Automated Ticketing of Email for ISP Customer Care

J. Buford[a], D. Bowie[ß], X. Huang[a] and M. Mady[a]

[a]*Verizon, 40 Sylvan Road, Waltham, MA 02451, USA*

E-Mail : john.buford@gte.com, mailto:xiaolan.huang@gte.com, mary.mady@gte.com

[ß]*Genuity, 3 Van de Graaff Drive, PO Box 3073, Burlington, MA 01803, USA*

E-Mail: dbowie@genuity.net

Email is emerging as an important channel for ISP customer care, and manual processing is labor intensive. A new system for automating the processing of large volumes of email received by an ISP customer care center is described. The system automatically determines the embedded email structure, extracts key email fields needed for aggregation and ticketing, performs keyword analysis, and invokes system utilities as needed. The production system processes up to 30,000 email requests per month, and has processed over 300,000 complaints to date.

Keywords: Customer Care, Service Management, Automated Email Processing, Spam, UCE

1. Introduction

Email is emerging as an important channel for ISP customer care. For many users it is more convenient than using a voice response system, and allows the user to provide details of the problem more efficiently than via phone. The interaction is asynchronous, so the ISP can prioritize and categorize according to its business process. However, manual processing of email can be expensive, particularly as volume grows. ISPs are experiencing significant growth in email complaints, particularly those related to UCE (unsolicited commercial email). The reason for this is due to the increase in UCE volume and the availability of tools that streamline the reporting of spam by users. The UCE situation leading to customer care is shown in Fig. 1. The source of UCE (spammers) have the practice of obtaining a number of transient email accounts, bulk mailing from these accounts (or forwarding via open mail relays) (Fig. 1, (A)), and then moving on to other accounts or ISPs when these accounts are closed. Spammers frequently manipulate email headers in order to make tracking difficult and to camouflage the nature of the bulk mail [1].

ISPs, systems administrators and users can employ filters to limit the exposure (Fig. 1, (B)). Filtering is not universally deployed and can't stop all spam. Users can respond to spam by complaining to the ISP or administrator of the host or network where the spam originated. There are a number of tools (e.g., SpamCop [2], Sam Spade, ORBS [3]) that simplify this (Fig. 1, (C)). The result is that the associated ISP can receive a large volume of email complaints which must be processed and tracked. The ISP must distinguish between an individual complaint (such as generated by SpamCop) and a spam incident, which is a set of complaints from the same host by the same spammer. This distinction is important, since many complaints relate to the same incident, and corrective actions taken by the ISP are at the incident level.

Unfortunately it is difficult to correctly group a set of complaints into the same incident. This is because the spam is manipulated during the bulk mailing to hide this, and complaints are presented in different forms by the complainants because of the different tools used. Failure to group complaints into the correct incidents can lead to incidents not be identified (and thus permitted to continue) or multiple identical incidents being ticketed and duplicately processed by Level 2 and Level 3 support staff.

O. Festor and A. Pras (Eds.) : 12th International Worshop on Distributed Systems: Operations and Management DSOM'2001 Nancy France, October 15-17, 2001.

Manual processing by the customer care center is labor intensive. Alternatively, some amount of automatic processing for routing and ticketing can be done by structural analysis, parsing, and keyword matching. Specific information (e.g., URL's, IP addresses, etc.) can be gathered to facilitate subsequent diagnostics. System utilities can also be automatically invoked to save time for level 2 and 3 customer-care specialists. The results of the utilities can be stored in the ticket with the complaint. Further automation could be obtained by the use of natural language processing of incoming complaints, but there is limited use of NLP for customer care email processing today.

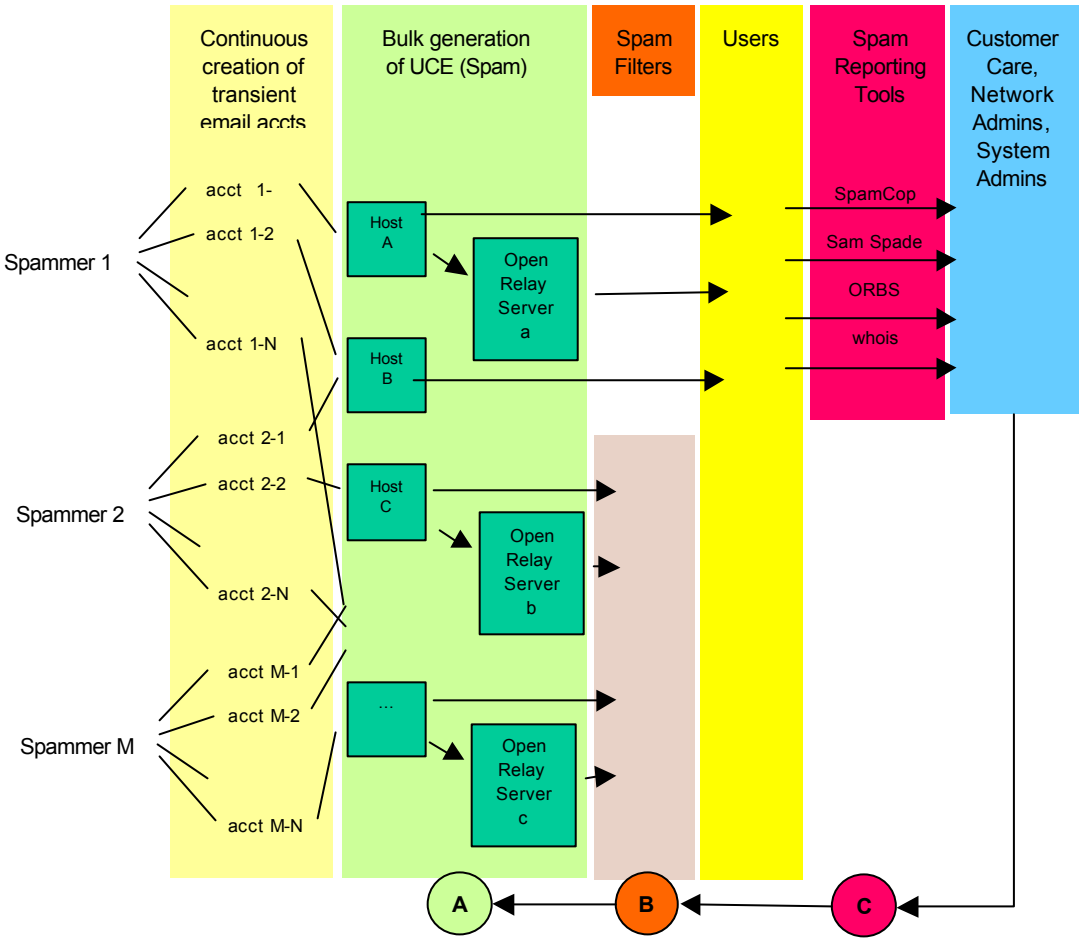

Figure 1 The customer care role in processing spam complaints involves correctly identifying which networks, hosts, and/or mail relays are involved (box A), identifying the spam incident that each complaint corresponds to, and managing the resolution process.

The Customer Support Center (CSC) for Genuity receives about 30,000 email complaints per month. The majority is forwarded complaints from individuals on the internet who have received UCE spam from sources that may be in its network. Another significant portion is from customers who have detected a security issue. Genuity validates each incident and takes corrective action if the problem is within its network. After some experience with manual processing of these complaints [4,5], a system SpamCheck™ was developed to replace the manual process [6] and is described in this paper.

The deployment context is shown in Fig. 2. Incoming email is divided into two categories. UCE complaints and security issues are processed by SpamCheck for ticketing. General support questions are currently ticketed manually. Tickets are processed by category or account by customer care specialists.

The SpamCheck system is concerned with categorizing the complaint so that it can be properly ticketed and remedies taken. There is an extensive set of business rules for this processing. Each email is multiple embedded or forwarded emails. Special processing beyond that normally needed for internet email is needed, and this processing is not done by any commercial email tools or spam analyzers today. There are many structural variations in the incoming email, and a portion of the embedded spam has frequently been manipulated by the spammer to make analysis and therefore tracing difficult. SpamCheck also performs keyword analysis.

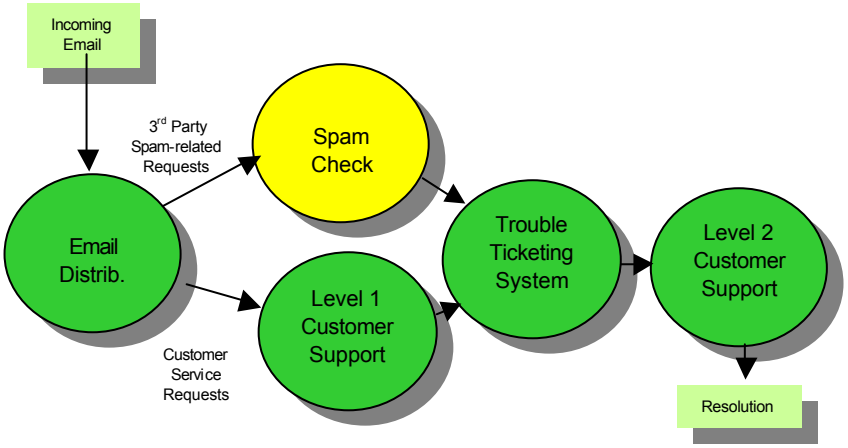

Figure 2 Email sent to the ISP's abuse account is forwarded to the SpamCheck system for automatic processing if it relates to UCE or security, or to the Level 1 customer support queue for manual processing.

As discussed later, automation of email complaint processing is complicated by a number of factors including: 1) complaints concatenate multiple text paragraphs and RFC 822 headers [7], 2) some spammers camouflage their messages to circumvent detection or correlation, 3) spam categorization depends in part on analysis of the content of the spam message. The system is designed to work correctly on 90% of the cases, and provides facilities for manual review of each instance.

A high level view of the system is shown in Fig. 3. On the left is a email forwarded by the ISP automatic email tracking system. It contains an embedded email from the source of the request. Typically the request email contains another email which would be the instance of the unwanted spam that the request concerns. After parsing the text both structurally and syntactically, key fields are extracted by the system and stored in a DBMS. A web-based reporting system (Figure 3, right) includes tools for rapid ticketing of groups of related requests and the ability to query the DBMS.

Section 2 describes the business process context in which SpamCheck is deployed. Section 3 describes the system processing. Section 4 gives an example. Section 5 describes the architecture and implementation, the paper ends with sections on related work and a summary.

2. Business Process

Genuity Customer Service Center (CSC) employs full-time staff to analyze and respond to a large volume of email that involve Genuity's customers and/or network. Like other Tier 1 ISPs, a frequent issue is the unwanted delivery of UCE which may have been sent via the ISP's network or hosted customers'

systems. Each incoming email requires review, logging, trouble ticket review, correlation with other complaints, response to the complainant, categorization, validation, distribution to appropriate staff, and resolution. Depending on the size of the ISP, routing between Level 1, 2, and 3 teams may be performed. Prior to the introduction of the SpamCheck system CSC had developed an effective manual process for handling email complaints, but a tool for automating a significant portion of the manual process was needed due to expected increase in volume and certain limitations of the manual process.

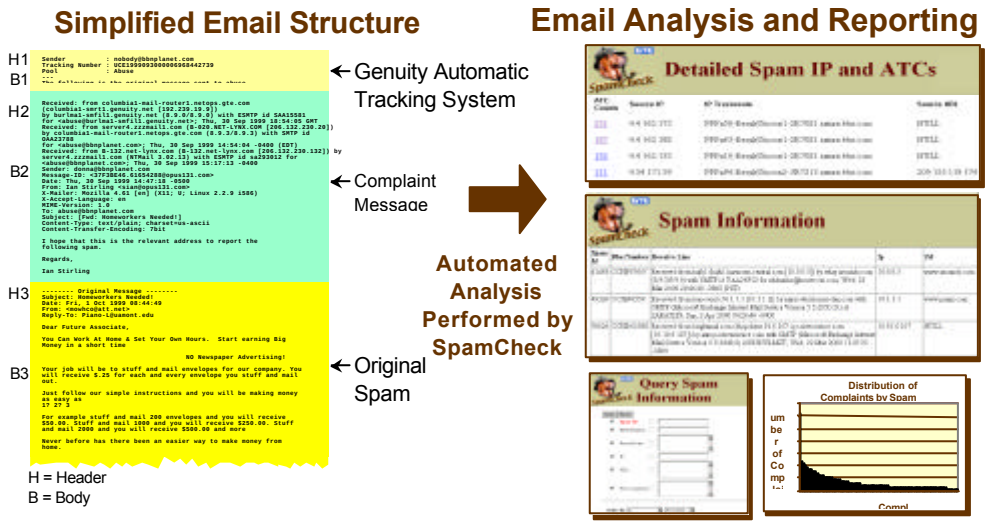

Figure 3 SpamCheck receives email service complaints, identifies embedded email structure, checks for header manipulation, and extracts key fields for analysis and reporting

Figure 4 shows the top level flow of the ISP customer care process for handling email requests.

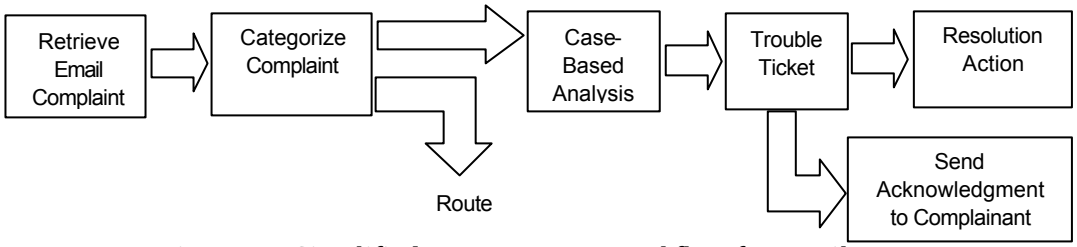

Figure 4 Simplified customer care workflow for email requests

The SpamCheck system is concerned with automated categorization and analysis of incoming requests, and grouping into ticketed incidents. Once the incident is ticketed and entered into the trouble ticketing system, then level 2 and level 3 support can be engaged. SpamCheck does gather diagnostic information that is useful for the level 2 and 3 support process. SpamCheck uses customer database in the trouble ticketing system to associate the request with specific hosting or network customers. Every request that is not related to the ISPs network or customers can be ignored, since the ISP has no direct control over the facilities used by spammer. All other requests are individually acknowledged via email. The SpamCheck categorization is used to determine the appropriate form of the acknowledgement.

Not all incoming requests relate to UCE. Another important set is security issues, such as attempts to penetrate a customer's host. SpamCheck uses keyword analysis to identify security related requests. The ticketing, routing, and resolution steps are different for the security cases.

The automatic processing system tracks related incidents over time to insure that the same problem is not multiply ticketed.

3. System Description

3.1. Overview

Internet email corresponding to IETF RFC 822 has a standard header and body separated by a blank line, the header has a list of field names and field values, and the body can be MIME encoded (IETF RFC 2045 [8]), HTML formatted, text formatted, or some combination of these. The RFC 822 header is specified by BNF syntax. When an email body includes one or more emails within its body, extracting the embedded email is difficult. The header for the embedded email is not easily distinguished from the enclosing body text because there is no standard syntax by which embedded emails are placed in the body of another email.

In help desk applications for ISPs, multiple levels of email embedded are typical. The SpamCheck system automatically disambiguates the structure of multi-embedded emails so that each email's header and body can be distinguished. This disambiguation is done by using heuristics such as searching for patterns that are most likely to correspond to the beginning of a header, or by removing delimiters and/or tags that are likely to make discovery of the header difficult. These heuristics were arrived at by processing many thousands of emails received by Genuity's help desk.

In addition to the difficulty in disambiguating the structure of the embedded emails, sometimes email headers are intentionally corrupted by spam sources in order to make it difficult to track them. The header might omit some required fields, it could have fake fields, or fields with incorrect values. These issues are well known in the spam fighting community, and we have incorporated these techniques in our overall process so that forwarded email complaints about spam can be processed automatically. Other complications due to the embedding of email include: the body may have special characters, possible forwarded email patterns (e.g., " >>" signs), various html tags(e.g., <html>, <body> etc.), hidden or extra blank lines, extra spaces that are generated by commercial spam tools.

When an ISP's help desk analyzes electronic mail customer requests regarding UCEs and other service disruptions, the SpamCheck parsing technique permits an automated processing system to analyze incoming email so that the ISP can ticket and resolve each complaint. The ISP typically maintains a tracking system in order to insure that each validated complaint is properly handled. The processing is complicated by several conditions, including: 1) analysis of the complaint may involve conditions that change over time and may no longer exist, 2) many complaints may deal with the same problem, but may be difficult to associate with the other complaints due to different complaint formats, 3) the customer network is changing during the time period that the complaint covers.

3.2. Structure of Email Complaints

Some examples of spam complaint email structure are shown in Fig 5. A typical case is a set of three concatenated emails conforming to IETF RFC 822. The original spam message is the inner most message (H3,B3) which may have been sourced from or via a host on the ISP’s network. This message is forwarded by an arbitrary spam recipient to the abuse complaint address (e.g., abuse@some_isp.net) in message (H2,B2). The automatic tracking system embeds each such incoming complaint in a third email (H1,B1) with some GTEI specific tracking numbers in B1. The embedded messages may or may not be encoded as mime-types (IETF RFC 2045).

The spam complainant may use a spam analysis utility such as SpamCop which creates an analysis email which includes portions of the original spam. Message (H2,B2) will come from SpamCop and B2 contains H3 (the header of the original spam), an exploded view of H3, and extracts from B3 (the body of the original spam).

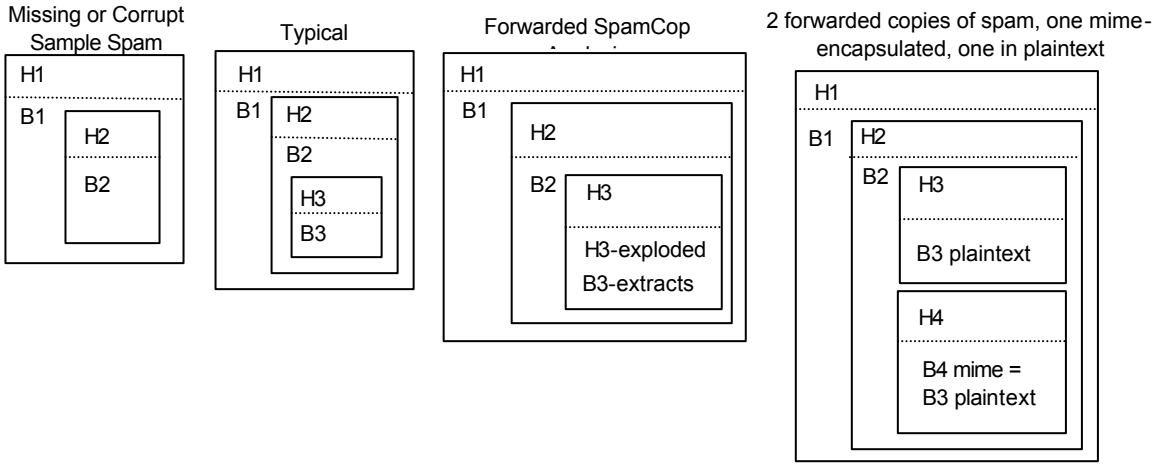

Figure 5 Examples of embedded email structure

Another possibility is that the complainant may omit the sample spam or include it in a corrupted fashion. Or the spam header itself may have been hacked to make it difficult to trace. In this case, the message (H1,B1) is the internal tracking labeling, and message (H2,B2) is the complainant.

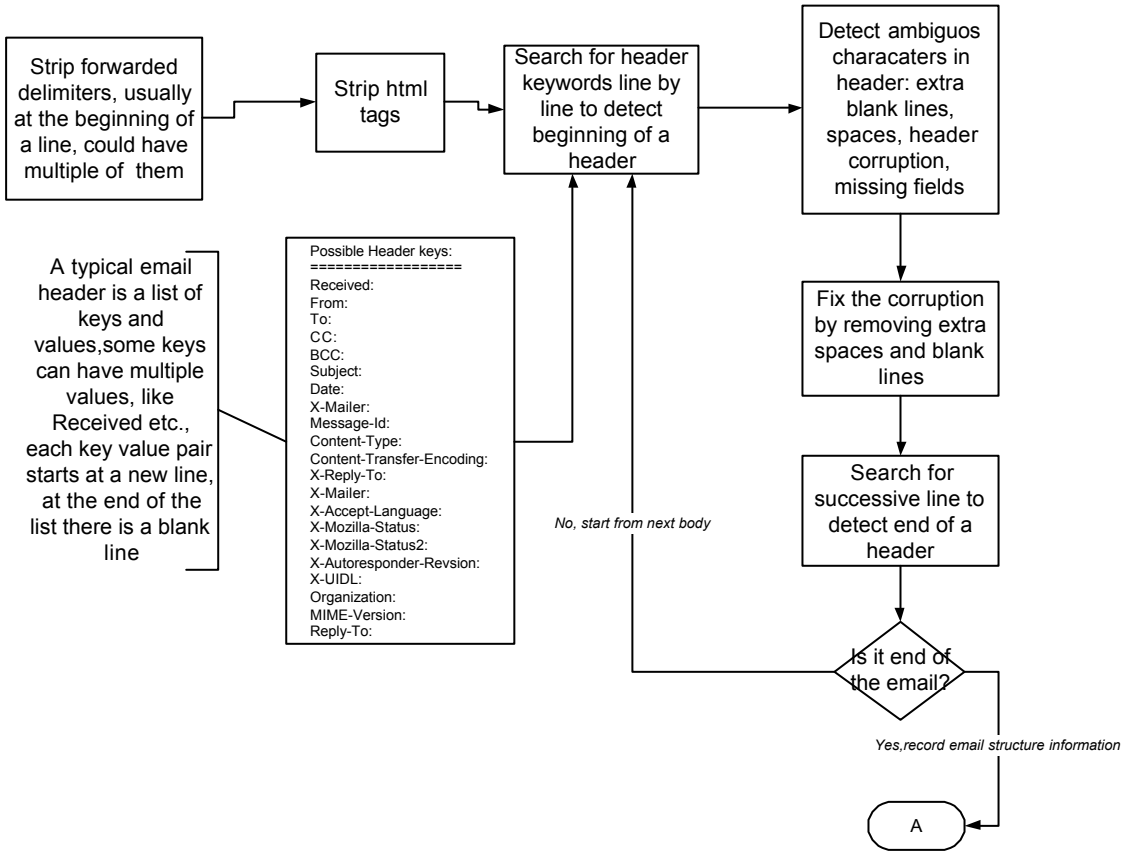

Figure 6 Top level processing flow for identifying structure of email and key fields

A fourth case is where two copies of the spam are forwarded, one in text/plain and the other in text/html, the latter being encapsulated as mime object.

Another case is where a spam is sourced from a non-hosted machine, but, because of configuration problem on an ISP-hosted machine, third party spam can be forwarded through the ISP-hosted machine.

There are variations of these cases which complicate parsing. For example, the forwarding of the embedded messages may have introduced blank lines between the header lines. Since a blank line is a separator between the body and header of RFC 822 email, this makes it difficult to isolate the body and header. Some of the embedded email messages may be encapsulated as mime-types, which requires mime identification and extraction.

A decision tree categorizing the different cases is used to drive the structural classification of incoming email. The top level flow of the processing of raw email is show in Fig. 6.

3.3. Keyword Categorization

In preparing the request for ticketing, it is useful to classify the request or problem. SpamCheck uses keyword analysis over the body of the embedded email to categorize the spam. This category and the first line of the spam are stored as part of the trouble ticket. The trouble ticketing system can then provide reports on the breakdown of request types. This is useful for spotting trends in the customer care traffic. Table 1 shows the current set of categories.

Table 1 Keyword categorization

CATEGORY	DESCRIPTION
Pornographic	All types of XXX ads
Stock	Pump and dump schemes
Vacation	Travel, timeshare and the like
MMF	Make money fast/multi level marketing
Casino	Online or physical, sports betting schemes
Web ads	Other spam
Security	Threats and related

3.4. Level 2 and Level 3 Facilitation

Once the email structure is determined and the key fields are extracted, various utilities can be run automatically to provide additional data that may be useful to resolving the problem. Examples are:

- Traceroute, nslookup, whois, dig – for identifying the host, the ISP, and the administrative address
- Host clock validation – for catching timestamp spoofs
- Hex address conversion – for converting IP addresses from hex format to dot format

Since many incoming requests may relate to the same spam source, the system need only invoke these utilities once per request group.

4. Example Processing Sequence

Structural analysis of a sample email is detailed in Figure 7. The left bar segments are the different embedded emails. This example excludes MIME content, html formatted, and corrupted headers.

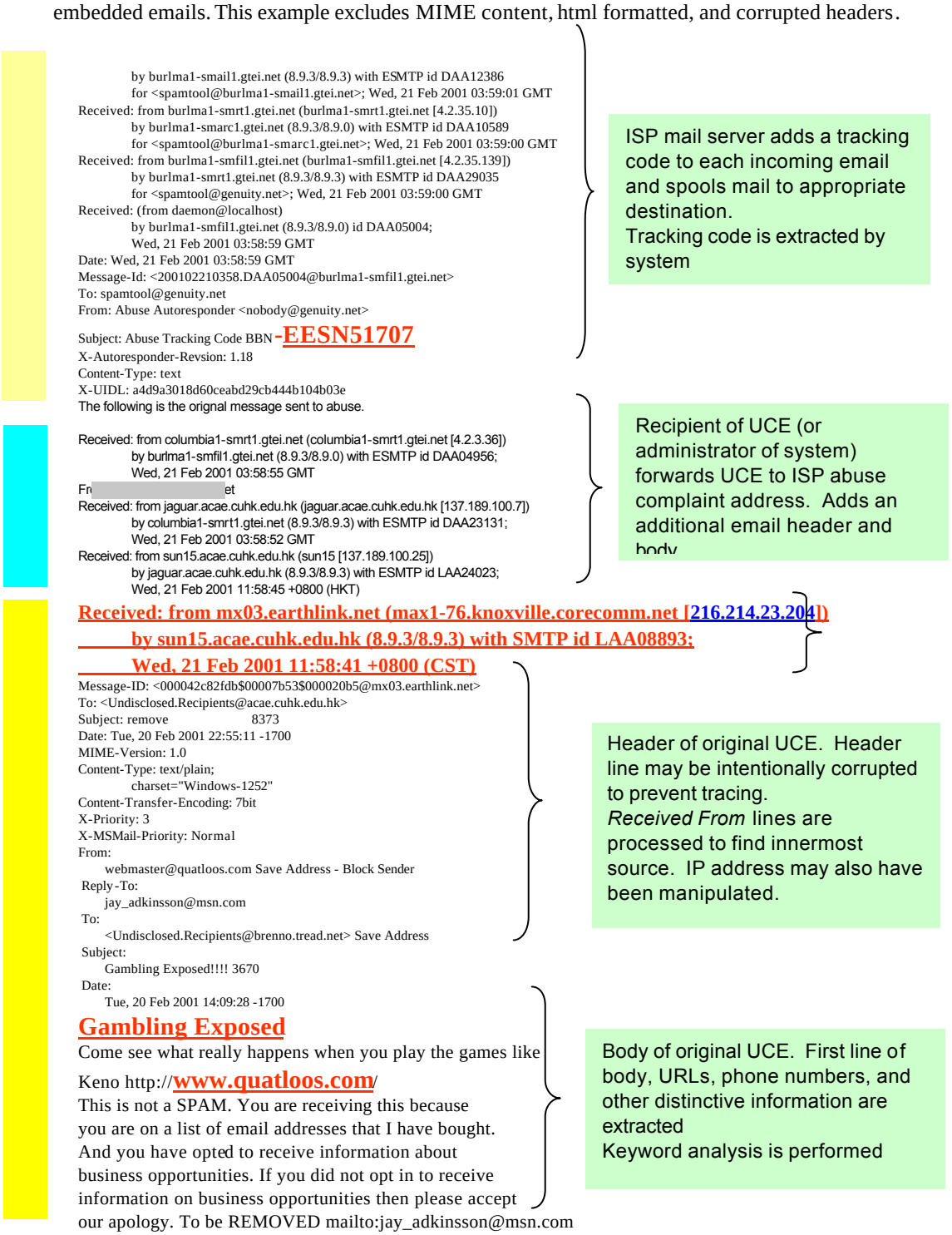

Figure 7 Simplified embedded complaint email

5. Implementation

5.1. Prototype

Because of the unique processing requirements and the thousands of cases to analyze, a prototype was created first in order to identify the cases and their frequency. Thousands of samples were processed by the prototype and analyzed. Histogram graphs were plotted showing categories and distributions. Many more special case instances were found than had previously been thought. However grouping opportunities were also verified, and the query system was designed around this after approval by CSC.

5.2. Architecture

The server-side architecture is shown in Fig. 8, the client architecture in Fig. 9. The web interface is implemented using Coldfusion, a product for server-side DBMS queries to generate web pages.

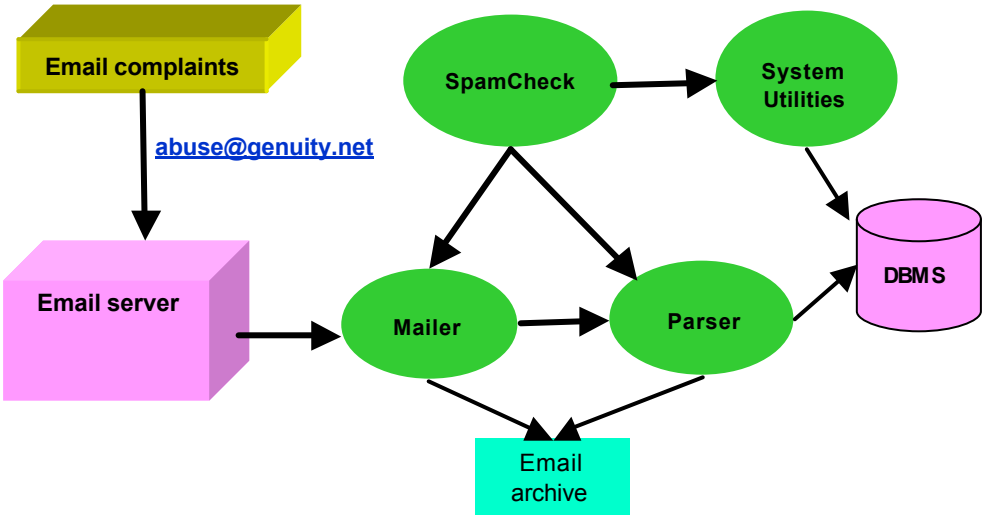

Figure 8 Backend system architecture

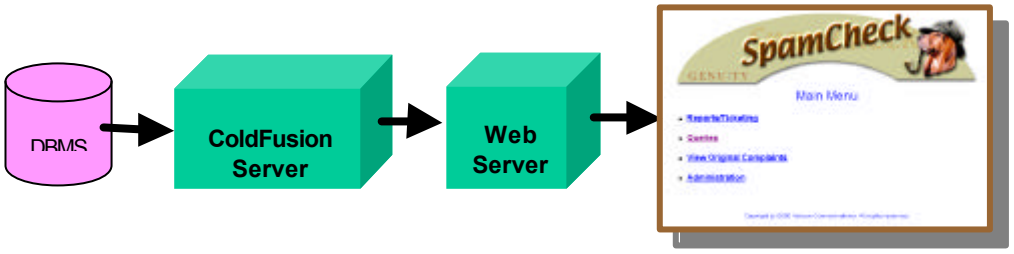

Figure 9 Front-end system architecture

5.3. Performance

Over 300,000 instances have been processed by the system and reviewed by CSC staff. The team monitors the system and has a copy of all incoming email being processed. 90% of incoming emails are successfully categorized. The system has been tested at throughput rates of 10x peak daily rates. Each incoming email is sequentially numbered by the CSC email distributor, and checking of missing or duplicate email is straightforward. Queries against 10,000 rows display in less than 20 sec on a dual 200MHz NT system (Table 2).

Table 2 Query response time vs. number of records

Number of Records	Query Response Time
8146	9 sec
14,362	12
20,578	15
39,226	22

5.4. Ticketing Statistics

Table 3 and Table 4 show measurements made by the system over different 30 day windows regarding the distribution of email by content category (compare with Table 1) and structure types (compare with Figure 5).

Table 3 Distribution of email content types in a 30-day period

Content Category	Count	%
Pornographic	12600	0.39
Stock	1496	0.05
Vacation	7341	0.23
MMF	603	0.02
Casino	7778	0.24
Web ads	1540	0.05
Security	1044	0.03

Table 4 Distribution of email structure types in a 30-day period

Structural Type	Count	%
Missing UCE attachment	31	0.001
Forwarded via SpamCop	36814	0.703
Simple embedded	6397	0.122
Plain Text + Mime	9066	0.173
Missing Header	76	0.001

6. Related Work

There are various tools and products for filtering spam before it reaches the end user. These tools are effective but there are still gaps that permit many spam sources to operate. After spam reaches the end users, there are tools such as SpamCop [2] that analyze the spam headers in order to the relevant ISP to contact. These tools do not address the handling of embedded email or the overall incidence identification and resolution process that the ISP help desk faces.

There are tools for parsing email and constructing mail processing applications such as JavaMail [10]. These tools do not support parsing of embedded email. Similarly applications for reading email also do not support parsing of embedded email.

7. Summary

The SpamCheck system has processed over 300,000 emails to date, is currently in use as a production system. 90% of incoming correctly formed email are handled correctly. The system provides a high degree of automation of a critical customer care function.

UCE tactics are continuously changing to counteract filters and reporting tools. Further improvement is possible in tracking trends in UCE, and generalizing the system to handle other categories of customer support email.

8. Acknowledgements

CSC team contributed feedback on each version of SpamCheck that led to its current instantiation. J. Doleac and M. Li contributed to the development of the prototype system. SpamCheck™ is a trademark of Verizon.

9. References

[1] Geoff Mulligan. *Removing the Spam--Email Processing and Filtering*, Addison-Wesley; Reading, MA, 1999.

[2] SpamCop. http://www.spamcop.com

[3] Open Relay Behavior-modification System (ORBS). http://www.orbs.org

[4] Current Processing Procedures (UCE Processing), Genuity, 9/23/99. Internal.

[5] Canned Messages, SPAM notifications to email to customers, Genuity, 9/23/99. Internal.

[6] Functional Specification for a SPAM Fingerprinting System, David Bowie, 6/15/98. Internal.

[7] IETF. RFC 822: Standard for the Format of ARPA Internet Text Messages

[8] IETF. RFC 2045: Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message

[9] Coalition Against Unsolicited Commercial E-mail (CAUCE) http://www.cauce.org/

[10] Sun Microsystems. JavaMail 1.2 API. http://java.sun.com/products/javamail/