*A'C 2O m,g 30 Pi!lS 11O.Oo
P.A*X.1'L 2O m*g 2o PiL1S 155.0O
M.E.R.I.D.I.A 1o m'g 3o PI‖S 147.oo

nice meeting you

Gavin Maynard
Plasterer
Clarion Drugs Ltd, Nagpur - 440 004 . INDIA, India
Phone: 273-191-1711
Mobile: 146-983-9711
Email: tuivdnam@dsl.nl

this message is for confirmation

Figure 1: Example of a zoned pharmaceutical email.

In the following, we give a more detailed description of how the zones for these two types of spam emails are defined and which zone variables are used. Figure 1 visualises a pharmaceutical email that is zoned.

4.1 Definition of a Zone

A zone can be defined as a region in a document that is allocated to a specific information in that document. For example, when having a text document with a price offer, this price offer can be annotated by a tag; a logical name for that zone could be "price offer". Of course, when having a huge document, zones can reoccur multiple times. The figure 1 illustrates an example of a pharmaceutical spam email in order to show what can be considered as a zone; each zone is represented by a colour.

The analysis on content and structure of pharmaceutical and financial spam emails showed a kind of redundancy in content rubrics and structures. These observations have given the main idea to the following concept: the most observed redundant rubrics in the spam email data set have been considered as the default zones.

Not all text of the email can be attributed to a zone, because often irrelevant text is included in spam emails. When having that case, no zone is attributed to it. After having defined a set of zones, these can be used for describing the content of a spam email.

Table 1 shows the most important zones we defined for both domains, the financial and pharmaceutical domain (but this does not mean that they have an equal number of zones, as already stated above):

Table 1: Most occurring zones.

Information	Offer
Additional Information	Product Description
Price	Expected Price
Name of company	Company Description
Testimony	Address
Mail Signature	Date
Symbol	Greetings
Forward Looking Stmt.	News
Stock Volume	Link

The annotation tags have logical names, so that mostly no supplementary information is necessary for understanding what the zone is about. Some examples are shown below to illustrate how a zone is defined.

- *Information*:
  The introductory information zone informs the reader on what the email is about
- *Offer*:
  The offer zone contains the offer itself, a promised product with its specific information
- *Price*:
  In the price zone, the actual price of a promised product is found (this zone can be considered as a sub-zone of "Offer")
- *Link*:
  This zone contains web-pages indicated in spam emails

Defining the zones alone is not sufficient for effectively realising Content Zoning, because they only indicate a position of an information in a text, but do not give information on the content of the text. For this, "zone variables" are introduced.

4.2 Definition of Zone Variables

A zone variable is defined as a parameter that describes specific information of a zone. By describing the zones with zone variables, more structured information can be extracted of the zone content. The zone variables are one of the most important factors in Content Zoning, because they contribute to the statistical evaluation and finally to the detection of similarities in spam emails. In this work, two kinds of variables have been defined:

Zone independent variables are defined as parameters which can be applied to every zone. Examples for zone independent variables are:

- position of the zone in the complete email
- length of the zone, expressed in number of characters
- number of words contained in the zone
- most occurring word (not considering stopwords)

Zone dependent variables on the other hand are defined as parameters, which are specific for a certain type of zone and hence can only be applied to zones of that type. Mostly, this kind of variables represents semantic parameters. The following list shows some examples for zone dependent variables:

- top-level domain for links (e.g. .com, .net, .lu)
- telephone numbers for the address zones
- value for price zones
- currency for price zones (€, £, $, etc.)

5 IMPLEMENTATION

The implemented system - CoZo - integrates a graphical user interface (GUI) which allows users to make their Content Zoning on emails. Most commonly used email types (HTML, EML, etc.) are supported by CoZo. Information that is not relevant for Content Zoning like email header or HTML tags can be automatically removed on user-demand, so that principally only the content part remains. We implemented CoZo in Delphi and Perl. The user can easily zone the content and the results are stored in XML files for further analysis.

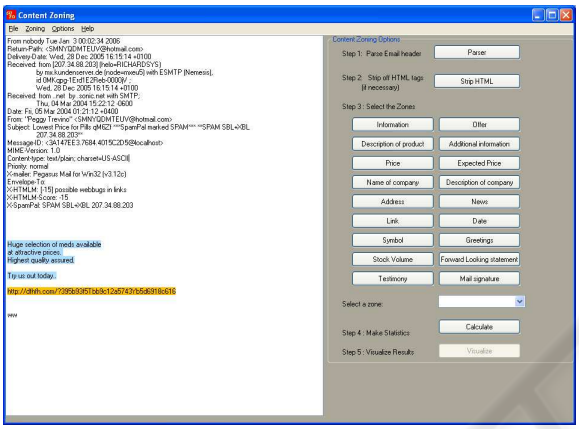


Figure 2: Screenshot of the CoZo Interface.

The figure 2 shows a screenshot of the CoZo application, where an example of selecting the zones is given. After charging the email onto the GUI, the different zones can be selected. To facilitate the zoning, 18 predefined zones have been implemented. However, if there is no convenient zone for a certain email region, the user has the possibility to create his specific zones or to leave that region unzoned. Each zone is visualised using a colour and CoZo calculates for each zone variables, e.g. the position of the zone in the email, the density of the zone, and the amount of characters in the zone. The output - definition of zones, calculated variables, and a picture of the zoned email with colours - is also stored in files for further analysis, e.g. picture matching.

The focus in the implementation lies on an application to test the idea of Content Zoning. Therefore, we have decided to use a manual zoning system in order to quickly zone many emails and compare these (proof-of-concept implementation). Furthermore, the application provides the output that we analyse in order to evaluate the Content Zoning approach for spam detection.

In order to use Content Zoning for real spam detection, it should be implemented as a part in a spam filter. It should zone the email automatically, calculate required variables, and evaluate the output to check whether the email is spam or not. Consequently, Content Zoning would be a technique for the evaluation of emails.

6 RESULTS

After finishing the implementation of CoZo, we tested the approach. We used a data set of emails captured from January 2005 to September 2005. The data set contains emails from various domains (e.g. credit offers, movie downloads, porn emails), but only emails from the pharmaceutical and financial domain were used for the tests of CoZo. For more details on the data set, we refer to the document (Brucks and Wagner, 2006).

For the tests, no supplementary zones had to be generated. This means that the 18 predefined zones mostly covered the email content. The zones of the emails from the data set have been coloured as in figure 1. After having realized the zoning, the statistics are calculated and stored in a XML-file:

```
...
<LengthOfZone>399</LengthOfZone>
<NumberOfSentences>5</NumberOfSentences>
<NumberOfWords>80</NumberOfWords>
...
```

We have obtained different results:

- It was possible to zone the emails and CoZo calculates the output without any problems.
- Pharmaceutical spam has mostly the same structure of zones. This means that after zoning, the order and the size of the zones are similar. Furthermore, the calculated variables of the zones are similar.
- Financial spam has a different architecture of zones but always contains some significant zones like e.g. offer, information, and stock details.
- Financial, pharmaceutical, and normal advertising emails have a different order and architecture of zones.

With these results, we can say that Content Zoning is a technique helping to classify emails as spam or not.

After zoning several emails, we obtained lots of pictures of zoned emails. These pictures are sorted by the average colour using ImageSorter[1]. The results are that the emails are clustered in normal non-spam advertising mails, pharmaceutical, and financial spam

[1] $http://mmk.f4.fhtw-berlin.de/?page\_id=40$

emails. Consequently, zoning of emails and automated ordering of the resulting pictures maybe helps identifying spam emails.

7 FUTURE WORK AND CONCLUSION

This project and the implementation of CoZo are not finished. With this status of implementation, we tested the approach of Content Zoning. One of our next challenges is to adapt the application so that it will be possible to have a semi-automatically recognition of the zones. Semi-automatically means in this case that most obvious zones are detected automatically, e.g. price, currency symbol, etc. This would simplify the zoning for the user, but the more complex zoning should still be realised by the user for getting more precise results. The complexity of this function is to detect the zones correctly even when having no legal formats. An example is “price” zone, where numbers are represented by characters, e.g. “0” replaced by “o”.

The next step is to implement an automated zoning so that the system calculates the zones and important variables without any input from users. Thereafter, the evaluation of the zoned emails must be automated and the system can be added to an existing spam filter.

There are different possibilities for classifying an email into the various types of spam emails using Content Zoning. One approach would be to define a similarity metric according to the calculated zone statistics (e.g. zone ordering, zone sizes, etc.). Another approach is to use picture matching to realise a comparison between two emails. To do this, the emails are e.g. coloured in order to denote the different zones. The actual comparison is of course not limited to basic (exact) picture matching but can as well include more sophisticated image analysis techniques. By doing this, we can perform a comparison of a new email to our existing spam email types (financial and pharmaceutical here) to decide whether or not the email is spam at all.

To conclude the article, we can say that Content Zoning is a promising technique for spam detection and supports existing spam filters with additional information. The results show that the picture of the zoned email and the calculated variables contain indicators whether an email is spam or not.

ACKNOWLEDGEMENTS

This project is a part of the TRIAS (TRIAS, 2005) project of the Computer Science and Communication research unit from the University of Luxembourg. We want to thank the staff of the MINE-team for their support and advise.

REFERENCES

Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G., and Spyropoulos, C. (2000). An evaluation of naive bayesian anti-spam filtering.

Brucks, C. and Wagner, C. (2006). Spam analysis for network protection. TFE Thesis - University of Luxembourg.

Feltrim, V., Teufel, S., Nunes, G. G., and Alusio, S. (2005). *Argumentative Zoning applied to Critiquing Novices' Scientific Abstracts. In Computing Attitude and Affect in Text: Theory and Applications*, pages 233–245. Springer, Dordrecht, The Netherlands.

Hilker, M. and Schommer, C. (2006). SANA security analysis in internet traffic through artificial immune systems. *Proceedings of the Trustworthy Software Workshop Saarbruecken, Germany*.

Rigoutsos, I. and Huynh, T. (2004). Chung-kwei: a pattern-discovery-based system for the automatic identification of unsolicited e-mail messages (spam). In *Proc. of the Conference on Email and Anti-Spam (CEAS)*.

Roesch, M. (1999). SNORT - lightweight intrusion detection for networks. *LISA*, 13:229–238.

Teufel, S. (1999). Argumentative zoning: Information extraction from scientific text. Phd Thesis, University of Edinburgh, England.

TRIAS (2005). Logic of trust and reliability of information agents in science. CSC, University of Luxembourg, Project description, Link: http://wiki.uni.lu/mine/TRIAS.html.

Wittel, G. and Wu, S. (2004). On attacking statistical spam filters. In *Proc. of the Conference on Email and Anti-Spam (CEAS)*.