An Exploratory Study into Automated Real-Time Categorisation of Engineering E-Mail

James A. Gopsill
Dept. of Mechanical Engineering
University of Bath
Bath, United Kingdom

Stephen J. Payne
Dept. of Computer Science
University of Bath
Bath, United Kingdom

Ben J. Hicks
Dept. of Mechanical Engineering
University of Bristol
Bristol, United Kingdom

***Abstract*—For large, spatially and temporally distributed engineering projects, e-mail is a central means for the discussion of engineering work and sharing of digital assets that define the product and its production process. The importance of communication and the value of its content for resolving issues *post facto* are universally accepted. More recently, the potential value of its content to predict events, issues and states *a priori* has been explored with some success. However, while in the former context (*post facto*) trends and patterns can be established through iteration and refinement over time; for prediction, heuristics need to be established in advance and closer to real-time analysis becomes necessary due to the critical and very often short timescales. It is this challenge of making predictions from the content of e-mail that is considered in this paper. In particular, the paper deals with engineering e-mail and the ability to automatically predict its purpose from its content rather than relying solely on the subject line.**

The work builds upon previous studies by the authors concerning the characterisation of the content of e-mail: what they are about, why they were sent and how the content is expressed. The paper summarises the previous work and looks at the potential of identifying the purpose of e-mail through the use of Naive Bayes and an adapted Latent Semantic Analysis approach. While the techniques have only been applied to an initial exploratory study of 98 e-mails, the results suggest the potential for automated real-time categorisation of engineering e-mails through achieving an accuracy of 66%. Such a capability would both support prioritisation of e-mail for engineers and macro level characterisation of project e-mail dynamics. The latter provides the opportunity for real-time analysis of an engineering projects status and correspondingly, modes of management intervention.

***Keywords*—Engineering Communication, E-Mail, Naive Bayes, Latent Semantic Analysis.**

I. Introduction

It is self-evident that e-mail has become a central means for the discussion of engineering work and sharing of digital assets[1] that define the product and its production process [1]. This is especially the case when teams become larger, increasingly multi-disciplinary and more distributed both spatially and temporally [2]. Delinchant et al. [3] argues that the prominence of e-mail is due to engineering companies offering support for the communication tool and its ubiquity across the engineering domain.

[1]Examples include: reports, calculations, photographs and results from simulations.

978-1-4799-0652-9/13/$31.00 2013 Crown

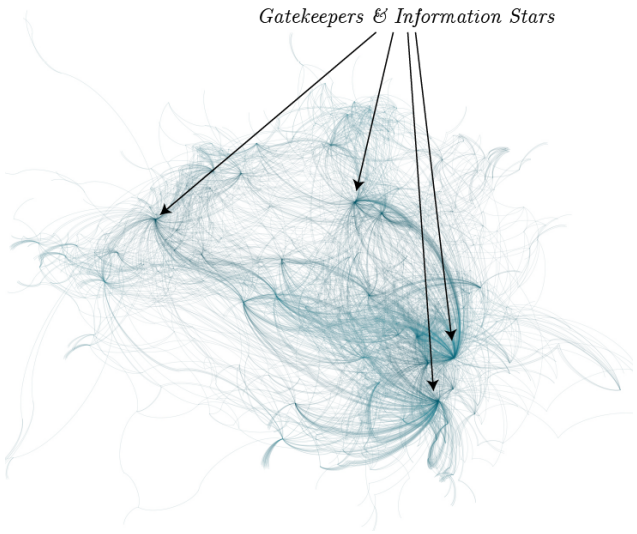

Fig. 1. Visualisation of an Engineering E-Mail Network

Engineering communication research has shown that the volume of communication is indicative of progress being made within an engineering project [4], [5]. In addition, Dong [6] reveals that almost all successful design teams have high-levels of communication as this helps maintain a shared understanding between the engineers. Although it may seem a positive step to encourage increased communication between engineers, there are a number of limitations of e-mail that need to be addressed both from Personal Information Management and Project Management perspectives.

A. Personal Information Management

Engineers typically have to send e-mails through a hierarchy of personnel before being able to reach the *right* people to share knowledge with [7], [8]. Engineering projects also contain key figurehead/expert engineers who are the *'go to people'* within the project. These are often referred to in the literature as *gatekeepers* or information *stars* as they fill one of two roles; 1) to know 'who knows' and therefore direct engineers to the relevant expert or 2) are experts in a particular field themselves [9], [10]. If one were to visualise the network of communication within an engineering project, it would appear similar to the visualisation in Figure 1, which highlights four *gatekeepers* from a project involving approximately 670 people

and an e-mail corpus containing 10,000 e-mails. Therefore, by increasing the communication activity within the network, the *gatekeepers* and information *stars* would likely start to receive an overwhelming number of e-mails, leading to the potential issue of information overload [11], [12].

Dealing with this overload is not always aided by the subject line, which may offer little or no contextual information with regards to the purpose of the e-mail. Rather, it often relates the e-mail to project dimensions (for example, Report on Part X) [13]. This leads to the engineers having to read the full extent of the e-mail to elicit the purpose and thereby determine how and when they should respond. Previous research has shown that there are a number of engineering purposes (for example, presenting an idea or asking for clarification) for sending e-mail. These are in addition to reasons such as Project, Customer and Supplier Management [14]–[17]. For these reasons, it is argued that the ability to categorise e-mails against their purpose in real-time would aid engineers in better managing their e-mail communications and task management [18].

B. Project Management

As previously stated, the volume of communication is indicative of progress being made within a project [4], [5]. Further, patterns of e-mail types (Problem Solving, Information, Management) have been shown to correlate with the project state and issues being experienced although this was performed *post facto* [14], [19]. E-mails form a major part of the explicit knowledge repository of an engineering company as e-mails often contain rationale behind *'why the product is the way it is'*, as well as the decisions made and insights/conclusions drawn from the discussion and aggregation of information [20], [21]. A number of studies have shown that engineers can use as much as 70–95% of previous designs to develop new products, which suggests that there is potential for an engineering e-mail corpus to provide useful information to future projects [10], [22], [23].

Currently, project design rationale capture tools have focused upon argumentative capture [24], [25]. The implementation of these tools has often led to engineers having an increased workload as engineers post-rationalise the design process once the project/task has finished [26]–[28]. Carlile [29] makes the comment that an engineer's rationale is often embedded in practice and therefore it is hard to recall and articulate, thus raising further issues with the current approaches taken to capture design rationale.

Again, it is hypothesised that by being able to automatically categorise e-mails against their purpose in real-time could lead to improved management of engineering projects. This is because the ability to aggregate e-mails based on their purpose could highlight potential signatures that may relate to project states and issues (such as highlighting a key issue in the product design and identifying when best to perform a design review meeting). In addition, it may provide additional insights into levels of shared understanding and identification of expertise amongst engineers.

C. The Research Challenge

Categorisation of text has been successfully applied in many areas such as sentiment analysis of tweets, search engines for web pages/documents and keyword generation using various techniques from decision trees, probabilistic modelling to matrix factorisation [30]–[33]. Therefore, there are challenges in selecting an appropriate technique for the given document dataset and ensuring that noise in the words used in the analysis is reduced effectively. When considering engineering e-mail content, noise arises for a number of reasons:

1) The differing word lengths of the e-mail.
2) The style and formatting used by the engineers.
3) The project phase in which the e-mail was composed in as the engineering context giving way to new terminology alongside terminology used in an alternative manner.
4) The need to remove *stopwords* [33].
5) Colloquialisms, synonyms and specific terminology used within the company.
6) Signatures and confidentiality statements that are commonly applied to the end e-mails.
7) Machine code generated by the e-mail clients used.

Given these characteristics, two techniques are explored, Naive Bayes and an adapted Latent Semantic Analysis approach. The dataset on which this exploratory study has been performed is discussed as well as cleaning of the content of the e-mail. This is followed by the reasoning behind and application of the two techniques. The results and discussion of the two techniques are presented followed by a brief outline of potential future work.

II. The Dataset

This section describes the data that has been captured and how it has been cleaned for use with the two techniques.

A. Data Capture

The e-mails that are to be used for the training and testing of the two techniques have been captured from a Formula Student project at the University of Bath. A team of 34 students were tasked to create a Formula Student race car to compete against other Universities at annual competitions. The students were required to use one of a number of e-mail templates made available to them. Each template labelled the e-mail according to its purpose (Table I) and was then copied to a shared mailbox for analysis. The data capture is currently on going. For this exploratory analysis a subset of 98 e-mails of original intent (i.e. not replies or forwarded e-mails) were used.

Although there are a number of labels available to the students, only 5 had a significant number of e-mails such that training of the techniques could be performed. They were 1) Project Management, 2) Information Request, 3) Idea, 4) Clarification and 5) Observation. These are the labelled e-mails that constitute the 98 e-mails in this analysis and Table II states the number of e-mails in the training and test sets respectively.

B. Data Cleaning

The Natural Language Toolkit (NLTK) was used to aid the cleaning of the e-mail content[2]. The regular expression "\b[a-z']{3,10}\b" was used to gather all words within the e-mail content. This was followed by use of the Porter Stemmer (within NLTKs toolbox) on all the words and then discounting

[2]http://nltk.org/

TABLE I. E-MAIL TAGS THE FORMULA STUDENT TEAM CAN APPLY (LABEL[1] *from* [16], LABEL[2] *from* [1])

E-Mail Label	Description
Idea[1]	Wants to show something potentially new
Help[1]	Wants to solve a process problem
Issue[1]	Wants to solve a product problem
Clarification[1]	Wants to double-check their knowledge on a subject
Observation[1]	Wants to highlight an artefact of potential interest
Confirmation[1]	Wants to ensure the artefact is correct
Comparison[1]	Wants to converge on a solution
Option Generation[1]	Wants to generate a number of solutions
Information Request[1]	Wants to receive information or be provided with its location
Decision[1]	Wants to propose a decision
Project Management[2]	Roles of Responsibility, Deadlines, Meeting Planning & Task/Process Management
Customer Facing[2]	Quotations, Customer Support, Sales and After-Sales
Social[2]	Evening Plans, Talking with Friends and *'the football last night'*

TABLE II. TRAINING AND TEST SETS

E-Mail Label	Train	Test	Total
Project Management	21	4	25
Information Request	24	4	28
Idea	17	3	20
Clarification	10	2	12
Observation	11	2	13
Total:	83	15	98

stopwords from the content using NLTKs *stopwords* reference corpus. In this exploratory analysis, no attempt has been made to remove e-mail signatures or confidentiality clauses from the dataset although it is highlighted that many of the students had minimal signatures and did not use any confidentiality clauses within this dataset.

C. Naive Bayes Classifier

Naive Bayes classifier determines the classification of a document by producing a likelihood estimate of the document being in each label based on the features assigned to the document [33, p. 249]. In this case, the features are the words, the documents are the email content and the labels the purpose of the email. The Naive Bayes classifier is initially trained on a set of e-mails for each label and then tested against the remaining e-mails for each label. This training calculates the probability of the words' existence in relation to that label. Thus, given a new e-mail, the classifier calculates the probability of the e-mail being associated with each label by the summation of the word probabilities.

The classifier makes the assumption that each feature is independent from one another, which may lead to the issue of double-counting. However, given the relatively small length of e-mails this may not be such an issue. Its suitability for this type of content is supported by [34] through their investigation into creating an anti-spam filter.

D. Latent Semantic Analysis

Latent Semantic Analysis (LSA) was originally patented back in 1988 and invented by [35]. Examples of its application are in identifying similarity in documents, marking of exams and indexing of webpages [31], [36]–[39]. The potential of using LSA within engineering documentation has seen some interest, see for example, [6], [40]. A detailed overview of LSA is given by [41] and is summarised in this paper in the context of applying it to this dataset. The process is to generate a matrix W of $m \times n$ where there are m words and n documents through the application of a term frequency-inverse document frequency calculation (tf-idf). Singular Value Decomposition is then applied to W, which generates three matrices $S \sum U^T$ (Equation 1).

$$
\begin{array}{c}
(S) \quad c_1 \quad c_2 \quad \dots \quad c_n \\
\begin{array}{c} w_1 \\ w_2 \\ \vdots \\ w_m \end{array}
\begin{pmatrix} & & & \\ & & & \\ & & & \\ & & & \end{pmatrix} \\
(\sum) \\
\begin{pmatrix} S_1 & 0 & \dots & 0 \\ 0 & S_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & S_n \end{pmatrix} \\
(U^T) \quad doc_1 \quad doc_2 \quad \dots \quad doc_n \\
\begin{array}{c} c_1 \\ c_2 \\ \vdots \\ c_n \end{array}
\begin{pmatrix} & & & \\ & & & \\ & & & \\ & & & \end{pmatrix}
\end{array}
\tag{1}
$$

S is an $m \times m$ matrix but only $m \times n$ is of interest, where there are n underlying *concepts* and the values within the matrix are normalised, $(-1 < i < 1)$ which informs us about the word's influence on that *concept*. $\sum$ is an diagonal matrix of Singular Values and informs us about the *concept's* influence on being able to recreate the original matrix (W). It is common convention for the Singular Values to be listed in descending order and therefore the concepts are ordered in their influence in regenerating the original matrix W. Finally, U^T is again a normalised matrix of size $n \times n$ and informs us about the *concept's* influence on the documents (doc). Once generated, the matrices are often reduced to only consider *concepts* 1 to k (i.e. the most important). Given a new document x, one can identify all the vertices for each word in S and then calculate the centroid for x. This is then compared to the other documents (doc) using the cosine distance with the smaller distance highlighting a greater similarity.

In this case, the aim is to be able to assign a purpose to a given e-mail. Therefore, the W matrix will be of words (w_m) against purposes of communication (p_n) and thus requiring a slight adaptation LSA. The tf-idf calculation used to calculate the word vector is described below (Equation 5).

To calculate the value between a word and purpose of communication ($W_{(m,n)}$), determine the frequency of word (w_m) in e-mail (e_i) and divide by the total number of words within that e-mail (e_i) for each e-mail in the training set of e-mails (E) that corresponds to purpose (p_n):

$$
\sum_{i=1}^{|\{e \in E : p_n \in e\}|} \left(\frac{f(w_m, e_i)}{\sum_{j=1}^{m} f(w_j, e_i)} \right) \tag{2}
$$

This gives us the weighting within that e-mail and all these weightings are summed together and divided by the number of e-mails in the training set that are of purpose (p_n):

$$\frac{\sum_{i=1}^{|\{e \in E : p_n \in e\}|} \left(\frac{f(w_m, e_i)}{\sum_{j=1}^{m} f(w_j, e_i)} \right)}{|\{e \in E : p_n \in e\}|} \tag{3}$$

This is then multiplied by the total number of e-mails in the training set divided by the summed frequency of the word across all the e-mails in the training set:

$$\frac{|E|}{\sum_{k=1}^{|E|} f(w_m, e_k)} \tag{4}$$

Giving $W_{(m,n)} =$

$$\left(\frac{\sum_{i=1}^{|\{e \in E : p_n \in e\}|} \left(\frac{f(w_m, e_i)}{\sum_{j=1}^{m} f(w_j, e_i)} \right)}{|\{e \in E : p_n \in e\}|} \right) \left(\frac{|E|}{\sum_{k=1}^{|E|} f(w_m, e_k)} \right) \tag{5}$$

This is mathematically not too dissimilar to the usual tf-idf calculations performed when comparing documents to documents. Here, the e-mails within the training set that have been labelled as a particular purpose are grouped and effectively treated as a single document. There is only a subtle but important difference in how the inverse document frequency is determined as it based on the total number of e-mails and not on the number of columns that contain a word (w_m), which is usually the total number of documents.

SVD is performed on this matrix to provide us with the three matrices $S \sum U^T$ (Equation 1) although documents doc are replaced with purposes p. Now, it will be the case that there is only a limited number of *concepts* generated (n) and for this case, the Singular Values are to be ignored and comparison between the matrices S and U^T will be made as the Singular Values provide an insight into the *concepts* influence on re-creating the previous matrix. What is of interest here is the subtle differences within the *concepts* that contribute to the purposes of communication. By not including the Singular Values, each *concepts* is treated as equal. Given an e-mail to categorise, the associated word vertices are selected, the centroid defined and compared to the purposes through the cosine distance measure.

III. Results and Discussion

Using the dataset described in section 2, the Naive Bayes classifier attained an accuracy of 40%. Looking at the most informative features (Figure 2, left-hand column) there are four words that could be relevant to an engineering project (make, i'v, system and request) but there remains considerable noise that could be altering the result. The right-hand column describes the main influence in differentiating between the various purposes of the words in the left-hand column, highlighting that there is a large disparity between Observation and Project Management and this is mainly due to words that could constitute noise, most probably from the various e-mail clients being used within the group.

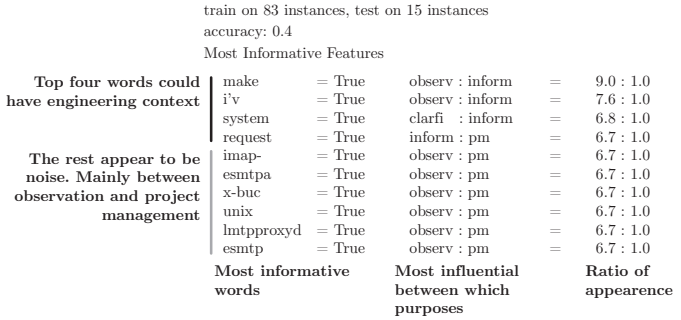

Fig. 2. Most Informative Features from Naive Bayes (Annotated output from NLTK Naive Bayes Classifier)

LSA achieved an accuracy of 66%. In this case, LSA is more effective at categorising e-mails by their purpose. This indicates that there are underlying associations between the words within an e-mail that constitutes it being of a particular purpose. The LSA result is very encouraging when considering that there remains a high-level of noise as the data cleaning performed does not address sources of noise 2, 3, 5, 6 & 7 described in the introduction. Furthermore, it is not too far away from the 80% human accuracy of sentiment analysis [42] and typical inter-coder reliability of 70% for conclusions to be able to be drawn [43], [44].

Figure 3 shows two graphs, the top graph highlighting the value of the Singular Value for each *concept* and the bottom graph showing the influence of each *concept* on each of the purposes. The Singular Values diminish rapidly and if taken into account for the vertices for the cosine distance measurement, the first few dimensions would have a dominant influence on the categorisation of the e-mail. Looking at the bottom graph it can be seen that each *concept* has a distinct influence upon one of the purposes. Thus, this confirms the initial assumption to not consider the Singular Value in the cosine distance measurement and to treat the *concepts* equally to enable clearer differentiation between the purposes.

IV. Future Work

The results are an initial indication that the categorisation of engineering e-mails based on their purpose could be achieved. This is especially so when considering more work can be done to improve the result, mostly involving the reduction in potential noise in the e-mail content. As the e-mail corpus being gathered increases, it may be possible to develop an LSA per engineer, that is, for the analysis to be performed on subsets of the dataset based on the engineers as well as purpose. This would define concepts specific to each engineer and could take into account their own writing style (2). In addition, the development of a custom dictionary for the e-mail corpus would aid in reducing machine code being captured (7) and enable like terminology to be considered the same (5). Further improvements can be made in identifying common features such as signatures, which can be removed from the analysis (6).

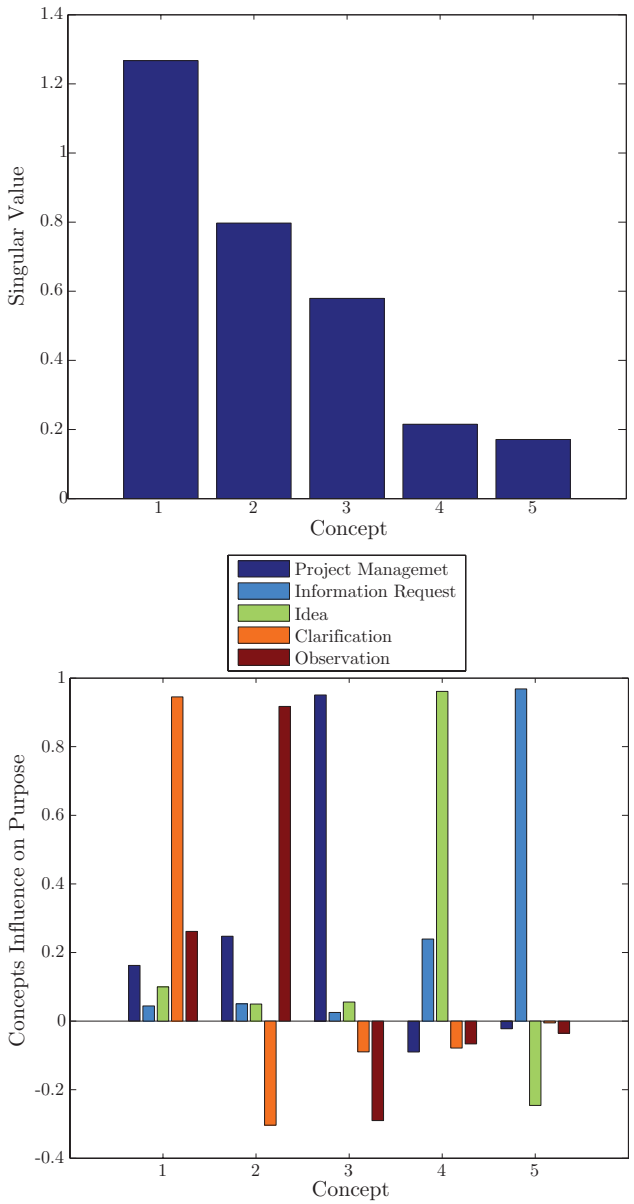

Fig. 3. Singular Values and Influence of the Concepts on the Purposes of Communication

V. Conclusion

This paper has discussed the importance of e-mail as a means of communication within engineering projects and even more so when a project becomes larger, increasingly multi-disciplinary and more distributed both spatially and temporally. Although high volumes of communication are seen to be indicative of a successful project and project progress, it is argued that there may be issues of information overload for engineers. In particular, the engineers classed as *gatekeepers* and information *stars*. E-mails also contain explicit rationale that can inform us *'why it is the way it is'*, yet the large volume of e-mail can present a challenge to aggregate and identify patterns.

Therefore, it has been proposed that being able to identify the purpose of the e-mail in real-time could aid engineers in their own Personal Information Management and aid in the identification of patterns/events within the engineering project leading to improvements in Project Management. To achieve this, an exploratory study using two techniques, Naive Bayes and an adapted Latent Semantic Analysis (LSA) approach, have been tested against a tagged engineering e-mail dataset. The results show Naive Bayes achieving an accuracy of 40% and LSA achieving an accuracy of 66%. This suggests that LSA is a technique better suited for the categorisation of e-mails against their purpose and is nearing comparability to human accuracy (typically 70-80%) despite the minimal data cleaning that was performed. Future work will look to reduce the noise further and at performing LSA on e-mails sent by specific engineers as a means at improving the accuracy of the categorisation.

Acknowledgment

The work reported in this paper has been undertaken as part of the Language of Collaborative Manufacturing Project[3] at the University of Bath & University of Bristol, which has support from the Engineering and Physical Sciences Research Council (EPSRC) (grant reference EP/K014196/1).

References

[1] J. A. Gopsill, H. C. McAlpine, and B. J. Hicks, "The Communication Patterns of Engineers within an SME 2012," in *International Conference on Engineering Design ICED'13*, 2013.

[2] J. D. Herbsleb and A. Mockus, "An empirical study of speed and communication in globally distributed software development," *Software Engineering, IEEE Transactions on*, vol. 29, no. 6, pp. 481–494, Jun. 2003.

[3] B. Delinchant, V. Riboulet, L. Gerbaud, P. Marin, F. Noel, and F. Wurtz, "E-cooperative design among mechanical and electrical engineers: implications for communication between professional cultures," *Professional Communication, IEEE Transactions on*, vol. 45, no. 4, pp. 231–249, Dec. 2002.

[4] J. Liebowitz and K. Wright, "Does measuring knowledge make 'cents'?" *Expert Systems with Applications*, vol. 17, no. 2, pp. 99–103, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417499000275

[5] A. Griffin and J. R. Hauser, "Patterns of Communication among Marketing, Engineering and Manufacturing-A Comparison between Two New Product Teams," *Management Science*, vol. 38, no. 3, pp. pp. 360–373, 1992. [Online]. Available: http://www.jstor.org/stable/2632480

[6] A. Dong, "The latent semantic approach to studying design team communication," *Design Studies*, vol. 26, no. 5, pp. 445–461, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0142694X05000050

[7] M.-L. Chiu, "An organizational view of design communication in design collaboration," *Design Studies*, vol. 23, no. 2, pp. 187–210, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0142694X01000199

[8] A. W. Court, "The modelling and classification of information for engineering designers," Ph.D. dissertation, Department of Mechanical Engineering, 1995.

[9] C. Tenopir and D. W. King, *Communication Patterns of Engineers*. Wiley-IEEE Computer Society Pr, 2004.

[10] C. Eckert, P. J. Clarkson, and M. Stacey, "Information flow in engineering companies: problems and their causes," *Design Management: Process and Information Issues*, vol. 28, p. 43, 2001.

[3] http://www.locm.org.uk

[11] M. J. Eppler and J. Mengis, "The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines," *The Information Society*, vol. 20, no. 5, pp. 325–344, 2004. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01972240490507974

[12] L. A. Dabbish and R. E. Kraut, "Email overload at work: an analysis of factors associated with email strain," in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, ser. CSCW '06. New York, NY, USA: ACM, 2006, pp. 431–440. [Online]. Available: http://doi.acm.org/10.1145/1180875.1180941

[13] B. J. Hicks, A. Dong, R. Palmer, and H. C. McAlpine, "Organizing and managing personal electronic files: A mechanical engineer's perspective," *ACM Trans. Inf. Syst.*, vol. 26, no. 4, pp. 23:1—-23:40, Oct. 2008. [Online]. Available: http://doi.acm.org/10.1145/1402256.1402262

[14] J. Wasiak, B. Hicks, L. Newnes, A. Dong, and L. Burrow, "Understanding engineering email: the development of a taxonomy for identifying and classifying engineering work," *Research in Engineering Design*, vol. 21, no. 1, pp. 43–64, 2010. [Online]. Available: http://dx.doi.org/10.1007/s00163-009-0075-4

[15] N. A. M. Maiden and B. P. Bright, "Recurrent communication patterns in requirements engineering meetings," in *Enabling Technologies: Infrastructure for Collaborative Enterprises, 1996. Proceedings of the 5th Workshop on*, Jun. 1996, pp. 208–213.

[16] J. A. Gopsill, H. C. McAlpine, and B. J. Hicks, "A Social Media Framework to Support Engineering Design Communication and its Reuse," *Journal of Advanced Engineering Informatics*, 2013.

[17] ——, "Meeting the Requirements for Supporting Engineering Design Communication - PartBook," in *International Conference on Engineering Design ICED'13*, 2013.

[18] J. Wasiak, B. Hicks, L. Newnes, C. Loftus, A. Dong, and L. Burrow, "Managing by E-Mail: What E-mail Can Do for Engineering Project Management," *Engineering Management, IEEE Transactions on*, vol. 58, no. 3, pp. 445–456, Aug. 2011. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5738325

[19] J. O. Wasiak, "A Content Based Approach for Investigating the Role and Use of E-Mail in Engineering Design Projects," Ph.D. dissertation, Department of Mechanical Engineering, University of Bath, 2010. [Online]. Available: http://opus.bath.ac.uk/21116/1/UnivBath_PhD_2010_J_Wasiak.pdf

[20] W. C. Regli, X. Hu, M. Atwood, and W. Sun, "A Survey of Design Rationale Systems: Approaches, Representation, Capture and Retrieval," *Engineering with Computers*, vol. 16, no. 3, pp. 209–235, 2000. [Online]. Available: http://dx.doi.org/10.1007/PL00013715

[21] G. Huet, S. J. Culley, C. A. McMahon, and C. Fortin, "Making sense of engineering design review activities," *Artif. Intell. Eng. Des. Anal. Manuf.*, vol. 21, no. 3, pp. 243–266, Jun. 2007. [Online]. Available: http://dx.doi.org/10.1017/S0890060407000261

[22] L. Freund, E. G. Toms, and J. Waterhouse, "Modeling the information behaviour of software engineers using a work - task framework," *Proceedings of the American Society for Information Science and Technology*, vol. 42, no. 1, pp. n/a—-n/a, 2005.

[23] G. Vijaykumar and A. Chakrabarti, "Understanding the Knowledge Needs of Designers During Design Process in Industry," *Journal of Computing and Information Science in Engineering*, vol. 8, no. 1, p. 11004, 2008. [Online]. Available: http://link.aip.org/link/?CIS/8/011004/1

[24] F. M. Shipman and R. J. McCall, "Integrating different perspectives on design rationale: Supporting the emergence of design rationale from design communication," *Artificial Intelligence for Engineering, Design, Analysis and Manufacturing*, vol. 11, no. 02, pp. 141–154, 1997. [Online]. Available: http://dx.doi.org/10.1017/S089006040000192X

[25] M. Klein, "Capturing design rationale in concurrent engineering teams," *Computer*, vol. 26, no. 1, pp. 39–47, 1993.

[26] R. H. Bracewell, S. Ahmed, and K. M. Wallace, "DRed and design folders: a way of capturing, storing and passing on-knowledge generated during design projects," *ASME International Design Engineering Technical Conferences, IDETC'04*, 2004.

[27] R. Bracewell, K. Wallace, M. Moss, and D. Knott, "Capturing design rationale," *Computer-Aided Design*, vol. 41, no. 3, pp. 173–186, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010448508001899

[28] Y. Zhang, X. Luo, J. Li, and J. J. Buis, "A semantic representation model for design rationale of products," *Advanced Engineering Informatics*, no. 0, pp. –, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474034612000985

[29] P. R. Carlile, "A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development," *Organization Science*, vol. 13, no. 4, pp. pp. 442–455, 2002. [Online]. Available: http://www.jstor.org/stable/3085976

[30] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 1195–1198. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753504

[31] S. T. Dumais, "Latent semantic analysis," *Annual Review of Information Science and Technology*, vol. 38, no. 1, pp. 188–230, 2004. [Online]. Available: http://dx.doi.org/10.1002/aris.1440380105

[32] M. Dredze, H. M. Wallach, D. Puller, and F. Pereira, "Generating summary keywords for emails using topics," in *Proceedings of the 13th international conference on Intelligent user interfaces*, ser. IUI '08. New York, NY, USA: ACM, 2008, pp. 199–206.

[33] S. Bird, E. Klien, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009. [Online]. Available: http://nltk.org/book/

[34] I. Androutsopoulos, J. Koutsias, K. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An evaluation of Naive Bayesian anti-spam filtering," *CoRR*, vol. cs.CL/0006, 2000.

[35] S. Deerwester, S. T. Dumais, G. W. Furnas, R. A. Harshman, T. K. Landauer, and K. E. Lochbaum, "Computer information retrieval using latent semantic structure," 1988.

[36] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9

[37] D. T. Haley, P. Thomas, A. De Roeck, and M. Petre, "Measuring improvement in latent semantic analysis-based marking systems: using a computer to mark questions about HTML," in *Proceedings of the ninth Australasian conference on Computing education - Volume 66*, ser. ACE '07. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2007, pp. 35–42. [Online]. Available: http://dl.acm.org/citation.cfm?id=1273672.1273677

[38] T. Miller, "Essay assessment with latent semantic analysis," *Journal of Educational Computing Research*, vol. 29, no. 4, pp. 495–512, 2003.

[39] T. K. Landauer, "Automatic Essay Assessment," *Assessment in Education: Principles, Policy & Practice*, vol. 10, no. 3, pp. 295–308, 2003. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/0969594032000148154

[40] K. Fu, J. Cagan, and K. Kotovsky, "A Methodology for Discovering Structure in Design Databases," in *International Conference on Engineering Design ICED'11*, 2011.

[41] A. Thomo, "Latent Semantic Analysis (Tutorial)," 2009. [Online]. Available: http://www.engr.uvic.ca/~seng474/svd.pdf

[42] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 347–354.

[43] L. T. M. Blessing and A. Chakrabarti, *DRM, a Design Research Methodolgy*, 1st ed. Springer, 2009.

[44] C. M. Snider, E. A. Dekoninck, and S. J. Culley, "Improving confidence in smaller data sets through methodology - the development of a coding scheme," in *DESIGN 2012*, 2012.