

# YOUR MARK IS MY DIRT: IMPACT OF EMAIL SIGNATURES ON DECISION MAKING

Dinesh Rathi, Michael Twidale, Vandana Singh

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

501 E. Daniel, Champaign, IL 61820, USA

{drathi, twidale, vsingh1}@uiuc.edu

## Abstract

*In order to text mine email data it is important to address the substantial amount of noise usually contained in the data. This noise can skew the results of data mining and so reduce the effectiveness and efficiency of decision support systems that use these techniques. Ideally the noise is removed in pre-processing. The paper presents a case study of a series of steps to progressively clean a set of email data, and examines the negative effects that a seemingly innocuous component of email, a signature, can have when email is repurposed for text mining. The paper advocates trying to understand the consequences of noise on the mining process through small scale pilot studies as a way to gain a better understanding of the real meaning of the results of mining algorithms. This approach also serves as an exemplar for understanding the implications of variations in input patterns for the interpretation of the results data mining.*

**Keywords:** Text mining, text data, email, data preprocessing, data cleaning, decision support system

## Introduction

The majority of data in an organization is text, forming up to 80% of the total (Karanikas and Theodoulidis, 2002; Leavitt, 2002). Text data in an organization consists of email, minutes of meetings, memos, reports, knowledge repositories, technical reports, customer correspondence, contracts, patents, etc. (Dorre, 1999). Email-based system use has increased substantially and many corporations are using email or online text messaging services to provide support to their customers. These text-based systems are an important resource which contain substantial amounts of data that could be mined to extract information and help organizations to provide better service. According to Tang et al. (2005), email is used in a large number of text mining applications including email analysis, filtering, summarization and newsgroup analysis.

Although there is great potential for the application of data mining techniques to text resources such as email, there are various problems that must be addressed. One is the quality of the raw data. It is tempting to offload this as a rather tedious problem of data cleaning that is currently done by selection of a number of semi-automated methods will hope they will eventually be done entirely automatically. In this paper we want to look at this problem in more depth and understand the consequences of dirty data and how progressive levels of data cleaning have an effect on the results of data mining. This may seem uninteresting – surely everyone knows that dirty data is bad because it causes meaningless or misleading results for data mining algorithms? However if, as we suspect, problems with data cleaning will be around for some time despite the best efforts of the developers of cleaning algorithms, then it is worth understanding their effects rather than just focusing on creating the ultimate data-cleaning algorithm. Outside test collections data is rarely pristine. An understanding of the consequences of dirty data on the results of data mining can greatly help the interpretation of those results, and help in identifying and diagnosing artifacts of poor quality data that may be rectified in subsequent rounds of data analysis. We do not claim to have a complete analysis of the issues, but do claim that more work in this area will have two main benefits. Firstly and most obviously it will help users spot potential problems and misleading results from data mining that can be verified by focused efforts at data cleaning. Secondly there is a more general contribution in informing the more considered interpretation of the results of data mining. We fear that those using data mining algorithms may be in danger of using them rather too uncritically. A more skeptical approach to interpreting the results, diagnosing problems and looking for side-effects

that may have accidentally caused those results (dirty data being just one) can greatly help in effective interpretation of the results.

## **Cleaning email for text mining**

Despite its inherently textual nature, email in its natural form is difficult to use for text mining. Typically, email is noisy, containing large amounts of potentially irrelevant and distracting or misleading information (Tang et al. 2005). Worse, depending on the nature of the text mining task, the very definition of which parts of the data are to be considered noise will vary, inspiring the punning title of this paper. For example, a researcher who wants to develop a social networking model will consider data in the header such as the sender's and receiver's email addresses and personal information in the signature as important for their research, while other parts of the email may create noise in their analysis. By contrast, people interested in understanding themes and patterns in an archive of emailed technical help requests (such as ourselves) are mostly trying to form clusters around content, and almost all header and signature information will be noise. Consequently an entirely automated data cleaning algorithm is unlikely to be feasible. Given the need for human intervention, it is helpful to understand the consequences of various kinds of noise in the raw data on the results of text mining, both in making tradeoff decisions about how much data cleaning will be cost effective, and in diagnosing from the results that more prior cleaning needs to be done on a subsequent iteration.

For simplicity, email can be divided into three major information blocks:

- 1) The header block, containing tracking and delivery information like to whom, from whom, time, date etc. (figure 1).
- 2) The content block, containing the main message or text of the email (figure 2).
- 3) The signature block, containing the sender's signature. (figure 3).

```
1 6459 RE: html assist 1970-01-01 00:00:00      2004-01-14 18:45:15
2 Content-Type: text/plain; charset="utf-8"\
3 Content-Disposition: inline\
4 Content-Transfer-Encoding: binary\
5 MIME-Version: 1.0\
6 X-Mailer: MIME-tools 5.411 (Entity 5.404)\
7 RT-Send-CC:\
8 RT-Send-BCC:\
9 Content-Length: 787\
10 > [abc@udc.com - Wed Jan 14 12:33:03 2004]:\
11
```

**Figure 1. Header Block**

```
1 > Hi Larry,  
2 > It is great to meet you, and hear from you. I think I  
3 > successfully moved all of my stuff links into the active  
4 > space. Would you please take a look at it? It's ABC408NE.  
5 > Everything seems to be working when I look at it from the  
6 > XYZ space.  
7 >  
8 >  
9 > If I have other needs, you can be sure that I will be in  
10 > touch.  
11 > --Ellen  
12 >  
13 >  
14 Hi Ellen,  
15  
16 Your webpages look really nice. All of the links are  
17 working so I think you are set. If you have any more  
18 problems or questions, please don't hesitate to ask.  
19 Best of luck and have a great time.  
20  
21 Larry
```

Figure 2. Content Block

```
1 Larry Singer  
2 User Services Office  
3 Department of X  
4 University of ABC, Catmat
```

Figure 3. Signature Block

Our aim in this paper is to give a case study of the impact of just the signature block on cluster results and the consequent impact of these results on decision making. We will show the level of dirt that is contained in emails, and how noisy signatures can skew the decision making process and reduce the effectiveness of a decision support system. The case will further strengthen the argument of the importance of text cleaning for data mining, a rather unglamorous but, as noted, essential activity. In order to inform work on developing better data cleaning processes, it is important to understand the consequences of any remaining dirty data. It is unlikely to ever be feasible to eliminate all dirt, and certainly considerations of cost will force researchers to deal with more dirt than they might like. Therefore we believe an acknowledgement of the existence of dirt in the data and an understanding of how it impacts the results of data mining is critical to a considered interpretation of those results.

Several papers have discussed the importance of the quality of text data in knowledge extraction and decision making, and the importance of data preparation (Cooley et al., 1999; Jung, 2004; Redman, 1998; Strong et al., 1997; Tayi and Ballou, 1998; Zhang et al., 2004) but do not provide much sense of the impact that un-cleaned data will have on results. This is a pervasive problem because as Zhang et al. (2004) note: “real world data is not pure” and poor data preparedness can lead to disastrous results and poor decision making (Rahm and Do, 2000). According to Rajagopalan and Isken (2001), “Data preparation, while generally recognized as an important part of the data mining process, has not received the same level of research attention as data quality or data characteristics”

Numeric data also requires data cleaning but involves different actions such as removing duplicates, maintaining schema integrity and reducing structural conflicts, (Parent and Spaccapietra, 1998). Data from the web, of which the majority is text-based is typically cleaned by removing banner ads, images, navigational guides, etc. (Yi et al., 2003). Email based data is typically cleaned by removing email headers, signatures, special characters, reply indentation, HTML elements and tags, extra spaces in words and lines, etc. (Tang et al., 2005).

Karanikas and Theodoulidis (2002), define Knowledge Discovery in Text (KDT) as the process of identifying valid, novel, potentially useful and ultimately understandable patterns in unstructured textual data. In KDT there are three

main steps: i) collect data, ii) pre-process the data, iii) apply text mining techniques. We focus on the second, data pre-processing stage of KDT, describing the data cleaning efforts (and failures) undertaken. In this paper we compare the results obtained after using a clustering text mining algorithm on two intermediate versions and the final pass of cleaning the data.

## **Problem Definition and Experiment**

Our objective is neither to evaluate the efficiency of any particular algorithm nor develop a new algorithm to clean the data. Rather we want to evaluate the impact of signatures on the clusters which in turn will impact decision making. It is obvious that uncleaned data has an impact on the results, but the question is *how* it impacts and how much. Given the costs of different data cleaning strategies, and that no one method can do everything, it is important to understand the tradeoffs of different investments of cleaning efforts.

We define a signature (figure 3) as the personal information of the users typically written at the end of an email after closing salutation remarks, which may include the name of the sender, address, phone number, affiliation and any personal quotations. For simplicity of processing we have not included the closing salutation such as ‘thanks’, ‘regards’ and ‘name of the person’ in defining the signature.

The study was conducted on data obtained from an organization’s Technology Support Group (TSG) help desk. This is an email based help support system. TSG provides support to users to solve their problems on practically everything related to Information Technology including email, other software, hardware, passwords, server space allocation, printing issues and other support such as equipment reservations and building access. Users who experience a problem can register it using an online form or by sending an email to *help@abc.xyz* (throughout the paper all identifying information such as the actual email address have been anonymized).

Raw textual data generally has to be processed prior to data mining (Rajagopalan and Isken, 2001). First, it has to be converted to the format that can be used by the text mining algorithm. Second, the data has to be cleaned of special characters and symbols as they might interfere with the functioning of the algorithm. For example “<”, “>” are regarded by the algorithm used in our study as special characters which can only be used as document separators. The moment the algorithm encounters these characters in the document it stops and generates an error message. Of course these characters are very common in email messages and so have to be removed. Third, the data has to be processed for missing values, missing data fields or duplication (Saravanan et al., 2003) and finally the data has to be cleaned of unwanted information or the results will be skewed.

The text mining software used in this study (Mei and Zhai, 2005) uses the concept of a theme. Themes are defined as “probability distribution of words that characterizes a semantically coherent topics or subtopics” (Mei and Zhai, 2005). The clusters of themes were extracted using simple probabilistic mixture model which consider “words as data drawn from a mixture model with component models for the theme word distributions and a background word distributions”. In this model each word in a document has the same weights (Mei and Zhai, 2005; Zhai et al., 2004). The mining algorithm requires four steps of preprocessing to convert the data into the format required:

Step 1: Put all the emails in a single document

Step 2: Give each email a separate identifier with a start tag of `<DOC UniqueId_Start Date and time_Closing Data & time>` and end tag: `</DOC>`.

Step 3: Each message in an email should be separated by tags: `<new_message>` at the start of a new conversation, and `</newmessage>` at the end of all messages except the last one.

Step 4: The last message also starts with `<new_message>` but the closing tag is `</new_message>` just before `</DOC>` (Figure 4).

```

<DOC 1805_2003-04-08 23:04:53_2003-04-09 18:37:19>
<new_message>
  Hey Tech Person!
  Signed into my e-mail this PM to find that the format has changed.
  .....
  Thanks,\
  Liz\
</newmessage>
<new_message>
  Hey Liz,\
  sorry for the confusion! It looks like you normally use ABCD express \
  .....
  Let us know if you are still having problems-Brad\
</newmessage>
<new_message>
  Liz, I'm forwarding this to you at your Express account
  since Brad's \
  .....
  John\
  John Chen\
  Assistant Director\
  Instructional Technology Office\
  Department of X
  University of Y, Catman\
  222 X Department Building, OP-22\
  222 W Peter St. Catman, AK 22222, USA\
  +1 222 222 2222 / +1 800 222 2222
</newmessage>
<new_message>
  We're still figuring out this situation ourselves. Besides calling \
  .....
  Sorry about these problems -- they're out of our hands!\
  John
</newmessage>
<new_message>
  Thank you so much for getting my e-mail back! I must admit, \
  .....
  Thanks again and you all have a good evening,\
  Liz\
</newmessage>
<new_message>
  I'm going to close this one, since this user's problem is solved and \
  .....
</new_message>
</DOC>

```

Figure 4. Document Format for Algorithm

**Email Data Overview**

The original data was in a MySQL database of over 550 MB, containing over 10,000 emails including (as found in subsequent processing) about 1000 spam emails. Before starting the data cleaning, data analysis (Rahm and Do, 2000) was done to understand the types of errors, inconsistencies and other noise in the email data, and the ways that might be used to reduce or remove those kinds of noise from the data. 50 non-spam emails were randomly selected and studied manually to understand the format of the email based ticketing system. We were able to identify the different blocks of email and, based on our understanding of the organizational context, interpret the meaning of these messages. One aim of our ongoing work is how to interpret the meaning of the results of text mining and how these might compare to a ground truth of a lavishly expensive human analysis of the data set. The following aspects of the email based data were considered noise for the purposes of our investigation:

- a) Spam Email.
- b) Attachments in the email. Pre-processing of the dataset received had converted these attachments into special characters and thus was important to remove them. The attachments in the emails were PowerPoint slides, pictures, screen shots etc.
- c) Every part of the email header except for the date and timestamp and unique IDs of the tickets.
- d) Duplicate reply or forwarded message contents.
- e) The signature of senders at the end of the email.

**Data Pre-processing for Noise Reduction**

Table 1 shows the various phases of processing the data which converted it into a suitable format for the text mining software, cleaned it up, and as a consequence substantially reduced it in size. Data cleaning is a costly, time consuming activity Buchner et al. (2000) and can be done using different approaches as highlighted by Tang et al. (2005) such as using existing software tool to remove signatures, writing your own code or cleaning manually. Our experiences confirm both the high effort involved and the advantages of combining different approaches.

We adopted three major approaches to clean the data:

- 1) Writing several Perl scripts to remove special characters and attachments.
- 2) Manual cleaning of non-standard patterns of special characters.
- 3) Visual inspection.

<b>Table 1: Stages of Data Reduction through Cleaning</b>		
Stage	Data Size	Remarks
0	580 MB	<ul style="list-style-type: none"> <li>- Original Size of the MySQL database</li> <li>- No processing done</li> </ul>
1	250 MB	<ul style="list-style-type: none"> <li>- Table and attribute selection done and all emails exported into one '.txt' file</li> <li>- Contained all the three blocks of email (as shown in figures 1, 2, 3)</li> </ul>
2	80 MB	<ul style="list-style-type: none"> <li>- Removal of header block using perl script</li> <li>- Four Step data formatting to meet algorithm requirements using perl script</li> <li>- Removal of attachments using perl script (partial removal could only be achieved)</li> <li>- Removal of duplicate message (forward message or reply message) using perl script (partial success in removal)</li> </ul>
3	40 MB <i>(Data Set A)</i>	<ul style="list-style-type: none"> <li>- Visual inspection to check formatting</li> <li>- Second stage of attachments removal using perl script</li> <li>- Second stage of removal of duplicate message (partial success in removal)</li> <li>- Second stage of header left over in email texts</li> <li>- Partial removal of spam mails with special characters</li> <li>- Removal of HTML tags, special characters, asterisks, etc.</li> <li>- Manual intervention to remove attachment characters</li> <li>- Visual inspection to check the quality of the data</li> </ul>
4	26 MB <i>(Data Set B)</i>	<ul style="list-style-type: none"> <li>- Third stage of character removal using perl script identified as noise after previous stage inspection</li> <li>- Removal of leftovers such as header, HTML tags, some signature blocks, special characters, asterisks embedded in the content block using manual and perl script.</li> <li>- Visual inspection to check the quality of the data</li> </ul>
5	11 MB <i>(Data Set C)</i>	<ul style="list-style-type: none"> <li>- Third stage removal of all types of spam emails using perl script and by manual intervention</li> <li>- Removal of signature</li> <li>- Visual inspection to check the quality of the data</li> </ul>

## Results and Discussion

### *Result 1: Data Preprocessing*

Figure 5 and table 1 present the different stages of the data cleaning. The figure shows the raw quantitative effects of data formatting and cleaning. Unsurprisingly, removing most attachments and quoted reply parts of emails (in stage 2) leads to a substantial reduction in the size of the dataset. Compared to this, other dirt like email headers and signature block are miniscule in size. This shows the skewness in the type of dirt in email based data as compared to other types of text based data like newsgroups which may have only XML or HTML tags and header related dirt. We found that that the scripts used were not able to fully achieve the desired results because the pattern of even similar kind of dirt is difficult to define in all possible ways. For example, signature blocks have different patterns like dashed lines, asterisks, tildes, quotations, etc. Even several passes of scripts and manual cleaning could not completely clean the data. Further cleaning was done to remove leftover header elements, attachments, forwarded

messages, a few remaining spam emails and other interwoven noises in the text such as HTML tags, special characters, lines, asterisks, etc. At each stage a visual inspection was done. Randomly picked emails had their content compared with the original email format in order to confirm the data integrity under the different processing stages. More careful work can lead to more effective data cleaning algorithms, but our experience confirms our suspicion that typically most data mining algorithms will need to deal with some dirt and so understanding the consequences of that dirt is important.

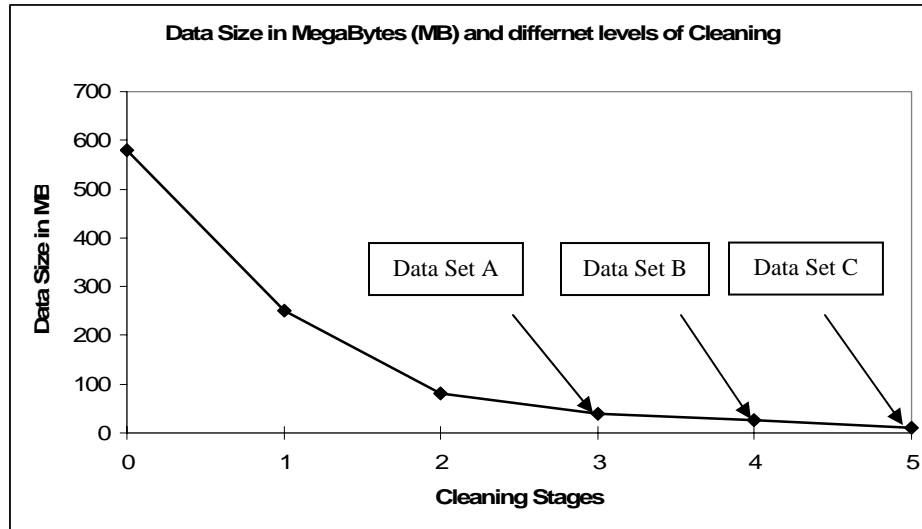


Figure 5: Data Reduction at different stages of cleaning

**Result 2: Clustering Results**

The algorithm parameters (table 2) were set to generate a total number of 30 clusters for each data set. LambdaB controls the appearance of non-informative stopwords such as "is", "are", etc. Several sets of experimentation led to setting lambdaB to 0.90. With this value the algorithm gives better theme generation, in line with the experimental findings of Mei and Zhai (2005). The need to tweak parameters to gain good performance with an algorithm is something of a black art, and likely to be far easier for the developers of that algorithm to do effectively than people who use algorithms developed by others.

Theme generation was a two stage process: first, the data was run through a stemmer, and then the output was used to generate theme based clusters. All three datasets were mined in the same way. Each cluster contains themes having a probability assigned by the algorithm. Each cluster has a total probability of 1. For brevity we just show the top 15 themes for each cluster.

Table 2: Key Parameter Input into the Algorithm
<p><b>Key Parameter input for Algorithm</b></p> <pre> index = "V.90-30-200-5-100%-korvetz/index.bsc"; /* The pointer to the database index file*/ cluster = 30; /*Defining the number of clusters the algorithm should generate*/ lambdaB = .90; /* control the strength of background model to absorb non-informative words from theme*/ it = 200; /* maximum Expectation Maximization iterations */ docGroup = "V.90-30-200-5-100%-korvetz/docids"; /* The pointer of docid files to indicate which docs are in which group */ trial = 5; /*To repeated several times to avoid local maxima */ output = "V.90-30-200-5-100%-korvetz/outPro"; /* To store the result output in which folder cutoff = 0.0001; /*Probability cut-off limit for themes */                     </pre>

Figure 6 shows a cluster from the final, cleanest dataset C. This cluster is relatively easy to interpret as being about applying for a job at TSG. In these examples the certain actual values have been removed for reasons of privacy and replaced by generic names (italicized). For example, Job Place stands for the name of the organization. Unitalicized words are actual words from the results. Note that some rather generic words still get through with the setting of lambdaB chosen.

<u>Themes</u>	<u>Probability</u>
<i>Level of Job</i>	0.017
<i>First Name1</i>	0.015
position	0.015
<i>Area of Job</i>	0.014
<i>Job Place</i>	0.013
interest	0.013
am	0.011
interview	0.011
<i>First Name2</i>	0.011
research	0.011
information	0.010
very	0.009
service	0.009
resume	0.009
assistantship	0.009

**Figure 6: A Cluster Result related to applying for a job position**

In the less clean datasets A and B, there are signatures in practically every email and their presence has a significant effect on the results. One effect is that whole clusters are created that are essentially about some ‘concept’ of signature. While perfectly correct, this is not particularly useful or interesting for data mining. Since only a limited number of clusters are created (the parameter was set to 30) this crowded out more meaningful clusters. Figure 7 gives three examples of signature related clusters from datasets A and B.

<u>Themes</u>	<u>Probability</u>	<u>Themes</u>	<u>Probability</u>	<u>Themes</u>	<u>Probability</u>
of	0.065	<i>Area Code</i>	0.036	<i>First Name4</i>	0.040
<i>First Name1</i>	0.062	<i>Street name</i>	0.036	<i>Last Name4</i>	0.040
<i>Last Name2</i>	0.059	<i>Building Number</i>	0.035	<i>Connector</i>	0.033
manage	0.043	<i>City of location</i>	0.034	<i>Group Affilication1</i>	0.029
<i>Phone Number</i>	0.041	<i>State of Location</i>	0.029	<i>Group Affilication2</i>	0.027
<i>Department Affiliation4</i>	0.037	east	0.024	<i>Group Affilication3</i>	0.026
<i>Group Affilication2</i>	0.037	fax	0.024	<i>Group Affilication4</i>	0.024
<i>Organization Affiliation1</i>	0.037	<i>First Name2</i>	0.020	<i>Program Name</i>	0.023
<i>Group Affilication4</i>	0.036	<i>Zip Code</i>	0.020	<i>Nick Name (First Name4)</i>	0.022
<i>Department Affiliation3</i>	0.036	<i>Department Affiliation2</i>	0.017	class	0.021
<i>Organization Affiliation1</i>	0.034	<i>Phone Number</i>	0.014	audio	0.021
<i>Department Affiliation2</i>	0.034	<i>Last Name3</i>	0.014	<i>Organization Affiliation2</i>	0.020
<i>Department Affiliation1</i>	0.033	<i>Department Affiliation1</i>	0.014	<i>Department Affiliation1</i>	0.019
<i>Organization Affiliation3</i>	0.031	<i>First Name3</i>	0.013	<i>Department Affiliation2</i>	0.018
<i>Group Affilication5</i>	0.028	edu	0.013	<i>Department Affiliation3</i>	0.018

**Figure 7: Three Signature Clusters (Actual identifying data has been replaced by italicized generic names)**

Additionally, the prevalence of signatures meant that elements in signature blocks increased the strength of certain themes in some clusters. This reduced the probability assignment of other less prominent, but perhaps more significant and meaningful themes, thus pushing them far down (beyond 15) in the cluster. As a result these signature-reinforced themes made clusters harder to interpret and also diluted the novelty of some of the clusters by introducing elements from the signature and thus making clusters less distinct.

The absence of signatures from the most cleaned data set (C) led to better and clearer clustering. The clusters are clearly identifiable and the probability of themes are not diluted or increased by the words frequently occurring in signature blocks. Comparing the results in figures 6 and 7, certain key words (which we cannot share for reasons of privacy) such as ‘Group Affiliation<sup>4</sup>’ (figure 7) and ‘Area of Job’ (figure 6) and similarly ‘Department Affiliation<sup>3</sup>’ and ‘Level of job’ are actually the same words both the figures, however their context is different. Thus, themes present in emails which had a certain context related to the email content were marginalized by the strong presence of the same words in the signatures which thus created poor clustering and led to altering the context of the theme. In order to further understand the impact of signature on cluster and subsequently on the interpretation of the results, the clusters were then given to two independent evaluators, E1 and E2. The evaluators were asked to assign them to one of three categories; definable cluster (they could guess what the cluster was ‘about’), non-definable clusters and signature clusters. Both evaluators were familiar with the organizational context from which the emails came. Table 3 summarizes their interpretation.

Data Set	Number of Clusters	Number of definable clusters		Number of non-definable clusters		Number of Signature Cluster	
		E1	E2	E1	E2	E1	E2
A	30	12	12	12	13	6	5
B	30	13	11	13	16	4	3
C	30	25	27	5	3	0	0

**Table 3: Comparative cluster results**

The result from the independent evaluators confirm the strong presence of signature based clusters (similar to figure 7) with 5 to 6 in dataset A and 3 to 4 in number in dataset B as compared to none in dataset C. As a result, we can (not too surprisingly) confirm the importance of data cleaning as advocated by many researchers. Of more interest, we can say *why* data cleaning matters, and the impact that imperfect cleaning has: the presence of signatures not only crowded out the number of potential meaningful clusters but also diluted the definable cluster number – the numbers of definable clusters were reduced to half the number as compared to clusters from cleaner data set (C).

Many corporations are using historic data either in text or in numeric format for making tactical and strategic decisions. However, text mining results depend on the quality of the data that is fed into the algorithm. Thus, poor data quality as shown by our experiment will lead to inferior extraction of patterns by the text mining algorithm which in turn will have significant impact on business decisions (Strong et. al., 1997). Huang et al. (1999) suggest that corporations have experienced adverse effects of decisions based on poor quality of information.

How can these impacts occur? We will illustrate using a small scenario from our text mining case. Assume that based on the results of text mining their email help data, an organization discovers that a key area for which users seek help is ‘networking’, and the organization has few networking experts. The organization duly recruits an expert in networking but they later find that this support person do not have much work to do while hardware related issues are still not resolved quickly. This means the organization has recruited a person with wrong skills. It turns out that the ‘networking’ theme arose not because users were seeking help on networking related but because ‘networking’ was used in the signatures of some help desk support staff or the users. Thus, signature elements have skewed the results and led to wrong decision making.

Exhorting organizations to only do data mining on pristine data is insufficient. We believe it is important to sensitize users in their interpretation of data mining results to be on the alert for aspects of those results that may indicate the presence of previously unconsidered problems with their data that if diagnosed can be remediated.

## **Future Work**

We plan to analyze the impact that different noise elements in email have on results, including the change in probabilities of themes in the clusters. We will conduct human evaluations of the results to try to understand their impact on decision making. In this study, signatures were removed but opening and closing salutation such as “Hi Bill” and “Thanks, Regards, Jane” were kept. We want to explore how feasible it is to automatically remove these too, and their effect on the final results. Given the difficulty of developing foolproof methods of data cleaning and their dependence on the purposes for doing the data mining in the first place, it would be helpful to develop a framework to enable more rational cost-benefit analysis of different levels and techniques of data cleaning.

Finally we plan to continue this approach of relatively quick and low cost analyses with somewhat easy to interpret problems with raw data and algorithm parameter setting as a way to systematize the detection and diagnosis of problematic initial results of data mining with a view to supporting their iterative remediation as part of a refinement process. That is, we doubt there is one right way to mine a data set, one right data cleaning algorithm to use or one right set of parameters to apply. Rather, those using data mining should always be continually refining their work, even when working on a single dataset. We believe that ongoing analysis in this vein of understanding the consequences of errors and hence being able to spot from the results when errors might have occurred can be useful. One analogy for this approach is from programming. In the ideal, one would write a program that ran correctly first time. More realistically, it makes sense to learn certain debugging skills to gain a sense of when something maybe has gone wrong with ones output, and what are likely causes, inspired by experiences with certain characteristic symptomatic error patterns in a range of contexts.

## **Conclusion**

We have outlined a case study of the impact of signature blocks on cluster results and their impact on decision making. By an investigation of the results of text mining of two intermediate stages of our data cleaning process and a comparison with the final stage, we can begin to show exactly what impact partially cleaned data can have on the results, which in turn will eventually impact interpretations and decisions based on that data. The paper highlights the effect of an individual’s mark or email signature as potentially being noise from the perspective of text mining, dirtying the data. We also investigated the consequences of particular error types and the difficulty of cleaning all possible dirt in email with one single approach. Cleaning of email requires both visual inspection as well as fine tuning of cleaning scripts to encompass the uniqueness of dirt in different emails. Of course that is infeasible in industrial-scale applications. But such small scale investigations into the actual meaning of mining results and the ways those results can possibly be misinterpreted can be a useful sensitizing activity and one we would advocate in any pilot study.

The work also shows how a small scale study with a relatively small sample size can help in understanding the underlying meaning of the data mining process, by going back to the original data and examining it critically in the light of iterative use of a data mining algorithm. By treating email signatures as a relatively easy to understand exemplar of the larger issues of data cleaning and the overall interpretation of the results of data mining and the sensitivity of those results to the input data and the parameters chosen, we wish to open up an exploration of the dynamic, interpretative, diagnostic and iterative nature of actual data mining practice and how this can be supported by subsequent research in analysis and design.

## **Reference**

- Buchner, A.G., Baumgarten, M., Mulvenna, M.D., Bohm, R., and Anand, S. S. “Data Mining and XML: Current and Future Issues”, in *Proceedings of the First International Conference on Web Information Systems Engineering (WISE '00)*, June 2000, pp. 131 – 135.
- Cooley, R., Mobasher, B., and Srivastava, J. “Data Preparation for Mining World Wide Web Browsing Patterns”, *Knowledge Information System*, 1999, pp. 1-27.
- Doree, J., Gerstl, P., and Seiffert, R. “Text Mining: Finding Nuggets in Mountains of Textual Data”, in *Proceedings of the 5<sup>th</sup> International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 398 – 401.
- Huang, K. T., Lee, Y. W., and Wang, R. Y. *Quality Information and Knowledge*, Upper Saddle River, NJ: Prentice Hall, 1999.

- Jung, W. “An Investigation of the Impact of Data Quality on Decision Performance”, in *Proceedings of the 2004 International Symposium on Information and Communication Technology (ISICT '04)*, 2004, pp. 166 – 171.
- Karanikas, H. and Theodoulidis, B. “Knowledge Discovery in Text and Text Mining Software”, [http://www.crim.co.umist.ac.uk/parmenides/internal/docs/Karanikas\\_NLDB2002%20.pdf](http://www.crim.co.umist.ac.uk/parmenides/internal/docs/Karanikas_NLDB2002%20.pdf), 2002.
- Leavitt, N. “Data Mining for the Corporate Masses?”, *Computer (IEEE Computer)* (35:5), 2002, pp. 22 – 24.
- Mei, Q. and Zhai, C. “Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining”, in *Proceeding of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, Illinois, USA, 2005, pp. 198 – 207.
- Parent, C., and Spaccapietra, S. “Issues and Approaches of Database Integration”, in *Communication of the ACM* (41:5), 1998, pp. 166-178.
- Rahm, E., and Do, H. H. “Data Cleaning: Problems and Current Approaches”, in *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* (23:4), 2000, pp. 1 – 11.
- Rajagopalan, B. and Isken, M.W. “Exploiting Data Preparation to Enhance Mining and Knowledge Discovery”, in *IEEE Transactions on Systems, Man and Cybernetics—Part C: Applications and Reviews* (31:4), 2001, pp. 460 – 467.
- Redman, T. C. “The Impact of Poor Data Quality on the Typical Enterprise”, in *Communication of the ACM*, ACM Press (41:2), 1998, pp. 79 – 82.
- Saravanan, M., Reghu Raj, P. C. and Raman, S. “Summarization and Categorization of Text Data in High-Level Data Cleaning for Information Retrieval”, in *Applied Artificial Intelligence* (17:5/6), 2003, pp. 461-474.
- Strong, D. M., Lee, Y. W., and Wang, R. Y. “Data Quality in Context,” in *Communications of the ACM* (40:5), 1997, pp. 103-110.
- Tang, J., Li, H., Cao, Y. and Tang, Z. “Email Data Cleaning”, in *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge Discovery in Data Mining*, Chicago, Illinois, USA, 2005, pp. 489 – 498.
- Tayi, G. K. and Ballou, D. P. “Examining Data Quality”, in *Communications of the ACM*, ACM Press (41:2), 1998, pp. 54 – 57.
- Yi, L., Liu, B and Li., X. “Eliminating Noisy Information in Web Pages for Data Mining”, in *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, 2003, pp. 24-27.
- Zhai, C., Velivelli, A. and Yu, B. “A Cross-collection Mixture Model for Comparative Text Mining”, In *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 743 – 748.
- Zhang, S., Zhang, C., and Yang, Q. “Information Enhancement for Data Mining”, *IEEE Intelligent Systems* (19:2), 2004, pp. 12 – 13.