International Journal of Computational Linguistics and Applications, Vol. 6, No. 2, 2015, pp. 29–42
Received 28/01/2015, Accepted 25/02/2015, Final 23/03/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

Footprints on Silicon: Explorations in Gathering Autobiographical Content

ESHWAR CHANDRASEKHARAN
SUTANU CHAKRABORTI
Indian Institute of Technology Madras, Chennai

ABSTRACT

As we interact with the world, we leave behind digital trails in the form of emails, blogs, tweets and posts, which serve as a rich source of data for generating our individual life stories, or autobiographies. Central to addressing the problem is the ability to discriminate content that is of autobiographical value from the rest. The features required for this classification task need to be discovered from the unstructured data, metadata, sentiments, properties of the social network and temporal properties of the interactions. In this paper we identify several dimensions of this problem, present some preliminary results on our explorations, and identify interesting research problems for the future.

1. INTRODUCTION

As more and more of human interactions occur online, a large amount of digital trails are left behind. One of the challenges that emerged in the UK Computing Research Committee's workshop on Grand Challenges for computing science in 2002, was entitled "*Memories for Life*." The idea was to analyze and use the digital data that people have about themselves which will soon be huge in size. An exemplar of this project was "*Stories from a Life*," which attempts to represent the stored memories in the form of stories. These stories could be generated under the supervision of the individual, or in an automated manner[1].

A systematic exploration of human interactions across the vastness of the Internet can serve as a rich source of data for generating individual life stories or autobiographies of people. Different sources like emails, blogs, tweets and posts on social networking sites can serve as potential sources of information about events of interest in a person's life. In this paper, we look at emails as the primary source of information and explore different methods to discriminate content that is of autobiographical value from the rest.

Emails are exchanged on a daily basis between several people, and contain varied types of content ranging from personal and professional messages to advertisements and spam. Over the course of a few years, there would be so much content buried in the inbox that correspond to important events or landmarks in the person's life. They could be about vacations, job promotions, cultivation of a new hobby, turning points in a person's life like the birth of new ones in the family and so on.

We are interested in building systems that analyze emails that get accumulated in a person's inbox over time, and identify those that could be of autobiographical value. We want to model the problem as a classification task, and try to use the unique structural properties of emails to gather autobiographical content from the collection of emails present in a person's inbox, in an automated manner.

Note that our system may not generate the summaries all by itself, but the gathered data will aid the user in creating one by prompting important details. The focus of our work is not the story generation part, but the crucial data gathering tasks that come before it.

2. DIMENSIONS OF THE PROBLEM

There are two methods that could be used by a person to generate an autobiography. The first is a top-down approach using cues like the person's hobby, biographical data, family members, and landmarks in personal and professional life to generate content. Research suggests that people use *interesting* events as "*anchors*" when trying to reconstruct memories of the past.

There has been work on probing the value of timelines and temporal landmarks for guiding search over subsets of personal content[2]. By using the landmark events that are identified by the person, a life-story can be created in top-down manner.

The second method is a bottom-up approach where we look at existing autobiographies and identify what discriminates events of autobiographical importance from the rest. We model the identification task as a classification problem, and are interested in looking at properties that could help in classifying autobiographical content present in a personal store of data, like a person's email inbox. We want to use a subset of emails that are labelled by the user and identify the features required for building our classifier. A summary of the gathered data can be used to generate the person's life story or autobiography with minimal supervision.

There are different dimensions which can be discovered from the unique structural properties of emails, in addition to textual content. We will look at some of these dimensions in detail, and look at how well they perform during classification.

2.1. *Text*

We would like to look at certain properties of the textual data present in emails exchanged by the user to aid our classifications.

Lexical features

We are interested in extracting textual keywords or subsets of words occurring in emails, that can describe the meaning of an email on the basis of properties such as frequency and length. We would like to look at the keywords present in an email and tell if it contains autobiographical content or not.

Language Function

Another interesting approach would be performing text classification using Language Function. The aim of Language Function Analysis (LFA) is to determine whether a text is predominantly expressive, appellative or informative. LFA is used to classify the predominant function of a text as intended by its author, and understand why a text was written[3]. It has been

observed that language functions relate to the writing style of a text, and they can be derived from statistical text characteristics obtained by using machine learning of lexical and shallow linguistic features like text type, writing style, sentiments and simple genre features.

Sentiment

The idea is that messages which contain high sentiments are likely to be expressing the user's opinion, stand or feeling about a particular topic in a conversation with another person through email, and are likely to be considered as containing information of autobiographical value by the user. Irrespective of the type of sentiment, negative or positive, an email displaying the user's sentiment regarding a topic is rich in content and could be classified as autobiographical.

2.2. *Mail network*

Our idea is that contacts with whom the user interacts on a one-to-one basis very frequently are generally close to the user, in a personal or professional nature. We could say that a contact is important if out of all the emails exchanged by the contact, most are one-to-one interactions with the user. This way we can compute the importance of a contact in a person's email network by looking at the sender-receiver characteristics of all the emails in the inbox.

2.3. *Email metadata*

Labels are assigned to each incoming email by the email client automatically, or manually by the user depending on different factors like the person sending the email, the words present in the subject and body of the email, and so on. Therefore, looking at the names of Labels or Filters that are assigned to an email can serve as a good indicator of the type of content present in an email, and could be used to distinguish between emails that contain autobiographical content and the rest.

2.4. *Time*

We would like to look at certain temporal properties of the user's interactions with other contacts to aid our classifications.

Threads

There are a lot of conversations or threads of emails present in a person's inbox where the user has exchanged emails with another contact or a group of contacts in succession. We would like to see if a large number of emails being exchanged between people on the same topic in a thread has an impact on the emails being of an autobiographical nature.

Burstiness

We could also look at the burstiness of emails being sent or received by a person, as an indication of the importance of emails. The periodicity, quantity and context shifts in the genre or type of content present in the emails being exchanged by the user can be used to identify mails containing autobiographical content. An example of a shift in context would be an email about travel tickets or say the birth of a baby, when all the previous mails were professional mails that were related to the person's work.

3. OUR APPROACH

We present a basic bottom-up scheme for performing the classification of emails containing autobiographical content by looking at different dimensions of emails, as our first work in a longer line of research. We build the classifier by mining discriminating features like textual keywords, threads, labels and mail network properties.

3.1. *Textual keywords*

The lexical features present in an email are obtained by tokenizing the textual data present in the "*Subject*" and "*Message Body*" components of the email. We perform keyword extraction on the textual content by removing stopwords, and using frequency of occurrence to rank the keywords. Then we perform

feature selection on the obtained keywords, by assigning scores to the different features using Information Gain values as the filter. Based on the scores, we rank the features and only keep the ones in the top which are the most distinguishing textual keywords in present in the email.

In Table 1, we list the top 50 words from the list of keywords that were obtained from the training set of user emails that gave the best performance when textual keywords were the only features used for building the classifier. These keywords can be used to quantitatively measure the lexical similarity of a new email with an email which is known to be tagged as containing autobiographical content or not.

Table 1. *Top 25 keywords from each class of emails obtained after feature selection using Information Gain filtering*

Autobiographical				
trip	undergraduate	going	tickets	booked
people	required	join	technology	feeling
confirmation	trekking	final	location	arrangements
places	finish	department	goal	inform
challenge	guidance	definitely	airport	dinner
Non-Autobiographical				
support	groups	stop	literature	subscribed
message	click	view	watch	reading
class	student	offers	matter	included
late	texts	encourage	bank	game
shop	anybody	read	story	software

3.2. *Mail network properties*

We compute the total number of emails exchanged by the user with the different contacts present in the user's contact list and also compute the number of these emails which are of a one-to-one nature with the user. A one-to-one email with the user is when there is only one recipient and one sender (which is always the case) and one of them is the user. Since we want to look at one-to-one interactions where the user is a primary part of the conversation, we make sure that the user's email ID is contained either in the "*to:*" or "*from:*" address in the email. We would not be interested in emails that are conversations between other

contacts and ignore emails where the user's email ID is contained in the "*cc:*" or "*Bcc:*" section of the email. We can use this information along with thread count of an email to capture the email network properties that are key to making a contact's email important to the user.

The features used to capture the email network properties are the number of overall mails sent by the e-mail's sender, number of recipients in the email and whether the user is a part of the email. Numerically,

Number of recipients = Number of email IDs present in "to:" address

We use the number of recipients as a numeric attribute to represent the number of contacts to which an email has been sent to. To see whether the user is a part of the email, we look at the email IDs of the sender and recipients to see if atleast one of them contains the user's ID.

We observed interesting properties in the mail network when we looked at the sender-receiver property along with the thread counts of emails. There were cases where we observed strong cycles being formed in the emails exchanged between the user and certain contacts. These were people to whom the user sent a one-to-one email and they responded back, and this cycle kept repeating for a number of times. We observed that these contacts were personal to the user and were of autobiographical value. There were other cases where emails were sent to groups of contacts, out of which just one or two would respond in a one-to-one manner with the user. We observed that these contacts were of a professional nature, where the interactions did not have as many cycles in their one-to-one interactions with the user, as observed in the previous case.

3.3. *Labels*

We collected the different label names that were present in our collection of emails and selected the most distinguishing labels. We observed that the labels that were most useful in the classification of emails containing autobiographical content were: "*Starred*", "*Important*", "*Sent*" and "*Inbox*".

These 4 label names were used as the metadata features for our classifications and we used them as f0,1g - features whose values are assigned based on whether the identified label names are assigned to an email or not.

3.4. *Threads*

We identify threads present in the inbox by looking at the "*Subject*" of emails. Emails that are part of the same thread contain the exact same "*Subject*" or with prefixes like "*Re:*" and "*Fwd:*" Numerically,

thread count = number of emails containing the same "Subject"

We compute thread counts for all the threads present in the Inbox and map the obtained values to the corresponding "*Subject*". We use this mapping to get the thread count feature for an email by using its "*Subject*" to perform a simple look-up. This gives us the size of the thread that the email is a part of, which is used as a numeric feature for our classification.

4. PRELIMINARY EXPERIMENTS

The corpus used for building and testing the model was obtained from the private emails of 3 different test users. Each user was asked to go through a representative set of emails present in his/her Inbox and tag them as containing autobiographical content or not. This kind of annotation helps us perform the task of gathering biographical content present in emails in a user-specific manner. It should also be noted that measures like inter-annotator agreement could not be computed as the emails used were private to a particular user.

From each user-tagged inbox, a random sample of 150 emails was used for the training phase and the classifications were done on another random sample of 80 test emails obtained from the user-tagged email inbox. We built 3 different classifiers, and report the average precision and recall values obtained for the 3 sets of user-tagged emails, when using di_erent combinations of features to build the classifier. We observed that

the average number of words present in the Email Corpus used for building our model to be 46746 words. Out of these, 3927 were distinct words, that were present in an English dictionary and were not stop words.

4.1. *Methodology*

The tagged email inbox of the test user is obtained in the form of a single *Mail Box* format file from the email client (Gmail, in our case). *Mail Box* (*MBox*) is a generic term for a family of related file formats used for holding collections of electronic mail messages. All of the messages present in a mail box are stored in a single. *MBox* text _le in a concatenated manner. We parse this file and separate the individual emails into different files of a similar format, to extract the features present in each email.

The typical information which we are interested in, that is contained in a *MBox* format file are the following:

- From
- Sent to
- Subject
- Metadata (labels, folders, date, time and other information)
- Body of the message

We conducted our experiments by using the following sets of features to obtain the vector representations of each email:

- Textual Keywords
- Email network properties
- Thread count
- Label or Folder name

4.2. *Results*

In order to reduce the bias in the choice of test and train data for building the classifier, we split the email corpora into 10 sets and run 10-fold tests using different train-test splits in the ratio 9:1. We compute average values of the obtained results for 3 different classifiers.

We used Naive Bayes, Random Forest and LibSVM classifiers in WEKA with default parameter values, to conduct our experiments on the tagged inbox of 3 test users using WEKA[4]. We look at the results that were obtained and compare the performance of different types of features and classifiers.

In Table 2, we have the results obtained for the different classifiers when using each of the individual type of features on their own. Here, P stands for Precision and R stands for Recall values that were obtained.

In Table 3, we have the results obtained when different combinations of features were used to build the classifiers. Here, *L* stands for Labels, *Txt* stands for Textual keywords, *Th* stands for Threads, *MN* stands for Mail Network, *All* stands for all 4 types of features, *P* stands for Precision and *R* stands for Recall values that were obtained.

Table 2. *Standalone performance of different types of features*

Classifier	Text		Labels		Threads		Mail Network	
	P	R	P	R	P	R	P	R
Naive Bayes	0.88	0.858	0.783	0.763	0.651	0.654	0.487	0.433
Random Forest	0.914	0.904	0.845	0.821	0.752	0.725	0.608	0.609
LibSVM	0.932	0.925	0.834	0.813	0.615	0.613	0.596	0.592

Table 3. *Performance of different combinations of features*

Classifier	L+Txt		L+Txt+MN		L+Txt+Th		All	
	P	R	P	R	P	R	P	R
Naive Bayes	0.868	0.846	0.874	0.854	0.879	0.854	0.888	0.867
Random Forest	0.948	0.946	0.942	0.942	0.952	0.95	0.951	0.95
LibSVM	0.952	0.95	0.959	0.958	0.908	0.904	0.921	0.917

We observed that the textual keywords and labels were most effective during classification when considered by themselves. Other features like email network properties and thread counts were not very good indicators on their own, but when augmented with textual keywords and labels, they were observed to give improved performances as seen in Figure 1.

We compute the average number of correctly classified instances observed for the 3 test user emails when using different

combinations of features and compare them to see how they differ, for the Naive Bayes classifier. In particular, we observe that text keywords and labels are strong indicators on their own. As a result, we consider the other 2 features: email network properties and thread counts on their own and see how the addition of labels and text keywords to these 2 sets of features impact the performance of the classifier, in a graphical manner in Figure 1. Here, *All* stands for all 4 features, *L* stands for Labels, *Txt* stands for Textual Keywords, *Th* stands for Threads, *MN* stands for Mail Network, *L+Txt+Mn* stands for combination of Labels, Textual Keywords and Mail Network, and so on.

A slight drop in performance was observed with the addition of thread count to the set of features in LibSVM, which was due to an increase in the number of misclassifications. This was due to the presence of certain test emails that had relatively high thread counts despite not having autobiographical content in them. Overall, the best performance was observed when all 4 types of features were used to build the classifier and the most discriminating individual features were observed to be text keywords and label names.

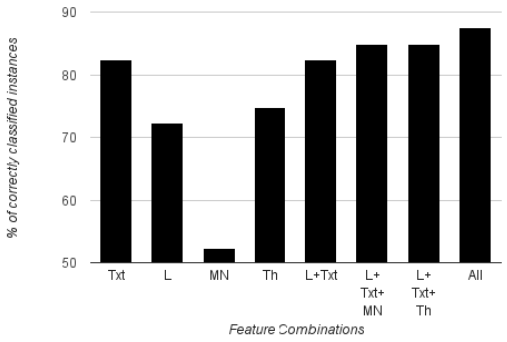

Figure 1. *Percentage of correctly classified instances for different combinations of features*

5. DISCUSSIONS

Most of the misclassifications were observed to be due to lack of information about the importance of certain class of emails or

contacts, that were not represented well enough in the training set of emails. For instance, there were emails from family members or close friends about events like functions, gatherings and meetings which were over-looked as they were small in number and were not captured well due to lack of emails corresponding to similar events in the training data. Also, the importance of activities or events that are related to a person's personal life in terms of his hobbies, interests and so on, were overlooked.

These misclassifications can be resolved by including a top-down perspective to our work. We can get the person to specify important details like hobbies, biographical data, family members, landmarks in personal and professional life, which they feel are of autobiographical value. These could serve as useful cues in identifying emails with autobiographical content that might get misclassified when using the bottom-up approach alone. We could look at a mix of top-down and bottom-up methods in the future to go about gathering emails with autobiographical content and, evaluate how the addition of top-down techniques help in resolving the previous misclassifications and, improve the performance of classifiers.

In our work, we used emails present in the inbox of 3 test users as the primary source of data for our experiments due to a lack of relevant datasets that are publicly available. There are several privacy and content related issues that restrict the availability of emails, and a potential future work would be to come up with a suitable and representative email corpora which can be made publicly available for future use.

Apart from emails, our idea can be extended to other rich sources of personal data like blogs, posts and interactions on social media, all of which have the potential to contain autobiographical content about a person. In the case of emails, we looked at features like labels and filter names, thread counts, one-to-one interactions with contacts in the mail network and so on, in addition to the textual content present in emails. It would also be an interesting problem to identify different social network and temporal properties of interactions in different sources and explore how they can leveraged to gather autobiographical

content for our main goal: generating the life story of a person with minimal supervision.

6. CONCLUSION

In this paper, we have looked at the problem of using online social interactions to create a person's life story. Central to addressing this problem is the ability to discriminate content that is of autobiographical value from the rest. We have identified emails as a rich source of information for the creation of a user's autobiography, conducted preliminary tests and presented the results of our explorations on emails. We observed that textual content and label names present in emails were the most informative individual features and that the combination of textual content, labels, mail network properties and threads gave the best performance, from our classification experiments on a bottom-up approach to autobiography generation. We have also talked about introducing a mix of top-down approaches in the future, to resolve the misclassifications that were observed, and interesting extensions to other sources of personal data like blogs and interactions on social media through posts, etc., to gather autobiographical content for the generation of a person's life story.

REFERENCES

1. Fitzgibbon, A. & Reiter, E. 2004. Memories for life: Managing information over a human lifetime. In T. Hoare & R. Milner (Eds.), *Grand Challenges in Computing Research* (pp. 13-16). Swindon: British Computing Society
2. Ringel, M., Cutrell, E., Dumais, S. & Horvitz, E. 2003. Milestones in time: The value of landmarks in retrieving information from personal stores. To appear in the *proceedings of Interact 2003*.
3. Wachsmuth, H. & Bujna, K. 2011. Back to the roots of genres: Text classification by language function. In proceedings of the *5th IJCNLP* (pp. 632-640).
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11/1.

5. Ulrich, J., Murray, G. & Carenini, G. 2008. A publicly available annotated corpus for supervised email summarization. In proceedings of the *AAAI EMAIL Workshop* (pp. 77-87).
6. Yang, Y. & Pedersen, J. 1997. A comparative study on feature selection in text categorization. In *ICML 1997* (pp. 412-420).
7. Kiritchenko, S., Matwin, S. & Abu-Hakima, S. 2004. Email classification with temporal features. *Intelligent Information Systems* (pp. 523-533).
8. Cohen, W. 1996. Learning rules that classify e-mail. In proceedings of the *AAAI Spring Symposium on Machine Learning in Information Access*.
9. Manning, C., Raghavan, P. & Schtze, H. 2008. Vector space classification. *Introduction to Information Retrieval*. Cambridge University Press
10. Garera, N. & Yarowsky, D. 2009. Modeling latent biographic attributes in conversational genres. In proceedings of the *Joint Conference of Association of Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, (pp. 710-718).

ESHWAR CHANDRASEKHARAN
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
INDIAN INSTITUTE OF TECHNOLOGY MADRAS,
CHENNAI 600036, INDIA.
E-MAIL: <ESHWAR@CSE.IITM.AC.IN>

SUTANU CHAKRABORTI
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
INDIAN INSTITUTE OF TECHNOLOGY MADRAS,
CHENNAI 600036, INDIA.
E-MAIL: <SUTANUCG@CSE.IITM.AC.IN>