

11

Archiving and Categorization of E-mail Systems

Ian H. MacGregor and Nicholas A. Wagner

Introduction

Imagine that a company called Custom E-Mail Corporation provides free e-mail to users of the Internet. They try to create appeal for their services by offering unique and fun domain names for their e-mail addresses, such as Internetjunkies.com, bears-fan.com, and darthvader.com. To make up for their costs, advertising within the web-based mail portal is utilized, along with advertising within e-mail messages their users send.

For about the last year now, Custom E-Mail has been exhausting their savings and has not been able to meet their expenses. A spokesman from a data analysis firm, Web Profilers, Inc., has been speaking with Custom E-Mail about purchasing the e-mail message archives of their users. The money that Web Profilers is offering is more than enough to pay off the debts of Custom E-Mail, upgrade their hardware, redesign their business model, and give everyone a raise. However, Custom E-Mail has learned that Web Profilers plans to use all the sent messages of the accounts to create profiles of the users. These profiles would contain the interests, habits, personal information, reoccurring topics, reoccurring subjects, and other contents of the users' sent mail. Web Profilers will then sell these profiles to whoever wishes to purchase them, such as government systems, marketing divisions, interested citizens, university research teams, and many more.

Custom E-Mail Corp is now faced with a huge dilemma regarding their users' rights. Many of Custom E-Mail's users have chosen this service as their main method of contacting people across the Internet. The data that their users send is sometimes career oriented, and many times involves their personal lives. Custom E-Mail has implemented a deficient privacy policy that states they will not disclose personal

information provided for the account and that the data stored on their servers belongs to them and is the property of Custom E-Mail. Web Profilers won't be labeling the portfolios by the account holder's name, but they will then own the personal data that was disclosed in the outgoing messages of the account holder.

What kind of responsibilities does Custom E-Mail have to its users? Is it legal to sell their data archives? What kind of threats to Custom E-Mail's account holders does a firm like Web Profilers present? What kind of ethical issues are present for each company?

This chapter will provide analysis of these questions and dilemmas, and prepare the reader to critically interpret the issues. The analysis will assume that the reader has knowledge of how e-mail systems work, including web-based e-mail communication systems, and POP3 and SMTP mail systems that require a program installed on the user's machine (such as Eudora or Outlook).

Archiving of E-mail Systems Defined

E-mail systems have become a staple in today's business culture, and are essential in ensuring productivity and efficiency. Many businesses, public schools, universities, and nonprofit organizations have made this form of communication just as important as the telephone or the fax machine, and in some cases, have replaced one or both altogether.

E-mail servers within organizations are usually programmed to retain sent and received messages within the organization's user account, in accordance with the disclosed statement of privacy or privacy policy. In some cases, actions and messages are backed up to comply with state laws and regulations such as those put in place by the Security and Exchange Commission (SEC Interpretation 20 May 2003). In many cases, the server's transactions are saved for a disclosed period of time, usually made clear in the provider's statement of privacy.

Who Archives Our E-mail?

Few people are aware of the fact that much of their e-mail is archived or backed up. This is usually done as a means to safeguard information in the event of a server crash or unexpected incident. Educational institutes, Internet service providers (ISPs), e-mail providers, and private and public sector companies are all organizations which store the e-mail they send to and from their customers and employees. Often, these organizations simply store e-mail for a fixed period of time with no intention for further analysis. Depending on factors such as financial status and new technology, certain organizations don't explicitly detail their intentions to analyze e-mail, and may find themselves in a position of ethical uncertainty when new opportunities and needs to start analyzing e-mail emerge.

Most companies are eager to store e-mail their employees send and receive. According to a survey conducted of 105 global corporations, 86% reported e-mail archiving as very or somewhat important. According to M2 Communications, only 37% actually had a policy about how to properly archive and secure this e-mail (5 September 2003). Companies understand the important role e-mail has assumed as a primary communications channel – with legally binding agreements,

sales contracts, etc. now being issued online – and the importance of being able to access this information at a later point in time.

Educational organizations have a rather interesting approach to storing e-mail. The University of Colorado System, for example, does not save students' messages, but occasionally backs up faculties' messages which can be turned into public records according to the University of Colorado's Use of Electronic Mail policy (Use of Electronic Mail 1997).

How Does Archiving Work?

E-mail archiving technology is a rapidly advancing science. For years, organizations were able to back-up e-mails using a traditional storage medium such as tape drives and hard disks. With the exponential rise in e-mail messages and the popularity of e-mail infiltrating every aspect of people's lives, the vast quantity of messages has started to overwhelm traditional storage methods. The ability to accurately file and retrieve important messages is simply too time consuming and expensive when such large volumes of messages must be filed. This has created a necessity for companies to explore new opportunities to help solve this data management issue. Now organizations have multiple options when deciding what method they want to pursue to archive e-mail.

A market has emerged that offers two systems for organizations wishing to archive e-mails more accurately. This includes sorting onsite and outsourcing the archiving. The product then can be programmed to sort e-mail in a way most beneficial to that organization's needs. This option is entirely in-house, offering organizations the most control over how their e-mails are stored, but with the additional expense of having to staff and maintain the equipment. Outsourcing the responsibility to a company that specializes in e-mail archiving is also gaining popularity. The outsourcing company charges a fee, usually more expensive in the long-run than an in-house system, but for many organizations, convenience makes this option most attractive.

Why Organizations Need Archiving

The need for e-mail archiving differs among organizations. As mentioned previously, companies are increasingly using e-mail as a formal form of correspondence. Invoices, legally binding agreements, and important discussions are all now communicated in e-mails. Government agencies like the SEC require these e-mails to be stored in original format for future access in the event of an audit or federal investigation (SEC Interpretation 20 May 2003).

Although there are no laws to prohibit the reading of an employee's e-mail and monitoring their actions on the computer, organizations establish their own internal 'laws,' or codes of conduct. This code of conduct is the privacy policy created within the organization. Courts have favored organizations over their employees when their policy openly disclosed the active monitoring of electronic communications. On the flipside, employees have won court cases due to some companies' lack of descriptiveness or absence of a policy altogether. In regard to e-mail, privacy policies should actively disclose their intentions to use retained data. The issue still stands though. If a

firm claims that e-mails sent through the server are company property, then employees' personal information that is disclosed is at the discretion of the company. In the case of Custom E-Mail Corporation, if they choose to sell their archives, then that information belongs to Web Profilers.

When Will Such Practices Start Affecting Users?

E-mail archiving is an issue that affects users in business today. It will continue to be a hot topic of discussion until new laws and universal policies surface. This is much easier said than done. In the future, archiving may be more prevalent, and new forms of personally identifiable technologies will be developed.

Persona Creation Through E-mail Categorization

What is Persona Creation through E-mail Categorization? The process of building a personality depicting one's personal traits, habits, and preferences is known as persona creation. Criminologists have done this for years as a way to identify criminals. Just recently, with the preponderance of e-mail archiving occurring in every aspect of e-mail services, the concept of persona creation will most likely be used in e-mail as well.

When many years worth of every type of correspondence is archived in a file - personal details, friends and family, likes and dislikes - all of which is stored in digital format, this information can, if properly analyzed, yield a very accurate representation of a person. This e-mail information is increasingly becoming available as more and more people use e-mail as a primary means of communication. The real concern is the lack of policies in place to safeguard this information; this is where real concerns of privacy exist.

As mentioned previously in this chapter, privacy policies are usually vague and difficult to understand, and this causes many people to underestimate, or be wholly unaware of, privacy rights with regard to their e-mail. In many cases, e-mail is not owned by the individual, but by the company. This e-mail is essentially the property of the e-mail provider, allowing them to determine what constitutes proper use. Usually, the user's privacy is protected but not indefinitely assured.

If, as described in the story of Custom E-Mail Corporation, a lesser known and less financially stable e-mail provider finds itself bankrupt, many policy assurances of e-mail privacy may be ignored. This could lead to a company purchasing all the archived e-mail accounts to analyze the information to build a persona. This poses a serious risk to privacy that the users may have no recourse to rectify.

The Latent Categorization Method

Dr. Kai R. Larsen and Dr. David E. Monarchi collaborated on a project in 2004 called, *A Mathematical Approach to Categorization and Labeling of Quantitative Data: The Latent Categorization Method* (Larsen and Monarchi). They collected abstracts of academic journal articles and studies and processed them through a series of applications developed in-house that mathematically categorized the subjects of the articles based on reoccurring words.

This process was created so that a collection of readings can be categorized by the most frequently occurring topics. Larsen and Monarchi used business related readings that were categorized into leaves with topic labels that included resources, knowledge, group, network, job, brand, funds, tax, etc. These seem to be obvious categories that would appear out of business related studies.

What would happen if this categorization process were applied to another form of information? What about corporate news articles, inter-office memos, or perhaps, maybe something a little bit more personal and sensitive. What about an entire e-mail system?

What Interests Do Companies Have in Collecting E-mails and Analyzing Them?

Through profiling using sophisticated software, e-mail can be analyzed to very accurately portray a user. This is extremely valuable information to a marketing company who could then use this information to offer targeted products and services. E-mail can accurately reflect an individual up to the point where a person could almost be perfectly identified, and thus could be marketed with incredible accuracy. The other result of profiling could be an accurate portrayal of a person's involvement in a social network. Products could be tailored to one's personal preferences or to close family and friends. A person might receive an e-mail right before his or her brother's birthday advertising a product he might be interested in having. This would provide a rewarding opportunity that would produce desirable information for which companies would pay copious amounts of money.

The Difference Between E-Mail Messages and Journal Abstracts

E-mail communication is generally considered to be a personal communication tool. Although it is also widely utilized in the business world, it can be considered the least formal of all the communication forms in the business world. This means that typos are common, as is slang, abbreviations, and other odd words will make their way into e-mail messages.

This is different from journal abstracts and summaries. Abstracts are generally written to give the potential reader a great deal of information in just a few sentences. They cover the main subjects, provide a brief conclusion, and describe the process of the authors' rationing. Above all, they are reread, edited, and corrected for errors. E-mails can receive such care as well, but most likely will not. Many e-mail users will jot down a few sentences to confirm an engagement, make their recipient aware of something, speak informally to a friend or a colleague, etc. Terms of endearment are also common, and such addressed messages are usually written as 'spoken' language.

The Use of Latent Categorization to Categorize E-mail Messages

With the assistance of two of the University of Colorado's Systems Division Professors, Dr. Larsen and Dr. Monarchi, the authors of this chapter applied the latent categorization process to categorized e-mails. The authors recruited volunteers to take part in this study. Seven web-based e-mail users volunteered their 'sent-mail' archives from the

University of Colorado at Boulder e-mail account, including the two authors of this study. 'Sent-mail' refers to the e-mail messages that the users send from their accounts. The authors chose these messages (as opposed to received messages that are personally filed) because these are the most personal uses of the system. Personal privacy infiltration is, after all, the main interest!

The text from the body of each e-mail was used in the study, and was output to a text file using Microsoft Outlook's 'Export' feature, after setting up Outlook to interface with the web-based e-mail accounts. Outlook was set to output only the bodies of the e-mails to the text file. This way, removal of the redundant e-mail headers and subjects were not necessary. Each message body was then separated with a group of values, or a flag, to identify the volunteer, and to identify the message number for the volunteer. The flag is a value that sends an internal message to the parser application used to scan the messages. In this study, the flag %!%N1!%!E1 was placed between each message body. The characters %!% were used as an assumption that the volunteers would not be typing %!% in their messages as a part of their conversation. The N1 and the E1 were code that was used to identify the volunteer and the e-mail message to the application.

Once the message body text files have been created, they were fed into a parser (a program that prepares data for processing), which is the first step in the categorization process. In the academic journal use, the parser was created to take out names and symbols. It had to be edited for this study, since e-mails have other forms of text within them that abstracts do not. For example, one of the volunteers had an e-mail signature that is a collection of text appearing at the bottom of each sent message that disclosed their webpage URL and Instant Messaging screen-name. The results would have been flawed should the e-mail signature have been included. In addition, sent messages in many cases include a copy of the previous sender's message as correspondence. For example:

```
> Windows Millennium Edition seems to be an unstable OS... can  
> you recommend something else?  
Sure, I think Windows 2000 or XP might be a good alternative.
```

These also had to be removed, so as not to have previously recorded transcripts making their way into the system again creating another instance. For the most part, these segments were easy to remove because the lines of text that were to be removed began with a greater-than sign (>). Other e-mail systems are known to use different symbols to distinguish the correspondence, such as # or %. The parser was modified to accommodate these systems as well.

The next step was to remove words that had no meaning or contextual value to categorize. A few of these words were: have, not, go, what, and a huge collection of others. Since they don't contribute to the outcome, proper nouns such as names, and pronouns such as he, she, and they, were removed as well to preserve privacy within the study. The correspondence that is left are the choice words to categorize, along with the extra garbled text and gibberish that got past the parser. As mentioned previously, e-mails are an informal

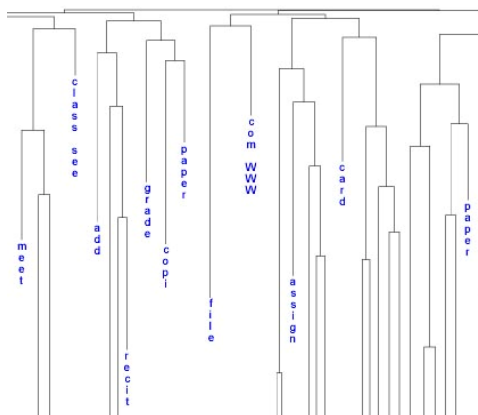
communication tool, and have many misspelled words that need to be addressed, in addition to words that aren't actually printed in dictionaries. Words that have separate meanings when used in online correspondence also need to be accounted for.

The next step in the process is to enter the previously parsed data into a database and point the words to the correct word stem. A stem can be thought of as the root of the word. For example, the word 'accredited' has a stem of 'accredit.' The word 'business' will have the root 'busi' to accommodate 'businesses.' This process can also be used to point misspelled words to their intended stem. So, if someone spelled business as 'busness,' then it can be directed to the 'busi' stem. Words that are made up, or don't appear in dictionaries could be pointed to their intended stem as well. The popular Internet expression, "LOL," could be stemmed to 'laugh.'

Once the word stemming task is completed, the mathematics-based categorization steps are run, which will not be covered in this chapter. Larsen and Monarchi's study provides a technical description of the mathematics involved, should the reader be compelled to investigate further (Larsen and Monarchi).

Results of the Categorization

Dendrogram. The figure below is a dendrogram that represents the analysis of the most frequently occurring topics within the categorized e-mail messages. According to the figure, Class See, Paper, and Grade were among the most frequently occurring because they are the closest to the top. This makes sense since most of these topics are frequently included in correspondence between the teaching assistants that were used as volunteers, as expressed in our limitations. Had the authors been able to recruit volunteers from a more diverse pool, these topics would have been much different.

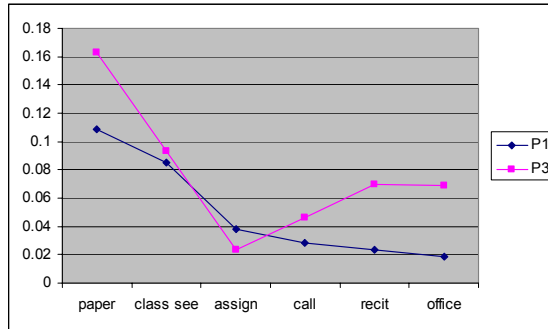


P1 Analysis. The data could also be used to analyze a single subject. Volunteers were not referred to by their names but by handles such as person 1 (P1), person 2 (P2), etc. The first volunteer (graph) is obviously a teaching assistant, and most likely was a key contributor to the frequently occurring topics. The corresponding number for each word in the graph is the percentage of the main stems that word comprises. For example, the word "paper" and its roots account for a little over 10% of all the stemmed words, i.e. 10% of all meaningful words and phrases in person 1's email relate to paper. When compared

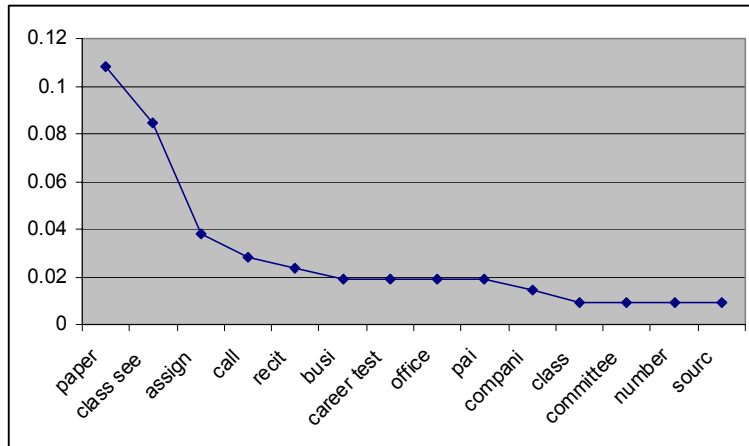
to other commonly occurring stems, such as assign, recit (recitation), class, and busi (business), it is possible to deduce that this person discussed papers relating to school and probably is a student that teaches a recitation. The appendix of this chapter lists the most frequently occurring word stems for all test subjects and the corresponding percentage that each word comprises of the entire subject's total e-mail.

P1 Similarity to P3. Person 1 and Person 3 are similar according to the topics that they share, as expressed in the accompanying graph. It can be assumed that P1 and P3 have similar lives outside of their Internet correspondence.

This was also confirmed by a factor analysis showing that P1 and P3 were quite similar, and that they were different from P4 and P5.



Limitations of the Presented Research. There are some issues that need to be disclosed regarding how this research was executed and interpreted. Due to tight time constraints accommodations had to be made. These were mostly volunteer-base issues, and issues relating to the used technology.



The size and diversity of the volunteer base could have been much greater. Ideally, 20 subjects should have been used from a variety of different types of users. All of the volunteers were junior and senior level college students, and five out of seven volunteers were involved in the University's academics as teaching assistants. One of the volunteer's

data had to be thrown out because the data set was simply too small. Working professionals would have made good subjects, along with college professors, and high school students. However, there were issues related to disclosure of business-related material and student-professor relationships that had to be honored.

The other problem was with the technology itself. Monarchi and Larsen's software was designed specifically for the formal text of academic journal abstracts, which is dramatically different from e-mail messages, as described previously. Ideally, many months of attention and fine-tuning would have been needed to make sure that proper categorizing and a better stem dictionary was created and used. In addition, the technology for Larsen and Monarchi's study has not been finalized yet. At the time of print, their study of the Latent Categorization Method is an ongoing and evolving project.

Business-Related Uses for Latent Categorization of E-Mail Archives

With the issue of personal privacy pushed aside or disregarded, the benefits for a system like this are enormous. One of these uses was already exemplified through the case of Custom E-Mail Corporation and Web Profilers, Inc. An obvious way to implement this categorization system would be to combine it with a free e-mail service. The company could then analyze the content of user emails and tailor advertising to each user based on prevalent categories in his or her email. Another use would be to have a system like this built to analyze individual e-mail accounts in order to monitor productivity in the workplace. The possibilities are endless, and as processing power increases and more data can be analyzed in a shorter period of time, the latent categorization of e-mails will most likely become a staple, or permanent feature of e-mail server software packages.

Other Uses for Latent Categorization Technologies: Social Networking

What if this technology can be modified and used for another purpose? For example, the use of analyzing social networks, as mentioned earlier. What if the categorization criteria were modified not to look for words and topics, but to look specifically for names and e-mail addresses? Frequent contacts would have a higher categorization occurrence than those of less used contacts. With an abundant amount of data, a rather accurate network of correspondents can be built and visually diagrammed! This would graphically detail how close e-mail users are to each other, and assumptions can be made about their 'offline-behavior,' or the extent of their relationships outside of the Internet.

The authors and Dr. Larsen consider this social networking use to be the most intrusive use of this categorization technology, and could inflict the most potential harm to those in the social network who are being profiled.

The Dangers of Combining E-Mail Archiving and Categorization

The potential consequences of network profiling are much more detrimental than a person simply being associated with a criminal in a network. What would happen if the host country of the business using this technology were to be occupied during wartime?

Say for example, World War III begins many years from now, and it involves countries in high power. Here, the term, 'high power,' means having a high use of technology, and the manpower to keep their information systems in prime, updated condition. The latent categorization method of categorizing e-mails and social contacts are in high use. Imagine that the data created by an activist force was lost to a foreign power. This foreign power now has a thoroughly detailed knowledge of the people in power of their activist organization, based on the frequency of contact, not to mention the most frequently occurring topics of the users of the system! It would be extremely easy for the foreign force to find out the mastermind(s) of their resistance, and terminate them.

The abundance of this data could inevitably be more revealing and dangerous to users rather than helpful to businesses. The scenario of World War III is perhaps a little extreme and a little paranoid, but the threat made present by the abundance of this data definitely outweighs the benefits created for a single company's productivity. This is an interesting take on the consequentialist view of ethical thinking. Archiving e-mails may seem like a good practice to help maintain a productive and safe work environment, leading to the end result of good management. However, beyond having a productive team and good management lies the implied risk of the lingering data, which could endanger those whose information is being held. In the user's best interest, archiving of messages should be avoided when embracing a consequentialist end-based view.

Conclusion

Issues and concerns of e-mail archiving stemming from ethical privacy intrusion are plentiful. As already discussed, the wealth of personal information stored in e-mail is vast; the scope can encompass both personal and professional topics, very private or open public information, or just about anything that is discussed on any given day. This creates a particularly vulnerable avenue for privacy intrusion, and very dangerous consequences if this data falls into the wrong hands.

While the threat to privacy in this respect is huge, other privacy problems can arise from opportunities involving analysis of this type of data. Well-intentioned, non-criminal use of e-mail could possibly pose more complicated threats to privacy due to the ambiguity of privacy protection guidelines. While malicious intent is defined through law, viewing or analysis for use in studies or business purposes is not. As discussed throughout the chapter the issue of e-mail analysis, in particular the process of e-mail categorization to create personas, has become a reality. This raises new questions about privacy invasion that are developing into pressing concerns because of the lack of safeguards such as clearly defined policies and regulation. Along with issues of criminal activity, this new method of e-mail categorizing will

increasingly become more relevant and threatening to the concept of privacy.

Works Cited:

Larsen, Kai R. & Monarchi, David E. "A Mathematical Approach to Categorizations and Labeling of Qualitative Data: The Latent Categorization Method." *Sociological Methodology*, Article : 20-129. (n.d.).

"Research and Markets: Many global corporations believe that email archiving is very important but very few organizations have a formal e-mail archiving policy in place." Proquest 395515261. M2 Communications Ltd. 5 September 2003. <<http://proquest.umi.com/>>

"SEC Interpretation: Electronic Storage of Broker-Dealer Records." U.S. Securities and Exchange Commission. 20 May 2003. <<http://www.sec.gov/rules/interp/34-47806.htm>>

"Use of Electronic Mail." *Administrative Policy Statements*. 1997. University of Colorado. 25 September 2004. <<http://www.colorado.edu/policies/General/e-mail.html>>

Appendix A: Persons vs. topics

Topic	P1	P2	P3	P4	P5	Total
Add	3.77%	0.00%	0.00%	1.51%	0.78%	6.06%
alumni	0.00%	0.00%	0.00%	12.56%	0.78%	13.34%
alumni univers	0.00%	0.00%	0.00%	1.01%	0.78%	1.78%
Assign	3.77%	0.00%	2.33%	3.52%	0.78%	10.39%
Attend	0.47%	2.94%	0.00%	1.01%	0.00%	4.42%
Busi	1.89%	0.00%	0.00%	3.02%	1.55%	6.45%
Call	2.83%	11.76%	4.65%	0.00%	3.10%	22.35%
Card	2.83%	2.94%	0.00%	4.02%	6.20%	15.99%
Career	3.30%	0.00%	0.00%	0.50%	2.33%	6.13%
Career test	1.89%	0.00%	0.00%	1.51%	0.00%	3.39%
Child	0.00%	0.00%	0.00%	1.01%	0.00%	1.01%
Class	0.94%	0.00%	0.00%	0.00%	0.00%	0.94%
class see	8.49%	2.94%	9.30%	1.51%	2.33%	24.57%
Com	2.83%	0.00%	0.00%	1.01%	0.00%	3.84%
com WWW	13.68%	2.94%	2.33%	3.02%	2.33%	24.29%
committee	0.94%	0.00%	0.00%	0.00%	0.00%	0.94%
compani	1.42%	0.00%	0.00%	0.50%	2.33%	4.24%
Copi	5.66%	2.94%	0.00%	0.50%	3.10%	12.20%
examin	0.00%	0.00%	2.33%	0.00%	0.78%	3.10%
File	6.13%	2.94%	0.00%	2.01%	6.20%	17.28%
Format	0.47%	8.82%	0.00%	0.00%	0.00%	9.30%
Girl	0.00%	5.88%	0.00%	0.00%	0.00%	5.88%

Good	0.47%	0.00%	2.33%	0.00%	0.00%	2.80%
good sound	0.47%	0.00%	0.00%	0.00%	0.78%	1.25%
Grade	0.47%	11.76%	20.93%	3.52%	0.78%	37.46%
Guy	0.00%	0.00%	0.00%	0.50%	2.33%	2.83%
Hei	0.47%	0.00%	0.00%	0.50%	3.10%	4.07%
Hope	0.47%	0.00%	0.00%	0.00%	1.55%	2.02%
Insight log	2.83%	2.94%	0.00%	2.51%	12.40%	20.69%
Love	0.47%	0.00%	4.65%	0.00%	0.78%	5.90%
love mom	0.00%	0.00%	18.60%	0.00%	0.78%	19.38%
Mail	3.30%	8.82%	0.00%	3.02%	1.55%	16.69%
Meet	5.66%	5.88%	0.00%	1.01%	4.65%	17.20%
number	0.94%	0.00%	0.00%	1.51%	0.00%	2.45%
Office	1.89%	0.00%	6.98%	0.00%	0.78%	9.64%
Pai	1.89%	0.00%	2.33%	1.51%	1.55%	7.27%
Paper	10.85%	5.88%	16.28%	4.02%	12.40%	49.43%
Print	0.47%	8.82%	0.00%	0.50%	0.78%	10.57%
Quiz	0.00%	0.00%	0.00%	2.51%	1.55%	4.06%
Recit	2.36%	2.94%	6.98%	0.00%	0.78%	13.05%
Resum	0.00%	2.94%	0.00%	0.50%	1.55%	4.99%
send univers direct	0.00%	0.00%	0.00%	1.51%	0.00%	1.51%
Sourc	0.94%	0.00%	0.00%	0.00%	0.00%	0.94%
subscrib	0.00%	0.00%	0.00%	2.01%	0.00%	2.01%
univers	0.00%	0.00%	0.00%	3.02%	3.10%	6.12%
univers origin	4.72%	5.88%	0.00%	33.17%	13.18%	56.94%
Work	0.00%	0.00%	0.00%	0.50%	2.33%	2.83%
	100.00%	100.00%	100.00%	100.00%	100.00%	500.00%