K. Kuutti, E. H. Karsten, G. Fitzpatrick, P. Dourish, and K. Schmidt (eds.). *Proceedings of the Eighth European Conference on Computer-Supported Cooperative Work, 14-18 September 2003, Helsinki, Finland*

Discovery of implicit and explicit connections between people using email utterance

Robert McArthur and Peter Bruza
Distributed Systems Technology Centre, Brisbane, Australia
{mcarthur,bruza}@dstc.edu.au

Abstract. This paper is about finding explicit and implicit connections between people by mining semantic associations from their email communications. Following from a socio-cognitive stance, we propose a model called HALe which automatically derives dimensional representations of words in a high dimensional context space from an email corpus. These dimensional representations are used to discover a network of people based on a seed contextual description. Such a network represents useful connections between people not easily achievable by 'normal' retrieval means. Implicit connections are "lifted" by applying latent semantic analysis to the high dimensional context space. The discovery techniques are applied to a substantial corpus of real-life email utterance drawn from a small-to-medium size information technology organization. The techniques are computationally tractable, and evidence is presented that suggests appropriate explicit connections are being brought to light, as well as interesting, and perhaps serendipitous implicit connections. The ultimate goal of such techniques is to bring to light context-sensitive, ephemeral, and often hidden relationships between people, and between people and information, which pervade the enterprise.

Introduction

Our information environment becomes ever more complex, but our ability to comprehend it does not keep pace. As a consequence, the connections between individuals, groups and information become lost, or forgotten, and individuals and groups become ever more isolated. The cost to the individual, and the enterprise, is a lack of awareness. The broad goal of the research reported in this paper is to discover appropriate and perhaps serendipitous connections within a given context thereby promoting the awareness of individuals to their environment: other individuals, groups and information.

It seems established organisational units often consist of informal social networks as much as of permanent formal teams. Like Nardi *et al*. (2000), we feel that “One of the most important resources we share with each other is access to those in our social networks”. Appropriately discovering and sharing context-specific personal networks is an important part of the life of people and knowledge in an organisation. More specifically, this paper is about finding useful and sometimes serendipitous connections between people by mining semantic associations from their email communications. Since using an individual’s email, and therefore their social networks, raises important issues of consent and privacy, these are separately discussed at the end of the paper.

A feature of current email use is its ubiquity: Ducheneaut and Bellotti (2002) commented that “while this [people spending a lot of time on email] may not be surprising for those who collaborate over distance, we have observed that even colleagues having offices next to each other, or sitting *in plain sight* of each other, still use email as a principal communication medium.” “Work objects are easily accessible while communicating over email in a way that they cannot be in most face-to-face encounters.” Their studies show how email exhibits characteristics as “a preferred medium for talking about work…[and] a valuable product of communication: email conversations, often standing as or evolving into work objects themselves.”

While email itself is a vitally important office communications medium that has a history of research (a very useful map of research in email is presented in Ducheneaut (2002)), studies using semantic information are few. Semantic knowledge is important for, as Kimble *et al.* (1998) note, “Linguistic and philosophical research has suggested that the interpretation of utterances depends not on isolated sentences but on the context….Wittgenstein, for example, asserts that we can only make sense of utterances and actions by seeing them within the contexts in which they were uttered or undertaken”. Their work examines a large organisation’s email overload problem, and their conclusions reiterate the importance of context, especially to the perceptions of users about the information which email (and other computer-mediated communication systems) provides.

Previous work on finding people in organisations has used a variety of methods. PeCo (Ogata and Yano, 1998) used email From & To headers to collect acquaintances and relationships between people, and collected ‘expertise’ by keywords extracted from the message text by morphological analysis. A similarity-based matching system completes their expertise management tool. However, those methods cannot use information in the email, such as may occur from person A to person B: “C said that D is too slow”. These references may be to other, unquoted, email messages, or to a discussion held outside the electronic realm, and are not represented in the Subject, From or To lines of the email message.

Schwartz and Wood (1993) mined 1.2M email headers to detect shared interests between people using graph theory. Specialised subgraphs for a person were aggregated to cover all their interests; specific interests could be determined only by starting with a specific set of people known to have that interest. The methods have the advantage that no processing, or even knowledge of, the message text (including Subject) is needed. However, it suffers from a lack of specificity, as does PeCo above, purely because it ignores the text – the context of the message. It also requires a known starting set for any specific interest area.

Kautz *et al.* (1997) used the co-occurrence of names in close proximity in Web pages as evidence of a direct relationship. They state "Searching for a piece of information…thus becomes a matter of searching the social network for an expert on the topic together with a *chain* of personal referrals from the searcher to the expert." Importantly, we agree that

> …manually searching for a referral chain can be a frustrating and time consuming task. One is faced with the trade-off of contacting a large number of individuals at each step, and thus straining both the time and good will of possible respondents, or of contacting a smaller, more focussed set, and thus being more likely to fail to locate an appropriate expert.

However, as Ogata and Yano (1998) respond, "…it may be difficult to find real private networks." Within a small-medium enterprise, the sources are less likely to provide evidence for a social network. For example, internal web pages are more likely to indicate organisational grouping rather than task- or interest-based relationships.

Ackerman and McDonald (1996) sought a surrogate for hallway talk for people seeking help: "Normally, one attempts to examine the documentation or other help sources, and then wanders out into a hallway in search of friendly colleagues." They collected databases of commonly asked questions that grew "organically" – collecting "organisational memory". This type of questioning and seeking is but part of a larger set; by itself it cannot find an answer, or person to answer, a question that is (probably) going to be asked only once and is possibly time-dependent as well. Unfortunately, these harder questions are all too prevalent. Instead of searching for an explicit answer, it often is necessary to search for someone who could answer, or direct to someone else who can.

This paper is about making people's work easier by finding who has the information, knowledge or expertise that will directly help. Other people in the organisation benefit from the sharing, and the person engaged in informing and being informed about the people in an extra-organisational network increasingly need tools to do that important work (Nardi *et al.*, 2000). Following from a socio-cognitive perspective, we propose a model called HALe which automatically derives dimensional representations of words within a high dimensional context space from an email corpus. These dimensional representations are used to discover a network of people based on a seed contextual description. Connections, perhaps explicit, sometimes serendipitous, are made by analysis of the explicit and tacit knowledge captured and represented in the high dimensional space.

The next section describes relevant theories of knowledge and of communicability of knowledge. Techniques for extracting and representing knowledge from email messages are then presented. These are then applied in an experiment with significant real-world data, showing one way of producing links between people based on semantic analysis. A brief discussion of privacy issues and further work concludes the paper.

Knowledge and Information

Nonaka and Takeuchi (1995) wrote a seminal book on organisational knowledge creation. Their knowledge creation spiral is often cited in the literature. They made a clear distinction between tacit knowledge (personal, context-specific, hard to formalise and communicate) and explicit knowledge (transmittable in a formal, systematic language), and their thesis is that "…the key to knowledge creation lies in the mobilization and conversion of tacit knowledge."

Four modes of knowledge conversion are created when tacit and explicit knowledge interact – socialisation, externalisation, combination and internalisation. "These modes are what the individual experiences. They are also the mechanisms by which individual knowledge get articulated and 'amplified' into and throughout the organisation." In this paper our interest is in externalisation, in which the tacit knowledge in the communications of the members of an organisation is made explicit. It is vital for an organisation as "Among the four modes…externalization holds the key to knowledge creation".

To make tacit knowledge explicit, it must first be identified, then represented. The final, full, step of externalisation is one that, as yet, is still squarely in the realm of the human. Computer systems such as the one described in this paper can assist the final step, but it is the human who accomplishes it and internalises it. Electronically, both the identification and representation is very difficult: it is hard enough for people to attempt to identify and articulate tacit knowledge.

We would like to note, in passing, that the relatively recent focus on "knowledge management" (KM) is not achievable without tacit knowledge and externalisation: "Organizational knowledge creation is a continuous and dynamic interaction between tacit and explicit knowledge" (Ibid, p70). While KM is important, it is how knowledge can assist individuals, who in turn create more knowledge, which is the focus of this work.

There exists a shared background in email messages. "Persistent talk [in email] provides the context for the solitary activity of viewing the content to which it relates….However, during our interviews we, in fact, saw many more examples of imprecise references that were immediately understood than long, drawn-out, explicit and literal descriptions or references." "…email conversations are grounded in sufficient mutual understanding to allow very brief, sketchy and implicit references to succeed without posing significant problems in interpretation.” (Ducheneaut and Bellotti, 2002).

We agree with Nonaka and Takeuchi in that “The semantic aspect of information [as against the syntactic] is more important for knowledge creation, as it focuses on conveyed meaning.” Freyd (1983) provides a socio-cognitive frame for a viable communal (i.e., shared or conveyed) knowledge representation:

> “what seems common to most of the main approaches to semantics is an assumption that values of semantics components, or features, are critical to word meaning. What is relevant to shareability theory is that a smaller number of features seem to be used than number of words.” (pp195-6)
>
> “I am arguing that a dimensional structure for representing knowledge is efficient for communicating meaning between individuals. That is, a small dimensional structure with a small number of values on each dimension is argued to be especially shareable, which might explain why such structures are observed.” (pp198-9)

Freyd’s suppositions on the dimensional nature of shared knowledge are compatible with a recent, three-level model of cognition by Gärdenfors (2000). How information is represented in this model varies greatly across three different levels. It is the conceptual level that is of relevance to this work.

Gärdenfors argues that the meanings of words come from conceptual (i.e. dimensional) structures in people’s heads. In addition, he adopts a socio-cognitive position that the meanings emerge from the conceptual structures harboured by individual cognition together with the linguistic power structure within the community. Of significant note is his adoption of Freyd’s supposition: social interactions will constrain these conceptual structures. This has implications for computer-based representations because it may mean that relevant dimensions are not represented, or the value in a dimension may not be weighted sufficiently.

This constriction of the dimensional structure by the individual for social interaction is important. We tend to economise our utterances, for example, by use of anaphora and liberal use of abbreviations made permissible by shared background.

Techniques for extracting knowledge from utterances

Representation

To bridge the gap between cognitive dimensional structures and actual computational representations, we propose using a variant of Hyperspace Analogue to Language (HAL) (Burgess *et al.*, 1998). HAL produces vectorial representations of words in a high dimensional space that seem to correlate with the equivalent human representations. For example, word associations computed on the basis of HAL vectors seem to mimic human word association judgments. HAL is "a model that acquires representations of meaning by capitalizing on large-scale co-occurrence information inherent in the input stream of language". It "…correlated with lexical decision latencies from a word priming task" and "…simulations using HAL accounted for a variety of semantic and associative word priming effects that can be found in the literature…and shed light on the nature of the word relations found in human word-association norm data." In short, HAL vectors seem to be promising computational representations of word meanings from a semiotic-cognitive perspective.

Utterances must be represented for computation so they can be mined. In light of the works of Perry (1997, 1998) and Gärdenfors (2000), and in accord with our semiotic-cognitive stance, we advocate representing words in utterances as dimensional structures. These are the basic carriers of the meaning of the word in question, but in addition, the dimensional structures have pre-semantic (i.e. what is needed to render a syntactic evaluation to an utterance) information embedded.

Prior work (McArthur and Bruza, 2003a and 2003b) has shown benefits of using these structures: pre-semantic information, in the form of part-of-speech, and LSA (see below) for generating post-semantic information by inference. An updated model for the automatic derivation of the dimensional structures from utterances is briefly explained below.

Vector creation

The basic carriers of meaning are the vectorial representations of words in an utterance. These vectors, created by our modified HAL, are input into the mining process.

Part-of-speech (POS)

POS (Part of Speech) tagging is a computationally efficient means of mapping arbitrary tokens into syntactic classes, determining basic linguistic information from a corpus. It is the means by which pre-semantic context can be automatically gleaned from utterances. The technology has matured to achieve high levels of precision. It is gathered by various methods (rule-based, probability-based, and memory-based being most common) all of which add part of speech tags—noun, verb, pronoun etc.—to the original text. LTCHUNK's (Mikheev, 2000; LTCHUNK) POS tagger was used.

Basic anaphora resolution

Anaphora is the co-reference of one expression with an antecedent (*cataphora* is co-reference with a following expression). The antecedent provides the information necessary for interpreting the expression. An example is between the two sentences: "A well-dressed man was speaking. He had a foreign accent." The term “He” in the second sentence is an anaphoric reference to the “well-dressed man” in the first sentence.

Anaphora is common in utterances and in email in particular. We do not attempt full anaphora resolution but implement an extremely basic algorithm: replace references to ‘I’, ‘my’ or ‘me’ with the first name of the sender of the email, and references to ‘you’ or ‘your’ with the first name of the receiver (if there is only one; the ‘Cc:’ metadata was ignored); these elements are part of the email metadata and easily accessible. No other anaphoric references are as easily determined, so terms such as ‘he’, ‘we’, ‘they’, ‘it’ etc. are left unchanged. We adopt this conservative approach as imprecise anaphora resolution would pollute the vector representations of some words with spurious dimensions.

HAL

Thus far, the exposition of information representation has centred largely upon aspects of Perry’s (1998) pre-semantic context. The second level of Perry’s three levels, semantic context, will now be addressed. This involves attaching meaning to syntactic structures.

A human encountering a new concept derives the meaning via an accumulation of experience of the contexts in which the concept appears. This opens the door to “learn” the meaning of a concept through how it appears within the context of other concepts. Following this idea, Burgess *et al.* (1998) developed a representational model of semantic memory called Hyperspace Analogue to Language (HAL), which automatically constructs a dimensional semantic space from a corpus of text. The space comprises high dimensional vector representations for each term in the vocabulary. Briefly, given an *n*-word vocabulary, the HAL space is a *n* x *n* matrix constructed by moving a window of length *l* over the corpus by one word increments ignoring punctuation, sentence and paragraph boundaries. All words within the window are considered as co-occurring with each other with strengths inversely proportional to the distance between them. After traversing the corpus, an accumulated co-occurrence matrix for all the words in a target vocabulary is produced. Note that word pairs in HAL are direction sensitive – the co-occurrence information for words preceding every word and co-occurrence information for words following it are recorded separately by its row and column vectors. By way of illustration, the HAL matrix for the example text “*The effects of spreading pollution on the population of atlantic salmon*” is depicted in Table 1 using a 5 word moving window (*l*=5). An example of reading the matrix would be that the word *spreading* occurs before *on* (is related to) with strength 4 (5 - 1 intervening word in the window).

Our pilot studies revealed that it was not useful to preserve order information, so, for our purposes, the HAL vector of a word is represented by the addition of its row and column vectors. The quality of HAL vectors is influenced by the window size; the longer the window, the higher the chance of representing spurious associations between terms. Burgess *et al*. used a size of ten in their studies (Ibid). In addition, it is sometimes useful to identify the so-called *quality properties* of a HAL-vector. Quality properties are identified as those dimensions in the HAL vector which are above a certain threshold (e.g., above the average weight within that vector).

	the	eff	of	spr	poll	On	pop	Atl	sal
The		1	2	3	4	5			
eff	5								
of	8	5		1	2	3	5		
spr	3	4	5						
poll	2	3	4	5					
on	1	2	3	**4**	5				
pop	5		1	2	3	4			
atl	3		5		1	2	4		
sal	2		4			1	3	5	

Table 1: Example of a HAL matrix

Developing dimensional representations of words in emails involves unique challenges, so we modified HAL accordingly. To distinguish Burgess *et al*.'s HAL and our model, we describe our model as HALe (for *e*mail) from this point.

The differences between HAL and HALe are:

1. HAL slides a fixed window across the text with all terms used: semantic information is not used. We use a smaller window (8 versus 10), and only terms tagged by POS as nouns (‘NN*’). This is based on earlier trials using verbs (‘VB*’) and adjectives (‘JJ”) as well as nouns. Also, nouns are of most interest as all people were tagged as nouns. A simple stop-word removal would not have the same effect, hence our use of POS tagging.
2. The strength of the association between terms is inversely related to their direct distance apart in HAL. Since HALe effectively ignores some terms (see above), it also ignores them as far as strength of association is concerned. So a noun followed by determiner followed by a noun would have the noun-noun association as though the determiner was invisible.
3. The sender and receiver (if there is only one) of the message is weakly associated with every word in the email. The rationale is that the author and receiver of the message should be associated with the communication. Often, no mention of their names occurs in the message itself (ignoring the signature), even with our anaphora resolution, so HAL on the message itself would not associate the people correctly.

In previous work (McArthur and Bruza, 2003a) a non-linear weighting function was used based on identifying syntactic structures (trees) using a shallow natural language parsing technique. We did not make use of such structures in this work since email messages tend to produce trees lacking branches, hence a loss of expressivity.

LSA

Until now, the creation of a high dimensional space, while interesting, is not necessarily better for the discovery of useful associations between people given a certain contextual description. A simple search using, for example, Microsoft Outlook®, could search for authors in the 'From' field and some text in the main body of the email. While the recall and precision of such searches may not be optimal, it is possible to retrieve 'information' and, with more work on the part of the user scanning many messages, perhaps even the same 'knowledge'. Latent Semantic Analaysis (LSA) (Landauer *et al.*, 1998) is a technique through which implicit associations between words that did not exist can be brought to light. Therefore, it is an interesting candidate for uncovering serendipitous associations between the names of people mentioned in emails in relation to a given contextual description (normally expressed by a few keywords). LSA represents the meaning of words as vectors in a dimensional space reduced by singular value decomposition (SVD). The meaning can be considered "as a kind of average of the meaning of all the passages in which it appears and the meaning of a passage as a kind of average of the meaning of all the words it contains" (p261).

The role of SVD is fundamental to LSA. The general claim is that similarities between words can be more reliably estimated in the reduced dimensional space than in the original one. The rationale is that contexts which share frequently co-occurring terms will have a similar representation in the reduced dimensional space, even if they have no terms in common.

For our purposes, the input to the LSA process is the *n* x *n* matrix produced by HALe. We did not normalize the values in the matrix as advocated in (Landauer, *et al.*, 1998) because pilot studies revealed a 6-9% improvement using un-normalized values. This is perhaps due to the smaller data set, but may also be due to the pre-semantic and semantic processing before HALe.

After dimensional reduction, the weight (*i*,*j*) may be non-zero, whereas it was zero before dimensional reduction. Where positive, it suggests that word *i* is implied within context *j*. In other words, LSA can discover implicit associations, or strengthen/weaken existing associations. Such behaviour is relevant for the mining of useful or serendipitous associations: those associations that appear after dimensional reduction, or are strengthened by it, may be suggestive of post-semantic context. Due to space constraints, the dimensional reduction process will not be described further; the details can be found in (Landauer *et al.*, 1998).

Connections: an Example System

Consider the situation in which Naomi is writing a company's annual report. She's interested in the highlights of the year for the 'Guidebeam' product. Guidebeam, created and developed by Peter, has been worked on by many people including Robert. For example, Rupert does all the business-development. Although Naomi works at the desk next to Rupert, and often socialises in the coffee room, she has forgotten that Rupert is involved. She emails Peter asking him to describe the highlights of Guidebeam's year. In truth, Naomi is more immediately interested in the business highlights than the development highlights, so it would be better to ask Rupert rather than Peter.

Naomi, like most people, is blessed with the ability to forget what she perceives to be unimportant, or at least to forget that she knows it. This is a feature of what some have termed "cognitive economy" (Gabbay and Woods, 2000). It may be behind the change in discourse structure brought about by information overload (Jones *et al*., 2001). Having forgotten Rupert's involvement, the information need Naomi has is primarily "tell me the highlights of Guidebeam in 2002". However, there is no-one or no system to ask such a question of, so her next question is "tell me who I can contact *now* to satisfy my original question."

The remainder of this paper describes a system that was built so that Naomi could answer such a question should, say, Rupert and Peter not be available (on holiday, perhaps not with the organisation anymore, or even deceased). These situations tax both the asker and receiver of questions: Peter would need to spend some time and effort determining the reason behind Naomi's request, and then answering it as best he could. It is a high cost solution.

Tacit knowledge extracted from the email utterances of Rupert, Peter and Robert assists Naomi to answer the 'who' question. It also shows the rich detail of explicit and implicit relationships that can support Naomi becoming more informed about Guidebeam and promote the internalising of the tacit knowledge from the relevant associations.

Data

The base data used for these experiments came from 14,424 email messages from the 'sent' and work-related folders of three individuals at our small-medium organisation: Peter, Robert and Rupert (see Table 2). Only messages from 2001 and 2002 were used since in our scenario, Naomi's need is for recent information. No attempt to separate private and work communications was made on the sent messages. Categories used by the individuals to store their email were not used: data was simply concatenated together into a single large email folder for each person. Not all messages relating to work have been kept by any of the three people. This is reflective of the usual state of practice in the real-world of email.

No standard test set of electronic mail exists (to our knowledge), since organisations, groups and individuals are reluctant to have private email made public. It behoves us to note, therefore, that the experiments cannot be independently verified as we also cannot make available the raw data. We strongly believe that any similar data will generate results that do not differ substantially from those shown here.

	Peter	**Robert**	**Rupert**	**Total**
2001 messages	710	1775	521	3006
2002 messages	1713	1190	1691	4594
2001 and 2002	2423 (528)	2965 (602)	2212 (430)	7600 (1560)

Table 2: Email data statistics

Numbers in brackets are "Usable" messages: those not from mailing lists (ignoring messages in the 'sent' folder where the receiver was the originator of the email), and where the unquoted section of the body of the message had less than 150 lines. The latter condition being used to eliminate large text documents as the primary focus was the direct communication as originally authored in the email.

Method

The email messages were manipulated in the following ways and order:

- Text pre-processing
  - Messages from mailing lists were discarded, as the focus was messages internal to the organisation
  - URI's were replaced with the term 'URL' otherwise the POS tagger erroneously separated out the components and tried to tag them
  - Shorthand words using quotes were expanded for the POS tagger: 'll (will), 've (have), I'm (I am), 're (are), and some words ending in 's (is)
  - Parts of messages whose MIME tag was not 'text/plain' were discarded
  - The identifying name of the sender and receiver of the message was made consistent where the change was unambiguous over the whole set of messages. For example, some messages to Robert were sent to '<xxx@yyy>' or 'Rob <xxxx@yyy>', while a large majority were sent to 'Robert <xxx@yyy>'; all messages of the first two kinds were modified to be the same as the latter in such cases.
  - Quotations of other messages were deleted since they usually occurred within the email collection already. Again, the focus was in the words the author was using for that particular message, rather than the context surrounding the message. Lewis and Knowles (1997) demonstrated that, in finding parents of messages to be threaded, words in the quoted

    part of emails are reiterated enough in the non-quoted section(s) to provide a useful level of similarity between the messages. We anticipate examining whether incorporating quotations between messages leads to any improvements.
  - Signatures and other trailing ‘garbage’ were deleted
- POS tagging
  - POS tagging was applied to all remaining message parts
- Anaphora resolution
  - Simple anaphora resolution changed words ‘I’, ‘my’, ‘me’, ‘you’, ‘your’ and ‘yourself’ in the POS’ed text. In some cases no first name of author and/or receiver could be determined so the term was left unchanged so as not to pollute the vector space. For example, in Robert’s total usable email, the count of anaphoric references by the ‘PRP’ POS tag is shown in Table 3 (bold indicates terms we resolved). Almost 50% of PRP-tagged anaphora was resolvable.

Word	Frequency	%
I	**1299**	**18.1**
you	**1270**	**17.7**
It	1137	15.8
we	1047	14.6
your	**557**	**7.8**
they	272	3.8
our	263	3.7
Me	**245**	**3.4**
My	**198**	**2.8**
their	177	2.5
…	709	9.8
Total	7174	100

Table 3: Anaphora references (Robert’s email, all usable messages)

- HALe
  - HALe was performed on the tagged messages and used to uncover explicit connections between people
- LSA
  - LSA was performed on the set of vectors produced by HALe to uncover tacit or implicit connections by creating new associations, or strengthening and weakening existing ones

Results

Table 4 shows some statistics of the email corpus used as input to HALe.

	Robert	Peter	Rupert	Total
Total words in the POS-tagged messages	49035	91788	37315	178138
Total HALe unique words	2623	1818	3010	5052
Total HALe unique words accepted by Unix spell	1750 (67%)	1203 (66%)	1898 (63%)	3043 (60%)

Table 4: Number of vectors produced by HALe

As a sample of the type of associations and explicit knowledge that HALe produces, Table 5 shows a single (un-normalised) vector for "guidebeam" from the combined emails of Peter, Rupert and Robert. People's names are in boldface.

peter:199, guidebeam:148, search:126, **rupert:114**, technology:114, installation:87, url:62, us:61, **robert:61**, com:60, government:59, query:55, management:54, categories:53, dstc:53, www.:51, engine:51, meeting:46, component:45, information:45, boeing:44, tool:42, catch-up:42, kernel:42, system:41, philosophy:39, p:39, knowledge:39, abc:37, chic:35, people:34, yp:33, panoptic:32, user:31, zen:30, **paul:31**, re:29, think:28, need:28, licence:28, **naomi:28**, partners:27, minutes:25, words:24, package:24, terms:23, project:23, gbst:23, ability:23, specs:23, citr:23, media:23, **justin:23**, @noptic:22, **yvonne:22**, **rob:22**, term:22, examiners:21, we:21, base:21, pb:21, agencies:21, acquiring:21, use:21, actions:21, team:21, +panoptic:21, google:21, and:21, data:21, way:21, web-based:21, libraries:20, time:20, feedback:20, review:20,site:20, recommendations:20, awards:20, article:20, keyword:19, development:19, yahoo:19, intranet:19, week:18, reformulation:18, anything:18, presentation:18, dll:18, proposal:18, **dave:17**, queries:17, results:17, capex:16, cheers:16, test:16,tuesday:16, box:15, **simon:15**, health:15, portal:15, doc:15, fee:15, niche:15, business:15, hib:14, ideas:14, context:14,web:14, file:14, website:14, work:14, idea:14, problem:14, organisation:14, advantage:14, view:14, access:14, talking:14, title:14, log-files:14, reason:14, …

Table 5: "guidebeam" vector produced by HALe

Observe that not all of the associations embedded in the above vector make sense to a wide audience, nor should they, as they are associations relevant within a certain context[1]. Nevertheless, some associations are clearly understandable to anyone (from Table 5: guidebeam is probably a search technology, with some relationship with government, and does queries or management using categories – all true); some relationships require general or specific domain knowledge ('abc', 'gbst' and 'citr' are organisations, p@noptic is a search engine); while some such as "dll" require specific knowledge probably only available to the owner of the email. Finally, serendipitous associations, one of the rationales behind our techniques, can be uncovered (see the discussion of "yukio" in the next section).

[1] For this reason, the developers of HAL refer to the matrix produced by HAL as a "high dimensional context space" (Burgess *et al.* 1998)

As Naomi is specifically interested in people associated with Guidebeam, Tables 6 describes the highest weighted associations for “guidebeam” from each person’s email, and the combination of all three email sets, along with the same vectors after applying LSA. Table 7 shows more detail of the changes wrought by LSA in uncovering implicit associations for “guidebeam”.

Analysis and Discussion

Let us return to the scenario where Naomi wishes to find out who are the major people involved in the Guidebeam product for her marketing report. Figure 1 is a network representation of some of the data from Table 6 (for clarity, the weights of the connections are not shown). It depicts a network of people surrounding the context description “Guidebeam”. The people who are most highly connected to the context – Peter (199), Rupert (114) and Robert (61) – are identified from the combined HALe vector for “guidebeam”. They are the highest weighted people, and they are also internal to the organisation and fortuitously have their email available. It is to this full organisational information space that Naomi would initially come, and the results could be immediately presented.

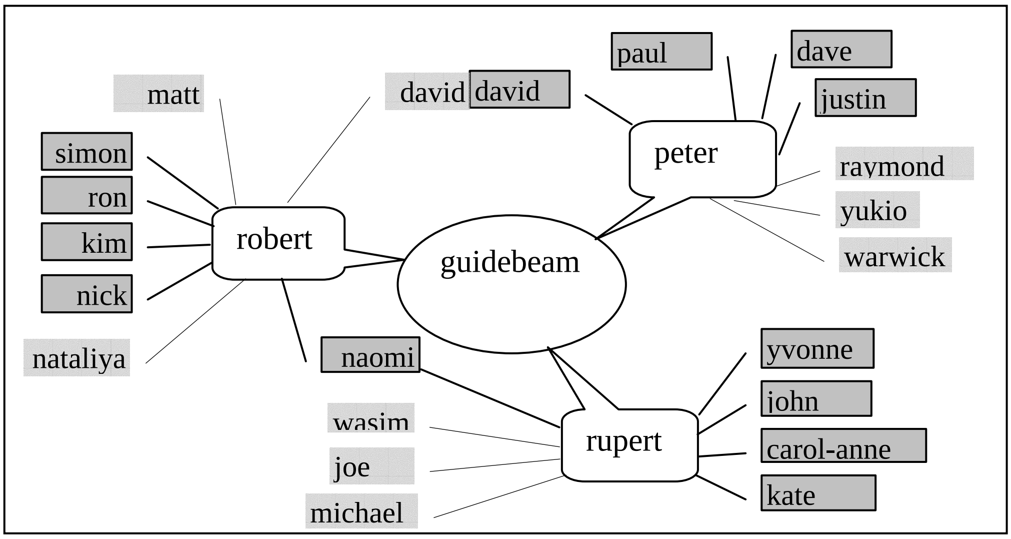

Figure 1: Context-sensitive network of people (Context = "guidebeam")

Radiating from these three key people are others associated with “guidebeam”. Those shown with a dark-grey enclosed box ▨ are derived in the same way as the key people, except they are taken from the “guidebeam” vectors of the respective person’s email: that is, from a localised space not from the global (combined) email space. For example, “naomi” can be seen in the HALe “guidebeam” vectors of both Robert (16) and Rupert (12).

Rupert				Peter				Robert				Combined			
HALe		LSA (6)		HALe		LSA (6)		HALe		LSA (6)		HALe		LSA (6)	
peter	105	rupert	75	peter	76	Peter	77	robert	28	robert	30	peter	199	peter	202
rupert	66	wasim	26	paul	31	Bruza	17	rupert	25	peter	28	rupert	114	rupert	95
robert	33	peter	25	rupert	22	raymond	12	peter	17	matt	10	robert	61	robert	70
yvonne	22	carol-anne	17	dave	17	robert	8	naomi	16	rupert	7	paul	31	bruza	39
rob	15	robert	15	justin	13	yukio	6	simon	15	lee	7	naomi	28	julia	39
naomi	12	lee	15	david	9	sweeney	5	ron	12	bruza	6	justin	23	lee	20
mcarthur	9	joe	9			warwick	4	kim	11	rob	6	yvonne	22	wasim	18
bruza	9	michael	9			mark	4	justin	10	mcarthur	5	rob	22	raymond	17
john	6	charley	9			dawei	3	nick	7	kim	5	dave	17	yukio	15
carol-anne	6	kevin	8			dave	3	rob	7	davies	3	simon	15	keith	11
kate	5	jacqui	7			andry	3			ron	2	nick	14	joe	11
…		…		…		…		…		…		…		…	
Avg	11		0.9		11		0.7		9		0.7		14		1.4
Stdd.	11		3.2		11		2.6		6		2		18		6

Table 6: “guidebeam” vector (people’s names only)

Rupert				Peter				Robert				Combined			
Additions		Largest change		Additions		Largest change		Additions		Largest change		Additions		Largest change	
wasim	26	peter	-80	bruza	17	paul	-30	matt	10	rupert	-18	julia	39	paul	-28
lee	15	yvonne	-20	raymond	12	rupert	-20	bruza	6	naomi	-15	wasim	18	naomi	-26
joe	9	robert	-18	robert	8	dave	-15	mcarthur	5	simon	-15	raymond	17	rupert	-19
michael	9	rob	-12	yukio	6	justin	-12	davies	2	ron	-10	yukio	15	justin	-18
charley	9	…		sweeney	5	…		david	2	…		keith	11	…	
kevin	8	kate	2	warwick	4	peter	1	nataliya	2	lee	2	joe	11	robert	10
jacqui	7	rupert	9	mark	4			ross	1	robert	2	jacqui	8	lee	15
teresa	7	carol-anne	11	dawei	3			philippe	1	peter	11	mark	8	bruza	30

Table 7: Change in “guidebeam” vector between pre and post LSA (people’s names only)

Naomi can see several details immediately: firstly, Robert, Peter and/or Rupert are the most likely people she needs to contact; secondly, she is aided to recall, by the connection between herself and Rupert and Robert, that she has had emails before about Guidebeam from these two people.

Figure 2 presents a similar view of the data to Figure 1[2], except the focus is no longer on the second tier of people, but on non-people elements. Naomi can see that Rupert’s work with Guidebeam has been with “chic” and “boeing” (organisations), and “management”, “installation”, “licence” and “meeting”.

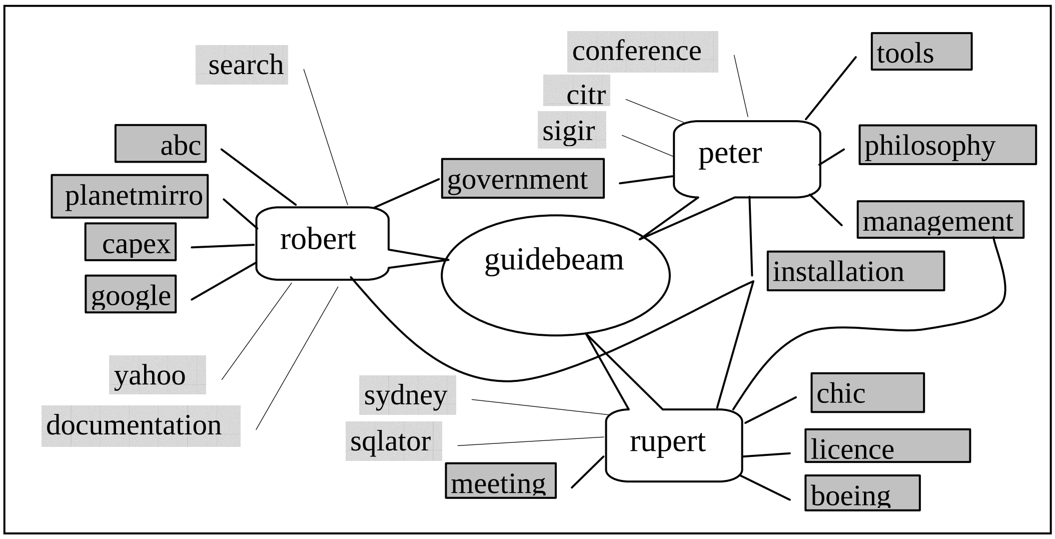

Figure 2: Context-sensitive network of topics (Context = "guidebeam")

In both figures, further important information is available. Thus far, what has been available to Naomi has been explicit knowledge: that which HALe has captured and made available. However, there is implicit or tacit knowledge also available. This is uncovered by techniques such as LSA. Table 7 displayed the changes in people’s names, and these are reflected in Figure 1 by light-grey un-enclosed boxes ▨.

The implicit connection between Robert and David, shown in Figure 1, links Peter and Robert by some association to David (again, without going into detail, there is a strong association between the three people). Similarly in Figure 2 Naomi can see that there is an implicit association from Rupert with Guidebeam and SQLator (they are both products that Rupert is associated with; being products many associated terms are in common – installation, licence, etc.).

Naomi is reasonably satisfied – she has identified the key players for Guidebeam within the organisation; been assisted to recall that Rupert, sitting next to her, is someone she should talk with; and been made aware of connections

[2] Space precludes the presentation of the base data from which this diagram is drawn

between people and topics associated with Guidebeam. However, two further examples can demonstrate serendipitous flows of information.

Peter has an association with Justin in connection with Guidebeam (Fig.1). The “Justin” connection in Peter’s email provides connections with Robert and Rupert. In turn, Justin in Rupert’s email and Robert’s email show strong associations with Nick. Again, the strong associations of Nick, in these two email sets, both agree on a connection with Kate. Thus a network of explicit and implicit connections can be formed, and Naomi can become informed that an association of some form exists between Justin, Nick and her colleague Kate. Note that no attempt to disambiguate the nature of the association is necessary, although it may be assisted by examining the vectors.

The second example shows how, if access is available to the original email messages, associations can be disambiguated and explained. From Fig.1., the LSA vectors offer the (possibly) implicit connections to “Warwick”, “Raymond” and “Yukio”. Naomi is curious about “Yukio”, so she sends an expanded query to a text retrieval system storing the emails, based on the LSA vector for Yukio. From the emails retrieved, Naomi learns that years before, Peter and Yukio were performing joint research on the “Hyperindex browser”, a forerunner of the Guidebeam product.

The last example used the LSA (or HALe) vector as the basis of an expanded query to a text retrieval system serving the email collection. The vector contains weighted associations to terms which the retrieval system can use to boost precision. Recent large-scale experiments using HAL vectors for query expansion in text retrieval has produced encouraging results with respect to precision (Bruza & Song, 2002). In practice, high precision can equate to lower cognitive load on the user as they are not confronted by large amounts of irrelevant material (in this case emails). Compare this to a “normal” search system which would process the 14000+ email messages and find that 1020 mention the word ‘guidebeam’ in the message body. Even a browsing system which measured, for example, how many messages mentioned “guidebeam” for each sender, would not uncover particularly useful associations (the rankings would be Peter >> Kevin > Rupert > Justin > Robert > Ron, Carol-Anne, Craig > … among a total of 34 unique senders evenly spread through the three personal email data sets).

Although all the three people whose email is being used have an organisational association with Guidebeam, it is not necessarily their major concern: only 1,020 of the 14,424 messages mentioned Guidebeam somewhere in the message body. The vectors with the largest dimensions in the combined messages of the three people, i.e. the words with the largest number of associations and thus used in the largest number of contexts, were peter (1275), robert (968), url (729), rupert (619), information (494), guidebeam (422), time (305), etc. We do not believe that the choice of only three email contributors, and their associations to Guidebeam, has hade any negative impact on the experiment.

The use of private email[3] is a difficult issue for privacy reasons. For this work, the use of the unmodified email was required as the purpose was to identify useful and interesting people. It would be feasible to produce the dimensional spaces from the original email messages, and allow people access to the spaces without providing access to the original text: an expertise management assistant could send an email to the 'owners' of particular emails that match an external search for specific expertise; it would then be the decision of the individuals concerned to answer the query. While fraught with dangers, we use the reasoning implied by Nardi *et al.* (2000), and believe that the various ends are worth the troubles.

Conclusions & Future Work

This paper is about finding useful and sometimes serendipitous connections between people by mining semantic associations from their email communications. To this end, the HALe model has been detailed which produces high dimensional representations of words within a context space defined by a collection, or sub-collection of emails. HALe can be used as the basis of a discovery mechanism to extract explicit associations between people given a seed contextual description (e.g., the name of a product, a particular person etc.). Latent Semantic Analysis (LSA) is applied to the high dimensional context space produced by HALe to "lift" implicit associations between people.

By means of a substantial email corpus, the discovery techniques have been shown to be feasible. We believe the techniques would scale to larger corpora, and probably be applicable to other utterance-based data sets such as mailing lists. The dimensional reduction aspect inherent to LSA has the potential to become a computational bottleneck for larger collections of email, but recent research from the knowledge discovery and data mining community has uncovered approximations of LSA, which seem to outperform LSA as well as being computationally tractable (e.g. Karypis & Han, 2000).

The scenarios presented in this paper are drawn from real-life experience in a small-to-medium information technology company. Anecdotal evidence suggests that useful, appropriate and at times serendipitous associations between people are being brought to light modulo the given context. Future experiments with larger datasets will feature more detailed user feedback. Our results show that the range of associations to people are very sensitive to the context space being used; for example, the network of people surrounding Guidebeam from Rupert's email is different to the network computed from Peter's email. Therefore we conclude that discovery of networks should not be restricted to a global email corpus, but should involve a mixture of information gleaned from the global corpus to

[3] or email considered private; some country's laws make the use of email conducted using an account supplied by the organisation the property of the organisation. There is still resistance to this.

investigate associations within particular sub-corpora. In this paper, the sub-corpora were static, but there is no reason not to employ dynamic corpora, like a set of emails retrieved from an intranet search engine based around a certain contextual description.

We feel that the techniques presented here form part of a solution to allow informal, ephemeral and mostly hidden networks of people to be discovered. Such “social networks” are integral to fostering collaboration in the enterprise, making use of all the resources possible. In short, we claim to have made a step to help enhance the awareness of individuals to their environment: other individuals, groups and information.

Acknowledgments

The work reported in this paper has been funded in part by the Co-operative Research Centre for Enterprise Distributed Systems Technology (DSTC) through the Australian Federal Government’s CRC Programme (Department of Industry, Science & Resources). We are also immensely grateful to Rupert for allowing us access to his email repository.

References

Ackerman, M. and McDonald, D. (1996): ‘Answer Garden 2: merging organizational memory with collaborative help’. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work* (CSCW), 1996

Bruza, P.D and Song, D. (2002): ‘Inferring query models by computing information flow’. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM 2002)* ACM Press, pp.260-269.

Burgess, C., Livesay, K. and Lund, K. (1998): “Explorations in context space: words, sentences, discourse”. *Discourse Processes*, v25, pp.211-257

Ducheneaut, N.B. (2002): ‘The social impacts of electronic mail in organizations: a case study of electronic power games using communication genres’. *Information, Communication and Society,* v5, n2, pp.153-188

Ducheneaut, N.B. and Bellotti, V. (2002): ‘Ceci n’est pas un objet? Talking about things in email’, forthcoming in a special issue of the *Journal of Human-Computer Interaction*

Freyd, J. (1983): “Shareability: the social psychology of epistemology”, *Cognitive Science*, v7, pp.191-210

Gabbay, D. and Woods, J. (2000): ‘Abduction’, Lecture notes from *ESSLLI 2000* (*European Summer School on Logic, Language and Information*). Online: http://www.cs.bham.ac.uk/~esslli/notes/gabbay.html

Gärdenfors, P. (2000): *Conceptual Spaces: the Geometry of Thought.* MIT Press, London, 2000

Jones, Q., Ravid, G. and Refaeli, S. (2001): ‘Information overload and virtual public discourse boundaries’. In *INTERACT, Eighth IFIP TC.13 Conference on Human-Computer Interaction,* Japan IOS Press

Karypis, G. and Han, E-H. (2000): ‘Concept indexing: a fast dimensionality reduction algorithm with applications to document retrieval & categorization’. University of Minnesota, Department of Computer Science, Technical Report #00-016

Kautz, H., Selman, B. and Shah, M. (1997): ‘ReferralWeb: combining social networks with collaborative filtering’. In *Communications of the ACM*, v40 n3, March 1997

Kimble, C., Hildreth, P. and Grimshaw, D. (1998): ‘The role of contextual clues in the creation of information overload’. In *Proceedings of the 3rd UKAIS Conference*. April 1998, Lincoln University, McGraw Hill, pp.405-412

Landauer, T.K., Foltz, P.W., and Latham, D. (1998): ‘Introduction to Latent Semantic Analysis’. *Discourse Processes*, v25, pp.259-284

Lewis, D. and Knowles, K. (1997): ‘Threading electronic mail: a preliminary study’. *Information, Processing and Management,* v33 n2, pp.209-217

LTCHUNK (software): online (6 May 2003) http://www.ltg.ed.ac.uk/software/index.html

Lund, K. and Burgess, C. (1996): “Producing high-dimensional semantic spaces from lexical co-occurrence”. *Behavior Research Methods, Instruments & Computers*, v28(2), pp.203-208

Mikheev, A. (2000): ‘Document centered approach to text normalization’. In *Proceedings of SIGIR'2000*, pp. 136--143.

McArthur, R. and Bruza, P.D. (2003a): ‘Dimensional representations of knowledge in online community’, in Ohsawa, Y. (ed.) (2003, in press) *Chance Discovery,* Springer-Verlag

McArthur, R. and Bruza, P.D. (2003b): ‘Discovery of tacit knowledge and topical ebbs and flows within utterances of online community’, in Ohsawa, Y. (ed.) (2003, in press) *Chance Discovery,* Springer-Verlag

Nardi, B. and Engström, Y. (1999): ‘A web on the wind: the structure of invisible work’. In Nardi, B. and Engström, Y. (eds) *Computer-Supported Cooperative Work*, v8 n1-2

Nardi, B., Whittaker, S., and Schwarz, H. (2000): ‘It’s not what you know, it’s who you know: work in the information age’. *First Monday*, v5, n5, May 2000. Online: http://firstmonday.org/issues/issue5_5/nardi/index.html

Nonaka, I. and Takeuchi, H. (1995): *The Knowledge-Creating Company*, OUP, New York

Ogata, H. and Yano, Y. (1998): ‘Collecting oganisational memory based on social networks in collaborative learning’. In *WebNet*, pp.822-827

Perry, J. (1997): ‘Indexicals and demonstratives,’ in *A companion to the philosophy of language*, Hales, B. and Wright, C. Eds. Oxford: Blackwell, 1997, pp.593-595.

Perry, J. (1998) ‘Indexicals, contexts, and unarticulated constituents’, in *Proceedings of the 1995 CSLI-Armsterdam Logic, Language and Computation Conference*. Stanford: CSLI Publications, 1998

Schwartz, M. and Wood, D. (1993): ‘Discovering shared interests among people using graph analysis of global electronic mail traffic’. In *Communication of the ACM*, v36, n8, 1993