The Emergence of A Peer Relationship in E-mail: A Longitudinal Study of a Student-Advisor Relationship

Paul N. Bennett
Microsoft Research
One Microsoft Way
Redmond, WA 98052
pauben@microsoft.com

1 Introduction

During the short handful of decades of its existence, the role of e-mail in society has quickly evolved. Initially, it aided in persisting asynchronous conversations across global distances. However, its nearly universal adoption and integration into both personal and work life necessitated information management technologies that alleviate the burden of dealing with quantities of e-mail – especially in time-critical scenarios. The RADAR project (Freed et al., 2008), of which Jaime G. Carbonell was one of the key leaders, was a major research project at Carnegie Mellon that recognized reducing e-mail overload required a holistic approach that integrated task management with e-mail, had the ability to recognize and prioritize new and ongoing requests, could perform automated planning to implicitly resolve contention for multiple resources (*e.g.*, rooms, equipment), would take proactive action on implicit requests, *etc.*. The project naturally benefited from Professor Carbonell's previous research experience – including text analysis, machine learning, information retrieval, and automated planning.

A variety of new research also developed around the same time or directly in the context of the RADAR project at Carnegie Mellon to contribute to this holistic approach. For example, Carvalho (2008) demonstrated membership in an ontology of speech acts could be accurately predicted using machine learning, and those predictions could be used to improve e-mail management. To aid task management and tracking, Bennett & Carbonell (2005a) demonstrated it was possible to effectively identify e-mails with action-items – explicit requests that required the recipient's attention. They went on to demonstrate that, unlike topic classification, action-item classification was one of the first of a growing class of text classification problems where the predictivity of a sequence of words (*i.e.* n-grams) on class membership could not be captured as a simple bag-of-words (Bennett and Carbonell, 2005b). Finally, Bennett & Carbonell (2007) used estimates of the sensitivity and variance of micro-level predictions (*e.g.*, "is this sentence an action-item?") to make more robust predictions at the macro-level (*e.g.*, "does this e-mail contain an action-item?").

Figure 1: An Illustrative Example of a Deeply Knowledgeable Advisor

Figure 2: The prototypical student may spend early years emulating the advisor – including superficial features of an advisor's appearance.

Figure 3: Count of e-mails sent between a student-advisor pair. Each mail was sent directly from the student/advisor with the recipient explicitly listed (*i.e.* not via a distribution list) in the "To" or "Cc" fields. Academic semester boundaries are denoted in gray – due to data loss a major time gap not represented in the e-mail archive is denoted by the black line.

However, in honor of Professor Carbonell's forward-thinking nature, I now will speculate on upcoming challenges in the domain of e-mail. In particular, while the majority of work has focused on managing the ongoing burden of mail or searching through past mail, we instead raise the question of whether e-mail can serve as a datasource for a retrospective and introspective examination of the user's individual relationships and their evolution over time. This is particularly of interest since the archival nature of e-mail and the fact that e-mail is now widely freely available means most people already have a substantial archive of e-mail. Tools that enable a user to gain insight into relationship management and growth, summarize periods of the user's life at both the aggregate level and within particular social circles, or simply enable a more detailed reflection could provide significant utility to anyone with an e-mail archive.

An example of such introspection is Wolfram's (Wolfram, 2012) introspective analysis using a variety of logged personal data (*e.g.* keystrokes, number of e-mails sent, calls made *etc.*). However, in this short paper, we go beyond that and consider the temporal change in a *one-one* relationship observed by e-mail. In particular, we choose to study the e-mail footprint of a PhD student and thesis advisor. For those who have never met such creatures, examples of a thesis advisor and PhD student are given in Figures 1 and 2, respectively.

The e-mail archive consists of mail between the student and advisor from when the student entered the university until when the student exited the university (several months after graduation). The student's and advisor's names will be withheld from the paper although the astute reader may be inclined to make guesses as to their identities. In order to remove general broadcast mail and focus on personal communications, the mail archive was restricted to the set of mails where either: (1) the student was listed in the "From" field and the advisor was explicitly listed in the "To" or "Cc" field, or (2) the advisor was listed in the "From" field and the student was explicitly listed in the "To" or "Cc" field. This helped eliminate mails to distribution-lists, *etc.* although the "To" or "Cc" field was allowed to have multiple names as long as either the student or advisor was explicitly listed and not simply a member of a group alias. Unfortunately, one period of time is missing from the archive due to data loss (a corrupt drive where an archive from that period was stored). It is possible that other mail was lost due to the use of other mailers at different points of time that may have different archive settings – in the analysis, I assume that these events cause mails to be missing at random so that the aggregate trends can still be viewed as a representative sample. Likewise this is only a sample of the digital footprint of the relationship since much communication occurred in person. Even with the data loss,

the archive spans many years – making this analysis unique relative to the literature. I will not specify exactly how many years, but let us call the total number of years the student spent at the university n. We will see later (see Figure 4) that $n > 6.5$ years.[1]

2 Results & Discussion

Figure 3 gives the number of e-mails sent per day by each of the student and advisor over the course of the graduate student's university career. To ease visual presentation, each point on the x-axis represents a unique day where mail was sent or received; we do not plot each date since on most dates e-mail was not sent or received between the pair. The semester boundaries are shown in gray with a black line depicting the time period where data was lost. Figure 4 aggregates the number of e-mails from each of the student and advisor on a per academic semester basis.

We note several trends. Initially, the advisor is the primary sender of e-mail. This is consistent with the traditional role of the advisor as providing more direct guidance early in a graduate's student career. During the 6th semester we see a shift where the e-mail is entirely sent by the student. After the 6th semester, it appears that the total volume of e-mail is more equally sent by both parties. We aggregate the numbers to examine whether this is the case. The aggregate numbers are given in Table 1. We see indeed that there is a massive shift from the advisor sending 2:1 mails before the 6th semester to the student sending slightly more in aggregate in the period starting with the sixth semester. Interestingly, it was during semester 5-6 that the student defined the key direction for his dissertation. We see here then digital evidence of the emergence of a peer relationship – where both sides seem to be contributing equally – around a key transition in research. At the end of the time at the university we see a notable increase from both the student and advisor. During this period of time the student and advisor were conducting a variety of new projects that benefited from their previous work together.

[1]Since the reported time to PhD graduation at the university is often said to be 6.5 years, we can also clearly conclude that the student is above average.

Figure 4: Count of e-mails sent between a student-advisor pair aggregated by academic semester. Due to data loss a major time gap not represented in the e-mail archive occurs between the columns denoted as 11 and 12. The first 11 columns are in fact the first 11 semesters (*i.e.*, 5.5 years) of the student's career as a graduate student.

Period	From Student	From Advisor
Total	69	80
Before Semester 6	21	40
Semester 6 & After	48	40

Table 1: Count of e-mails originating from the student or advisor overall and relative to the transition from the pivotal semester 5 to 6 where the student's major dissertation thrust was defined.

Finally, we can analyze the trend in word usage over time as well. In particular, after removing words that are globally frequent in English,[2] we looked at the rank correlation of word frequency with time for all words that occurred in at least three calendar years. We did this to identify words whose usage increases with time or decreases. We note two interesting cases. The top word from e-mails sent by the advisor that decreases in frequency is "progress" (Spearman's $\rho = -0.83$). The concerned reader may hypothesize that the advisor ceased to use the word "progress" having given up hope that the student would make any. However, examination of the e-mail corpus reveals that in early years the advisor was both characterizing the student's progress for fellowship applications and in promoting group discussions whose goal was to help shape direction. As the student's direction became

[2]We used the top 1K words of Bing's web n-gram list (Microsoft, 2010) as a stopword list.

set, the word disappears from the corpus at a similar timepoint to the transition in frequency of e-mail. Likewise, the word in e-mails sent by the student that *increases* the most in frequency is "attached" (Spearman's $\rho = 0.68$). In this case, examination of the corpus reveals the natural hypothesis is correct: the student increasingly sends attachments consisting of write-ups, papers, and results as time progresses. Again, this word begins appearing in the corpus near the same critical point in time. Thus, the textual analysis lends support to the frequency analysis that a key shift in direction and the relationship occurred around the same point in time.

3 Conclusion

In this paper, we considered e-mail centric technology directions. In particular, we discuss and conduct a case study of e-mail archives as a valuable source of introspective data. We primarily considered very broad temporal trends in this paper, but obviously richer tools could be used to perform a deeper textual analysis as well. For example, studies of topic specific evolution in individual relationships, peer-peer, and subordinate-supervisor relationships could prove interesting. This would extend work such as Danescu-Niculescu-Mizil's *et al.* study of power relationships and linguistic accommodation in public Wikipedia logs and Supreme Court transcripts (Danescu-Niculescu-Mizil et al., 2012) to private domains as well as to cases where the relationship may vary over time. Finally, as another measure of success, when presented with the analysis the student was not only interested in the results, but found the data-derived summary provided a more complete picture than memory – perhaps because the research-focused mind may sometimes forget to take note of all of the little data points that add up to big trends.

References

P.N. Bennett and J. Carbonell. 2005a. Detecting action-items in email. In *In Proceedings of the 28th Annual International ACM SIGIR Conference.*

P.N. Bennett and J. Carbonell. 2005b. Feature representation for effective action-item detection. In *In Beyond Bag-of-Words Workshop at the 28th Annual International ACM SIGIR Conference.*

P.N. Bennett and J. Carbonell. 2007. Combining probability-based rankers for action-item detection. In *In Proceedings of HLT-NAACL 2007.*

Vitor R. Carvalho. 2008. *Modeling Intention in Email.* Ph.D. thesis, Carnegie Mellon University. CMU-LTI-08-007.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708.

M. Freed, J. Carbonell, G. Gordon, J. Hayes, B. Myers, D. Siewiorek, S. Smith, A. Steinfeld, and A Tomasic. 2008. RADAR: A personal assistant that learns to reduce email overload. In *Proceedings of the 23rd American Association of Artificial Intelligence Conference.*

Microsoft. 2010. Microsoft web n-gram service (public beta). http://web-ngram.research.microsoft.com/info/ (Retrieved March, 2014).

Stephen Wolfram. 2012. The personal analytics of my life. http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/.