Inferring Ongoing Activities of Workstation Users by Clustering Email

Yifen Huang, Dinesh Govindaraju, Tom Mitchell, Vitor Rocha de Carvalho, William Cohen

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

1 The Problem

We are interested in automatically discovering the key ongoing activities of a workstation user, such as committees to which she belongs, writing projects in which she is involved, etc., based on the contents of her workstation. The thesis underlying our research is that this collection of user activities can be automatically inferred from the variety of data available on most users' workstations, including their emails, files, online calendar, and history of web page accesses. Knowledge about the user's activities could be used in a variety of ways, such as cross-indexing email, calendar events, files, and web accesses according to activity, or automatically producing a 'briefing folder' for each meeting on the user's calendar. We describe here our initial research on inferring such activities by examining only the user's email. In particular, we describe a variety of unsupervised clustering methods designed for clustering emails by user activity, and the use of information extractors and pretrained classifiers to infer additional information about each discovered cluster. Experimental results are presented for emails from three users.

2 Approach: Clustering, Classification, Information Extraction

Designing an algorithm to cluster user emails involves several design choices. First, we must choose how to represent emails. We represent each email using both header features (denoted by H) and body features (denoted B). Header features include the subject line, email addresses, domains of these email addresses, and words judged to be proper nouns. Body features include the bag of words found in the email body. A second design choice is the clustering algorithm. Here we explore two classes of algorithms (for additional details, see [1]):

1. *EM-based mixture of multinomials algorithms.* (EM) Discovering clusters is formulated as the problem of identifying the components of a mixture distribution assumed to generate the data. An EM algorithm is applied to iteratively compute a locally maximum likelihood assignment of emails to clusters, given a target number of clusters, *k*. We consider two approaches to generating the initial assignment of emails to clusters. Random Initialization (RI) randomly assigns each email a probability distribution over *k* clusters. Distant Initialization (DI) generates *k* distinct initial groups of emails, where each group consists of five similar emails. These *k* groups are chosen using a heuristic sampling method designed to find groups with maximum inter-group distance. One variation on this algorithm involves using email thread information to blend the posterior class probabilities of individual documents within the same thread (method BV, see [1]).

2. *Bottom-up agglomerative clustering* .(BU) In contrast to EM, BU finds clusters bottom-up, by iteratively merging email documents into trees that represent clusters. The process begins by considering only the subject lines of emails, grouping those emails containing identical subject lines (which are typically from the same email thread). Pairs of subtrees with the largest cosine similarity are then merged, iterating this process until all emails are merged into a single tree that forms a hierarchical clustering. The *k* desired clusters are determined by partitioning the tree at points where the leaves have the smallest cosine similarity.

After clustering the email, the program post-processes each cluster by collecting various statistics (e.g., who sent the most email within this cluster?), applying information extractors to the emails (e.g., to obtain the names and dates mentioned in email bodies), and applying a previously-trained email classifier to determine which emails within the cluster correspond to requests (e.g., meeting or information requests). The purpose of this step is to automatically

construct a structured description of the user activity associated with the cluster, for use by an intelligent workstation assistant, as illustrated below.

3 Experiments and Results

We evaluated the results of our clustering algorithms over three email corpora from three authors. Two corpora, DG (486 emails from 15 folders), and YH (623 emails from 11 folders) were sorted beforehand by their users. The third corpus, TM (1148 emails from 1 folder) was an unsorted collection of all of the user's emails. We first evaluated the clustering algorithms by their ability to reconstruct the folders for corpora DG and YH, given the union of emails from all folders within the corpus. Table 1 summarizes the accuracies reconstructing these folder assignments, for several algorithms, feature sets, and initialization methods. Variances reported in the table indicate differences in accuracies over 10 trials of the same algorithm arising from different random starting points. As shown in the table, each of these algorithms results in accuracy substantial greater than randomly assigning emails to folders, and the best performing algorithm is EM-DI on DG corpus and EM+BV-DI on YH corpus.

Table 1: Results on DG and YH corpora

Algorithm	Feature set	Initialization	Accuracy on DG	Accuracy on YH
BU	B	-	0.55	0.41
BU	HB	-	0.72	0.49
EM	HB	RI	0.60 ± 0.18	0.41 ± 0.14
EM	HB	DI	**0.79 ± 0.08**	0.48 ± 0.13
EM+BV	HB	DI	0.75 ± 0.14	**0.50 ± 0.07**

We used the TM corpus to explore the feasibility of automatically constructing structured descriptions of the dominant activity associated with each discovered cluster. In a qualitative evaluation, user TM found each cluster was related to one or more of his ongoing activities (e.g., a committee, family email, etc.), but that each cluster also contained 20-50% of extraneous emails not affiliated with the main activity of the cluster. Despite this non-homogeneity in the clusters, the post-processed descriptions of each cluster often produced quite reasonable descriptions of its dominant activity. Figure 1 shows the automatically constructed description for one cluster of TM emails related to a research project named CALO, involving research on intelligent workstation assistants. The keywords, emails, and extracted names were found by the user to be highly related to this CALO activity.

ActivityCluster5 (105 emails)

- ***Keywords***: CALO, TFC, SRI, examples, heads, labeled, Leslie, HMM, contacts, email, task
- ***Primary Senders:*** mitchell@cs.cmu.edu(39), lpk@ai.mit.edu(7), mccallum@cs.umass.edu(6)
- ***User Activity Fraction***: 105/1448=.072 of total email
- ***Intensity Of User Involvement:*** user authored 37% of email traffic; (default 31%)
- ***Extracted Names:*** Leslie(23), Rebecca(21), Carlos(12), Ray(10), Stuart(9), William(9), April(9), …
- ***Extracted Dates:*** Wed(39), Tues(33), Fri(25), Mon(23), Thurs(20),… Feb 18 (16)
- ***Extracted Times:*** 5pm(24), noon(14), morning(8), 8am(7), before 5pm(7),...
- ***RequestEmails:*** <emailA>, <emailB>, …

Figure 1: Automatically constructed activity description

In summary, our clustering algorithms produced email clusters that reproduced email folder assignments for DG and YH at accuracies of 40-80%, and produced clusters that TM found qualitatively aligned with recognizable activities. Despite the non-homogeneity of these clusters, they often produced reasonable structured representations of the dominant user activity associated with the cluster, such as the one shown in Figure 1. In future work we intend to explore the use of social networks of senders and receivers to refine clusters, and to explore algorithms that jointly cluster calendar entries, files, and web accesses related to emails.

This work was supported by Darpa under the CALO and RADAR project contracts.

1. Yifen Huang, Dinesh Govindaraju, Tom Mitchell. *Learning Ongoing Activities of Workstation Users by Clustering Email.* Internal CMU CALD Working paper, http://www.cs.cmu.edu/~hyifen/publication/EmailCluster04.pdf, 2004.