A Topic-based Forensic Analysis and Visualization of an Email network: Application to the Enron Dataset

Casey Kalinowski
University of Lynchburg, VA, USA
Kalinowski_c@lynchburg.edu
M. Zakaria KURDI
University of Lynchburg, VA, USA
kurdi_m@lynchburg.edu

Abstract

This work is about visualizing an email network with graphs. This visualization is based on the email's topics. So, the first part of this work is about exploring three rule-based methods and an unsupervised method of topic detection applied to a large dataset. Keyword or Term Frequency (TF) method is used as a baseline for comparison. Latent Dirichlet Allocation (LDA) combined with WordNet as well as two versions of conceptual topic detection, both involving a version of keyword extraction combined with WordNet, are also compared. Our results show that LDA combined with wordnet has the highest precision but a comparable F-measure to the conceptual approaches. Through a series of examples, we then demonstrate how annotating the emails with topics is a good way to shed light on the underlying professional and social relationships within the email network, which can provide substantial help within application contexts such as forensic investigations. This annotation is also showed to help in providing quantitative feedback about the performance of the topic detection algorithms.

Keywords

topic modeling, Social Network Analysis (SNA), Social Network Visualization, Term Frequency (TF), Latent Dirichlet Allocation (LDA).

التحليل الشرعي المرتكز على الموضوع والتصور لشبكة البريد الإلكتروني: التطبيق على مجموعة بيانات Enrom

هذا العمل يدور حول تصور شبكة البريد الإلكتروني مع الرسوم البيانية. يعتمد هذا التصور على موضوعات البريد الإلكتروني. الجزء الأول من هذا العمل يدور حول استكشاف ثلاث طرق تستند إلى قواعد وغير خاضعة للرقابة للكشف عن الموضوع يتم تطبيقها على مجموعة بيانات كبيرة. يتم استخدام طريقة تردد الكلمة أو المصطلح (TF) كخط أساسي للمقارنة. يتم النظر أيضًا في مقارنة تخصيص Latent Dirichlet Allocation (LDA) مع WordNet بالإضافة إلى إصدارين من اكتشاف الموضوع المفاهيمي، وكلاهما يتضمن إصدارًا من استخراج الكلمة الأساسية المدمجة مع WordNet وتعتبر أيضا بالمقارنة. تظهر نتائجنا أن LDA مدمجة مع wordnet لديها أعلى دقة ولكن مقياس F قابل للمقارنة للنُهج المفاهيمية. من خلال سلسلة من الأمثلة، نوضح بعد ذلك كيف يكون التعليق على رسائل البريد الإلكتروني مع الموضوعات وسيلة جيدة لتسليط الضوء على العلاقات المهنية والاجتماعية الأساسية الكامنة داخل شبكة البريد الإلكتروني، والتي يمكن أن توفر مساعدة جوهرية في سياقات التطبيق مثل التحقيقات الجنائية. يتم عرض هذا التعليق التوضيحي أيضًا للمساعدة في توفير ملاحظات كمية حول مدى أداء خوارزميات اكتشاف الموضوع.

1. Introduction

Different approaches were proposed in the literature to identify communities or groups of users within social networks. These approaches rely mostly on the connectivity of the users, which are represented as edges within a graph. The purpose of this research is to develop a program that will analyze the Enron email dataset to find patterns of networking between the employees of the Enron Corporation involved in the fraudulent acts of the company leading up to the 2001 bankruptcy. More specifically, it is designed to help identify and rank the actors involved in a given process such as accounting, stocks, IT, and management. Although it is possible to find some information from the organization's documents or employees about who is involved in a given process, it is hard to trust this information as some people hesitate to provide the real information during a criminal investigation. The second goal of this work is to group the actors involved in a process and understand the nature of their relationship that can be strictly professional or sometimes personal.

The approach proposed here consists of annotating the emails of a large dataset set using different approaches. The main technical challenge with the task of topic detection consists of the large size of the data that cannot be annotated by hand to train a supervised machine learning algorithm.

Several previous works in the literature viewed the relationships between the members of a social network from a quantitative angle [1], [2], [3]. For example, if two persons exchange 100 emails, 100 would be the weight of the connection between these two persons. However, this weight does not give any idea about the nature of the relation between the involved persons. In this paper, these relations are viewed from a qualitative angle, where the topic or subject of the connection is considered. Combined with the frequency that can help build a weighted graph that shed light on the true relationships within an organization. For example, if two persons exchanged 100 emails out of which 56 are about accounting and 44 are about IT that gives a much better idea about the nature of these persons. In many contexts, such as criminal investigation, learning about the actual hierarchy of actors involved in a process such as accounting or management can be hard to find.

The work presented in this paper contributes in two main areas. First, it proposes an approach based on the integration of keywords with two methods based on conceptual information from WordNet as well as an LDA combined with wordNet. After a comparison of these methods with a simple Term Frequency (TF) approach, it demonstrates how topic detection of emails can greatly

reduce the time of criminal investigations, where a large dataset must be searched. This is done by effectively visualizing the data using a topic graph that gives a precise idea about the interactions between the involved persons within the network.

This paper is showing how topic graphs can be used on real data collected from a real criminal investigation. It is also significant in that the technique can be applied to other social networks besides email, thus making it very useful in many application fields.

2. Topic Detection and Modelling

Topic modeling is based on three fundamental assumptions. First, every document has its internal (latent) topical structure. Second, this structure can be inferred from the document algorithmically based on the vocabulary used in each document. More formally, a collection of documents $D=\{d_1, d_2, \ldots, d_n\}$ may cover a set of topics $T=\{t_1, t_2, \ldots, t_m\}$. In a normally constituted data, we usually have $m<n$. Third, since words are the main indicators of topics, it is easy to imagine a mapping between topics and words such that $t_i=\{w_1, w_2, \ldots, w_k\}$. Topic detection in document *d* consists of algorithmically assigning a set of one or more topics (T_d) to every document using a function such that $F(D)\rightarrow T_d$ where $T_d \subseteq T$.

Information Retrieval (IR) and topic identification are tightly related. Given the size of the documents collection to deal with in IR, unsupervised techniques such as document clustering were used since the early stages of this field's investigation [4]. In such approaches, documents dealing with similar topics are supposed to fall within the same cluster. For example, [5] used Self Organizing Maps (SOM), while [6] used a combination of Latent Semantic Analysis (LSA) and K-means. A major disadvantage with clustering is that the machine-built clusters are not easy to interpret by human users, which limits the application of these techniques. Many works have used LDA for identifying the topics at the level of sentences in written texts [7]. Later many other applications to topic identification started to emerge like opinion mining [8], text summarization [9], analysis of open-ended survey questions [10] and machine translation [11].

3. Social Network Analysis (SNA) and Link Mining (LM)

Social networks are becoming a central part of modern society. Therefore, it is essential for a wide number of applications to extract information from these networks effectively. Link Mining (LM)

is concerned about exploring social networks from the angle of the relationship patterns between the entities. Hence, LM focuses on discovering explicit links between the social network's entities represented as a graph [12]. So, for each user of the social network (in our case the email set), the connections of this user are analyzed usually from two angles: volumetric and temporal. Volumetric concerns the number of emails sent or received by a given user. The temporal analysis concerns the response time of such interaction, which is about the time lapse between sending an email to an email address and getting the response to the same email address. These analyses are used to calculate different statistical scores that can help rank and group users. Such links usually exhibit patterns of relationships. Several key indicators were proposed in the literature such as the degree of centrality of a node, clustering coefficient, betweenness centrality, and the social score [1], [2]. Graphs are a natural tool to visualize social media, as it is easy to represent a person or an entity as a node and a connection as an edge.

Unlike previous works, the relations will not be explored here from a binary point of view (there is a relation or not). Rather, they will be explored from a continuous topical point of view, where the nature of the relationship between the involved entities is taken into consideration.

4. The Enron emails set

The Enron email set is a large dataset that is publicly available. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's scandal. It is made of about 500.000 emails (1.32 GB of raw data) from 158 users, exchanged by the Enron's corporation employees during the period of 1998-2002. Given its importance, this corpus has been explored by a wide number of works from disciplines such as SNA, text mining and authorship attribution [13], [1], [14], [15]. This corpus is the right dataset for this study for different reasons. First, it is a realistic dataset about a real forensic investigation case. Second, this dataset is large enough and comes from a representative number of adult users from different age ranges and genders (almost half of the emails are written by males).

An analysis of the email lengths shows that shorter emails are more frequent than longer ones (figure 1). The lengths of about 44% of the emails are within the range 1-30 and the lengths of about 60% of the emails are within the range 1-50. On the other hand, only about 24% of the emails are within the range 100+. These numbers are coherent with the fact that email exchanges within a professional context such as Enron are usually concise. Hence, given this variation in lengths, a

good method for topic classification should be able to process texts with different lengths ranges and to be especially good with short texts.

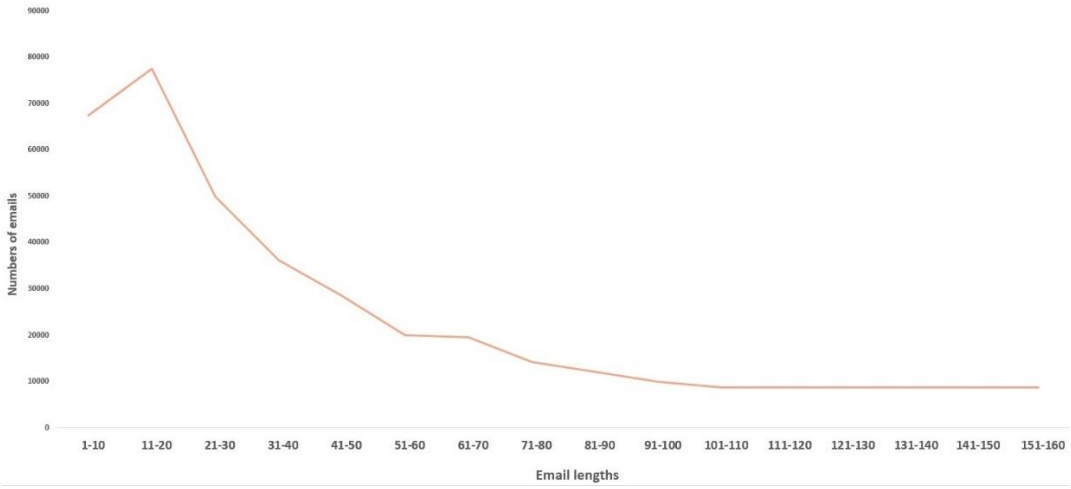


Figure 1. Number of emails by length range within the Enron corpus

The emails are pre-processed as follows. Since the emails contain information in various formats, all emails are reformatted into one uniform JSON file. All the history of previous emails is removed. The removal of everything but the latest response helps reduce the noise related to previous emails from the same exchange. Also, the subject and the author of the email are not used. The body of the emails is cleaned from any non-ASCII characters.

To evaluate the system, the authors with the help of two CS students manually annotated a set of 1999 emails. A schema of eight topics was proposed: meeting, management, IT, leisure, stock, accounting, law, and miscellaneous. The corpus was annotated by four different annotators using the same program that was written specifically for this task.

5. Experimental setup

The data is analyzed in four different approaches and then compared to measure their effectiveness and processing runtime. The first approach is a simple keyword search, which will serve as the basis for comparison for the other three methods. The second and third approaches utilize NLTK's WordNet in a modified and more complex versions of the keyword search, while the fourth version uses a hybrid of LDA and WordNet.

WordNet is a thesaurus-like and psycholinguistically motivated lexical database that was developed at Princeton University [16], [17] (see [18] for an introduction). Within this database,

the synset is the smallest unit and represents a meaning of a word (possibly among others). In addition to a word's explanation and usage examples, WordNet also covers semantic relations such as synonyms, antonyms, hyponyms (a word that has a more specific meaning than another. For example, *lion* is a hyponym of *animal*) and hypernyms (*animal* is the hypernym of *lion*). In the three approaches proposed here, WordNet plays a key role in the mapping between the words, that are too specific, and the concepts that are generic by nature and can be expressed by multiple words.

5.1 Term Frequency (TF)

Words and concepts are tightly interrelated. Hence, it is possible to use keywords as an indication of a topic in a text. This method is usually referred to as Term Frequency (TF) and it was one of the first used methods in Information retrieval [19]. It consists of finding the most frequent word(s) in a document and mapping it into the list of topics. It is widely used in text processing software for keyword search. This method has well-known limitations in the literature [20]. However, given its simplicity and popularity, it is used here as a baseline for our comparison. Some examples of mappings between topics and keywords are provided in table 1.

Table 1. Example of keywords mapping into topics

Topic	Keywords
Management	Management, Guidance, Supervision
Accounting	Accounting, Finances, Balance of Payment
Meeting	Meeting, Board Meeting, Conference, Convention
Stock	Stock, Shares, Growth Stock
Leisure	Leisure, Free Time, Vacation
Legal	Law, Tax Law
Other	No keywords

5.2 Concept based search

The main limitation of keyword-based information retrieval is that only the texts that contain words that match the searched keyword will return positive results. There is a great potential to improperly assign a topic or fail to assign one altogether. To go beyond simple pattern matching, semantic relationships can help widen the spectrum of relevant words found in the text. Hence, a generalization of the extracted keywords is carried out with a list of the keyword hyponyms that is obtained from WordNet.

The searched concepts are defined in two ways. First, the **semi-automatic synset** selection consists of getting every synset for every keyword using WordNet. Then, using WordNet again, every synset for every direct hyponym of each of the original synsets is found and mapped to the keyword. The intention of creating such a large list of synsets is that the increased range of words increases the chances of properly assigning a topic. The second approach, the **manual synset selection,** is a stripped-down version of the first. One specific synset is chosen for each topic and then WordNet is used to make a list of all the direct hyponyms for those specific synsets of the topic keywords. In both versions, when needed, additional synsets can be added by hand to enrich the concept. For example, the topic "accounting" can have the synsets "transaction.n.01" and "value.n.02" added onto the hyponym list in addition to the ones WordNet automatically added. Both of these methods provide us with the final list of target synsets to search in the text.

After defining the searched concepts, the search is carried out as a binary decision process. For every searched concept, the program decides if the text is relevant for this concept or not. WordNet is again used to match the word against a defined concept. The program will read each email once for every topic keyword. As each word in the email is read by the program, the same WordNet function as before is called to find each synset associated with that word. Since the program cannot determine the context of the current word being analyzed, all the synsets of that word are used to match against the current target keyword synset (table 2). It then compares each of those new synsets of the current word in the email to each of the synsets in the current topic's synset list using WordNet's *path_similarity()* function.

Table 2. Example of how each word in the email has all senses of the word analysed against each topic synset

Current word in email	**Current topic being scored**
Transaction	Management
Associated Synsets	**Current Synset being scored**
"expedition.n.01", "expedition.n.02", "expedition.n.03", "excursion.n.01", "dispatch.n.03"	"administration.n.01"

The function returns a score based on how similar a synset is to another synset. If the resulting score is higher than .35, the program assigns that word in the email a point. To prevent topics with larger synset lists from getting more points than those topics with fewer hyponyms, no more than

one point can be assigned per word. After all the words in the email have been compared to all the topic's synsets, the score is added up and divided by the total number of synsets in the email to give the weighted average for that topic. The process then repeats for the rest of the target topics. Since we have decided that an email can have multiple topics assigned to it, a minimum threshold was created to prevent the program from assigning too many topics to an email. The weighted average score for a topic must be greater than or equal to 0.01 to be assigned as the topic to that particular email (see table 3 for an example). Any email where all the topics fail to meet that minimum threshold are assigned the topic of "Other."

Table 3. Example of mapping between the keywords extracted from an email and topics

Email	Topic scores	Chosen Topics
Rick, Attached is the spreadsheet that contains the capital deployed as of June 2001. Let me know if you need anything else. Thanks, Rob Rob Brown Manager, Enron Corp. Financial Accounting Reporting Off. 713.853.9702 Cell 713.303.4497	{'meeting': 0.0, 'accounting': 0.014084507042253521, 'management': 0.0, 'stock': 0.001006036217303823, 'leisure': 0.0, 'IT': 0.0, 'law': 0.0}	Accounting

5.3 LDA

Presented for the first time in [21], Latent Dirichlet Allocation (LDA) is one of the most popular methods for unsupervised topic detection. Based on the Bag of Words approach, LDA is a generative model that builds on the following intuitive idea: every document is made of a mixture of topics and every topic has a distribution of words that is associated with it. The plate notation of the LDA model is provided in figure 2.

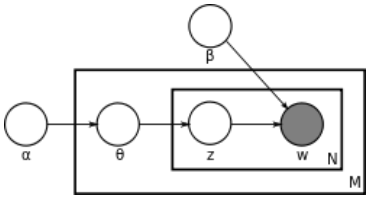


Figure 2. Plate notation for the LDA model

Where:

— α is a Dirichlet prior that represents the topic distributions per-document. A high value for α indicates that each document is likely to contain multiple topics.
— β is a Dirichlet prior that represents the word distribution per-topic. A high beta indicates that each topic will be made of many words.

— θ_i represents the topic distribution for document *i*. One can think of this parameter as document-topic matrix.
— φ_k represents the word distribution for topic *k*. It consists of rows defined by topics and columns defined by words.
— z_{ij} represents the topic for the j-th word in document *i*
— w_{ij} represents a specific word *j* in document *i*.
— M is the total number of documents within the corpus.
— N number of words in a document.

The Gensim library was used to build the LDA model. To make LDA more effective, the vocabulary dictionary of the model is made up of every noun and plural noun that appears in the email messages except for stopwords. The NLTK built-in Part-of-Speech Tagger was used to identify the nouns. To increase the speed of the program, a list of stopwords is created using NLTK's default stopwords. Each email, being used to create the vocabulary dictionary, passes through a Regular Expression to check validity. Words, that do not match the regular expression and thus are not real words, are added to the stopwords list. As the stopwords list grows, the program speeds up because it knows which words it does not have to pass through to the NLTK POS tagger. A training set of emails is then fed to train the LDA model and produce a set of topics. Each topic produced by the LDA model is a list of keywords. Therefore, the raw topics are not usable for this type of analysis in their current state. Hence, alterations must be made before they can be used in the topical analysis module. The top keyword from each topic list is added to the new topic set. A simple check is done to avoid topic keyword repetition in the new set (see figure 3 for an example of the above steps). The final step in preparing the LDA topics for the analysis is the transformation of the topics into WordNet readable synsets. To do this, every hyponym of every synset of the keyword set is found from WordNet. Because the synsets are being chosen for every sense of the word, the resulting dictionary of synsets is significantly larger than that of the WordNet version where a specific synset is chosen for each topic.

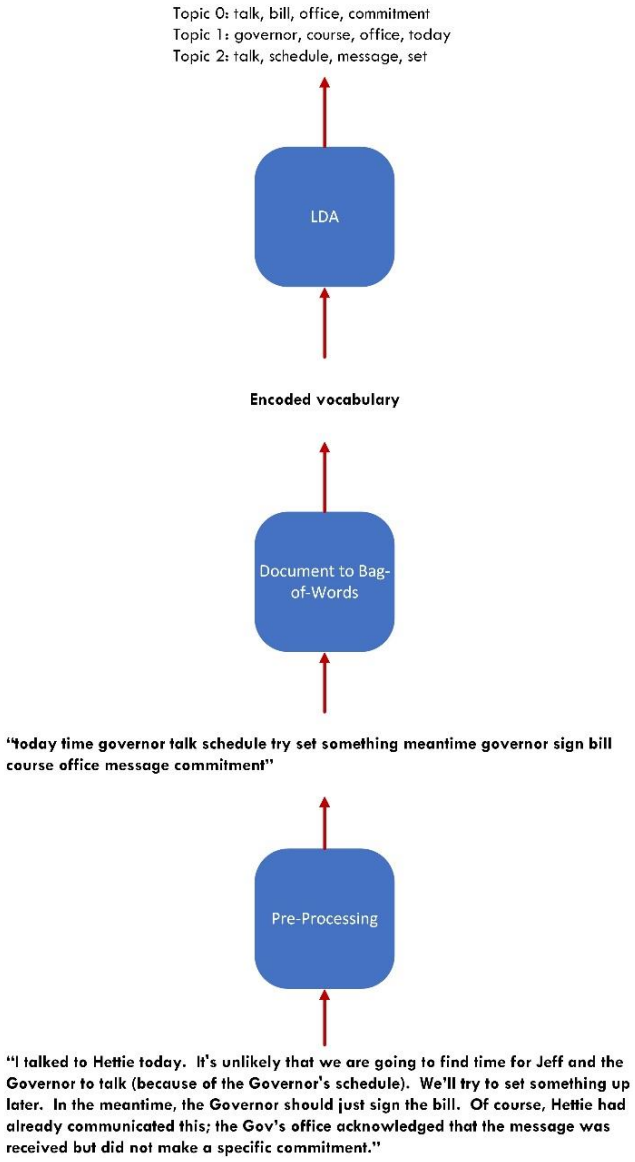


Figure 3. Example of an email topic detection with LDA

6. Evaluation and Results

Given the size of the Enron dataset, it is essential to do performance analysis in terms of processing time. A runtime comparison is depicted in figure 4.

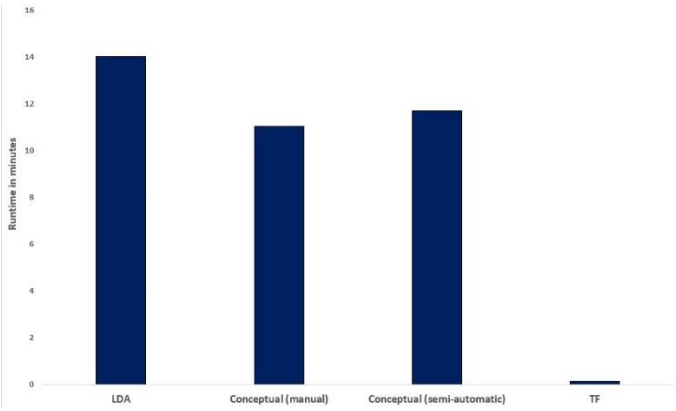


Figure 4. Runtime comparison of the four approaches to complete the topical analysis of the 1999 test emails (in minutes)

As seen in figure 4, the runtime is proportional to the complexity of the method. So, TF is the fastest method given the simplicity of processing, as it analyzed the 1999 test emails with a runtime of 8.88 seconds. On the other hand, LDA, being the most computationally complex approach, requires more time than the three other methods (about 14 minutes). Finally, the difference between the two conceptual methods on the one hand and the LDA approach is not substantial, with the manual synset selection method taking 11 minutes and the semi-automatic synset selection method taking 11 minutes and 45 seconds.

In addition to the runtime, it is important to compare the performance of the adopted approaches. For the testing of the accuracy of the topical analysis, a separate program was created that compared the topics of the 1999 hand-annotated emails to the same 1999 emails that have been run through the different discussed topical analysis programs. The evaluation was performed using the following three following metrics: recall, precision, and F-measure. Since we are dealing with a multiple topic setup, these metrics were calculated based on [22]. Figure 5 depicts the performance of the four adopted measures.

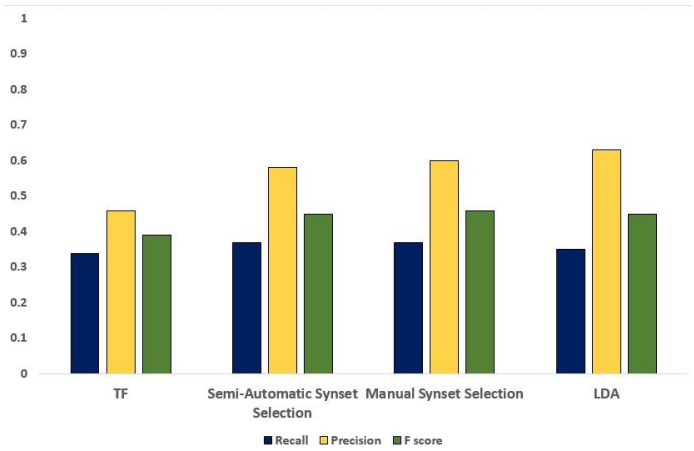
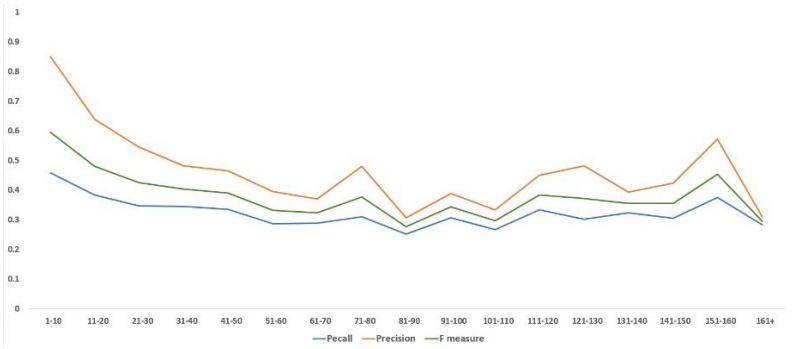


Figure 5. Accuracy Comparison of the Topical Analysis Programs using the 1999 part of the data set that was annotated by hand

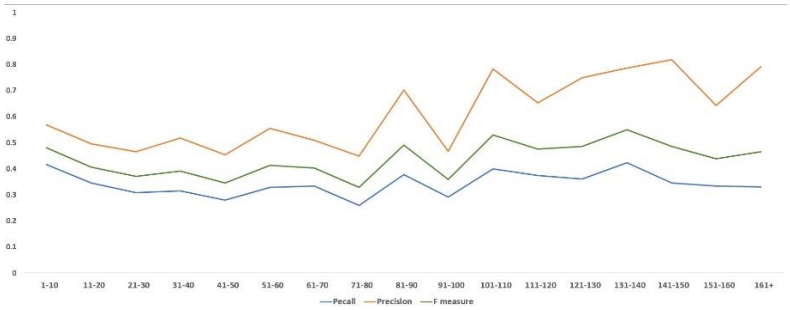
The accuracy of the TF method is the lowest in the three considered measures, with the recall being 0.34, precision being 0.46, and F-score being 0.39. The two conceptual approaches have comparable F-score (0.45 for the semi-automatic and 0.46 for the manual), this score is driven by the small difference in terms of precision (0.58 for the semi-automatic and 0.60 for the manual). LDA has comparable F-score to the conceptual approaches 0.45. However, its recall is slightly lower (0.35), while its precision is slightly higher (0.63).

As we have seen in section 4, there is an important variation in email lengths within the Enron corpus. Therefore, it is important to study how the four different approaches are sensitive to the variation in the texts' lengths. To account for that, we depicted the three adopted measures of performance for every approach by a varying length of the texts (figure 6). The lengths are measured by the number of words. We adopted the word as a measure of length as the lexicon is the basis of topic classification within our four approaches. We also used sixteen gradual length ranges by 10 words, between 1 and 160, as these ranges seem to reflect the variations of lengths as discussed in section 4.

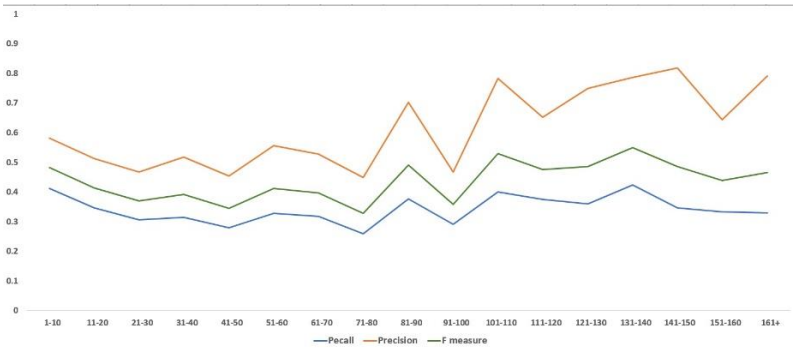
TF



Semi-automatic synset selection



Manual synset selection



LDA

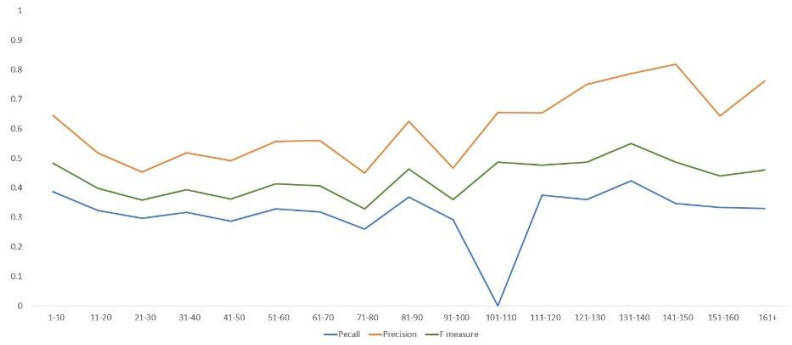


Figure 6. Accuracy Comparison of the Topical Analysis Programs by email lengths using the 1999 part of the data set that was annotated by hand

7. Discussion of the results of topic detection

The results depicted in figure 5 are coherent with our initial expectations. TF has the lowest F-score. This score is driven by a low recall, which is justified by the limited capacity of generalization of the keyword search. This is because TF has a great potential of a mismatch of the topic. The generalization brought by WordNet usage by the three other methods is rewarded by a higher recall, especially with the conceptual methods. On the other hand, the complex processing of LDA gives the highest precision of all the approaches. A final observation about all the approaches, it is easy to see that recall is lower than precision. This suggests that the mapping between the keyword part of all the adopted methods is problematic, as it is hard to build a robust representation of the topics. However, this incomplete representation of the topics is better at building the boundaries between the topics.

The results by email lengths show different patterns. For the keyword-based approach (TF), the tendency is clear: the more words we have, the lower the performance. This is because the increase in size means an increase in the number of candidate keywords. Consequently, within the range 1-10, TF has the highest performance. This can explain the poor overall performance of this approach. On the other hand, the three other approaches seem to better resist the variation in length. Finally, except for TF, precision seems to be more prone to variation with the length increase than recall. Since precision is about adding a new class that is incorrect (e.g. saying that an email about management is also about leisure), the risk of incorrectly adding a new class increases with the increase of the number of words.

8. Visualization of the graph of topics

Following the topical analysis of the emails by the four different approaches comes the creation of the network graphs. This is done with the library NetworkX in Python. In addition to helping draw the graph, this open-source library provides functions for standard graph algorithms as well as different measures of centrality used in SNA. The program goes through every email and gets the “to address” and the “from address” of the email. These will be used to create the nodes and edges of the graph. As the program reads through the emails, it adds weight to the edges between the nodes. If there is no existing node for the “from address” it will create one. It will do the same for the “to address”, but with this one condition changed: if there are multiple recipients of the email, it will check for an existing node for all the “to addresses” and create new ones if necessary. This

allows for the creation of nodes for every recipient of the email. Once the nodes are created, an edge is generated between the sending node and all the recipient nodes with a starting weight of one. NetworkX can detect if an edge already exists, so if there is already an edge, the weight of that edge is increased by one. In addition to keeping track of the weight of each edge, NetworkX can keep track of custom attributes. Using this feature, the weight of each topic is saved to the edge in addition to the edge weight. For example, an edge between two nodes could have the following attributes: 'weight:45, meeting:6, accounting:2, management:7, stock:0, leisure:5, IT:0, law:0, other:25'. These added attributes allow for creating graphs based on different topics. The data visualization tool used to create the examples in this paper is the open-source software Gephi.

To give an idea about the usefulness of the topical annotations of the connections, let's first start with a simple example. Suppose we have the following situation: *a* and *b* exchanged 100 emails out of which 25 are about accounting. Let's suppose as well that *a* and *c* exchanged 60 emails, out of which 40 are about accounting. If we take the absolute numbers, *b* is more connected with *a* than *c*. However, if we are interested in emails about accounting then *c* is more connected with *a*.

Figure 7 A general graph and a topic graph

As we can see in figure 7, the topic graph shows some underlying patterns in the social graph that cannot be seen by observing the general graph.

To give examples of a larger scale, let's start with the graph depicting all the connections within the Enron dataset, presented in figure 8.

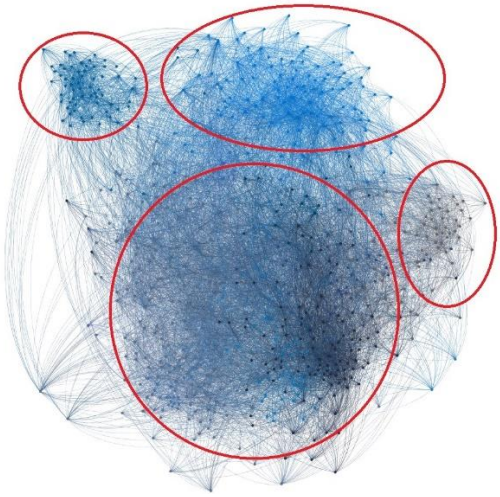


Figure 8 Graph depicting all the unlabeled connection of the Enron dataset

As we can see in figure 8, this graph just gives a very general idea about the connections within the network. One can see there are some clusters (surrounded by red circles) of different sizes. At this level of generality, it is hard to tell if these clusters are about groups of people of similar interests or what is exactly the nature of the interactions within each of these clusters.

If we use the 1999 portion of the dataset. The top 500 nodes (by degree) are provided in figure 9A, along with the top 500 nodes (by degree) with all the nodes of Enron employees that were found guilty highlighted in red that are provided in figure 9B. Despite the reduction of the number of emails, the produced graphs are not very useful as they still give a very general overview of the interaction patterns within the network.

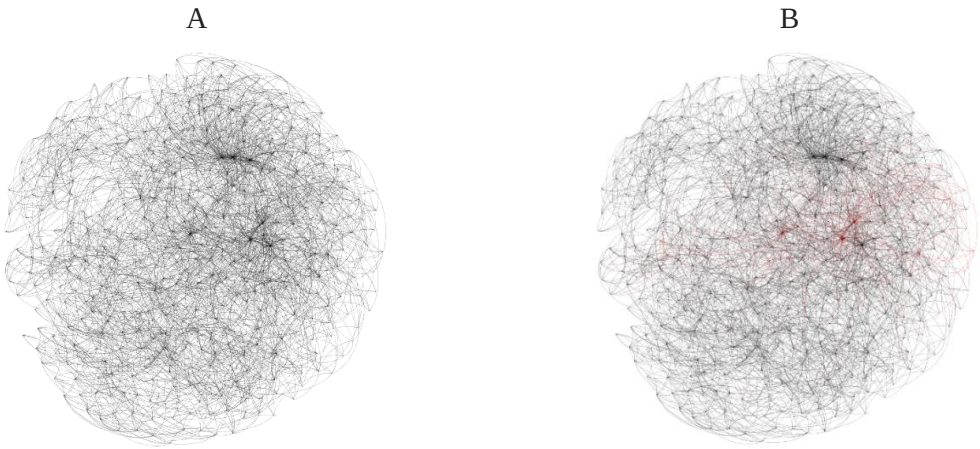


Figure 9 (A) Top 500 nodes without topic – (B) Top 500 nodes without a topic, with the nodes of Enron employees that were found guilty highlighted in red

Another example is provided in figure 10. It is about the top 21 nodes (by degree). This graph showcases how one can find the biggest contributors to a social network. Some of the larger nodes are encoded as follows: A: Jeffrey Skilling, B: Kenneth Lay second email address, C: Jeff Dasovich, D: Tim Belden, and E: Kenneth Lay's primary email address. A surface-level look at this network graph shows that not only do Jeff Dasovich and Tim Belden communicate with a lot of people, they also communicate with each other much more than anyone else. In this example, one might also be interested in who the unlabeled node is that communicates with Kenneth Lay's primary email address. Further investigation reveals that it is his secretary, Rosalee Fleming.

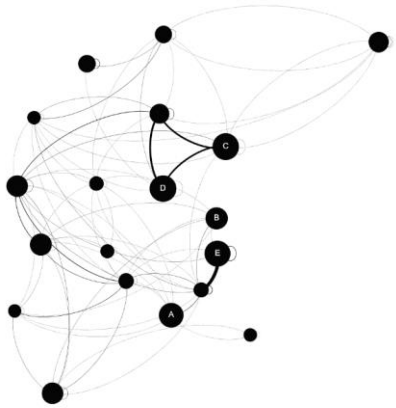


Figure 10. Top 21 nodes with all topics from the full corpus

With the classic method, we can have a more precise view by selecting the weight range (e.g. the most or least connected range). However, despite its precision, this graph does not give any information about the nature of the interactions between the nodes.

In figure 11, we have the employees with the highest count of unique email recipients with the topic of *law* as analyzed with the manual synset selection method. This is an example of how one could analyze a graph to see who talks to most people about a certain topic. In addition to that, we can see who they talk to the most by seeing the larger edges.

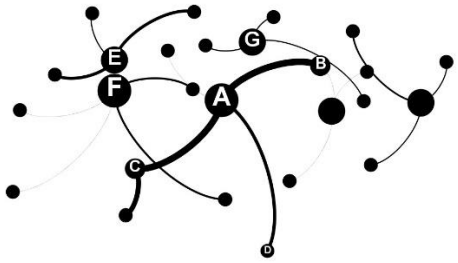


Figure 11. Topic Law from the entire corpus, with A - michelle.akers@enron.com, B - phillip.allen@enron.com, C - eric.bass@enron.com, D - anne.bike@enron.com, E - kayla.harmon@enron.com, F - jeff.dasovich@enron.com, and G - araceli.romero@enron.com

In addition to being useful for understanding the social and professional relations underlying the network, graph visualization is useful for evaluating the performance of the adopted topic detection algorithms. To do so, we generated the meeting of the *meeting* topic using the 1999 human-annotated emails as well as the equivalent graphs generated by three topic detection methods as an example (figure 12). To make the comparison easier, we selected specific nodes that seem important on the human graph and highlighted them in other graphs. The red node represents the user Steven Harris. The red node is present in all the graphs, but we can see that it is much less important in the semi-automatic and keyword search methods than it is in the human method and the manual synset method. Furthermore, the blue node represents Mike Grigsby. This node, which is prominent in the human graph, is only present in the manual synset graph. The yellow node that represents Mitch Robinson is found in only the human graph, meaning that all the other methods failed to categorize any of their emails with the topic "meeting". The orange node is based on what is most prominent in the manual synset method. As we can see, the orange node is also found in the human graph and the semi-auto graph, but not in the keyword search graph. It is interesting to note that what seems central in the WordNet methods is much less prominent in the human graph.

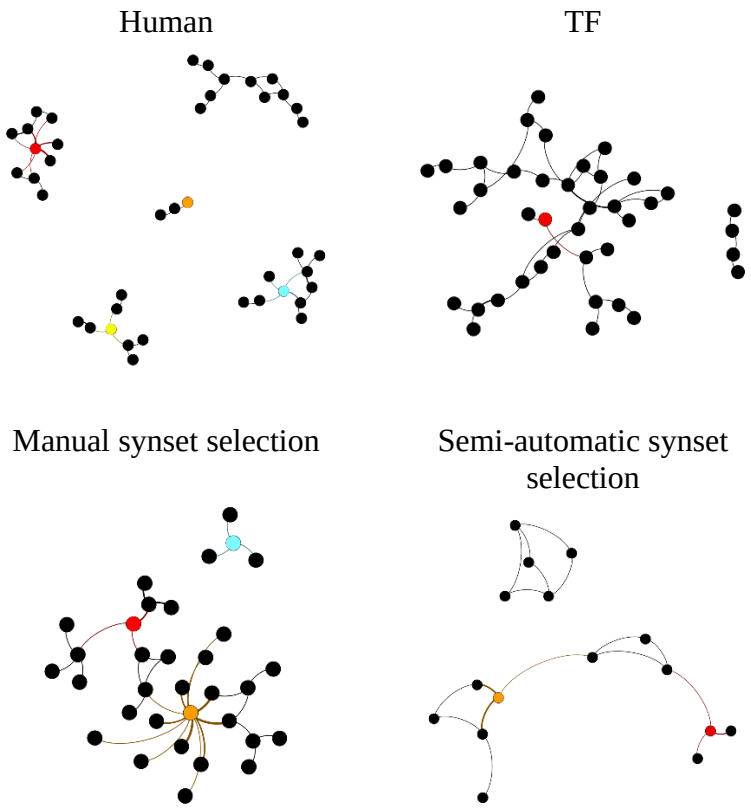


Figure 12. Comparison of partial graphs generated with three topic detection methods with the graph based on the human labeled data

9. Conclusions

This paper is about visualizing social networks with graph topics. It is made of two main parts. In the first part, we compared four different approaches to topic detection using the Enron dataset. The results showed that LDA outperformed the three other approaches especially in terms of precision. In the second part of the work, we demonstrated how topic graphs can shed light, not only on the quantitative aspects of the relations between the users but also on the nature of these relations through the topics. We also showed that the graph generation method can also bring interesting information about the qualitative performance evaluation of the topic labeling methods.

Several perspectives of this work are being explored. First, a larger scale comparison of different rule-based and unsupervised machine learning approaches is being carried. In addition, a deeper exploration of the role topics within the graph is also being conducted. Finally, the application of

the concepts presented in this paper is being considered to other types of social networks like Twitter and Facebook.

10. References

[1] J. Diesner and K. Carley, "Exploration of communication networks from the Enron email corpus", In Proceedings of Workshop on Link Analysis, Counterterrorism and Security, Newport Beach CA, 2005.

[2] J. Diesner, T. L. Frantz, and K. M. Carley. "Communication networks from the Enron email corpus". Journal of Computational and Mathematical Organization Theory, 11:201–228, 2005.

[3] Ryan Rowe, German Creamer, "Automated Social Hierarchy Detection through Email Network Analysis", Joint 9th WEBKDD and 1st SNAKDD Workshop'07 August 12, San Jose, California, USA, 2007.

[4] Gerard Salton, ed., The SMART Retrieval System. Englewood Cliffs, N.J. Prentice Hall, 1971.

[5] Everitt B.S., Landau S., Leese M., Stahl D., Cluster Analysis, 5th Edition, Wiley, 2010.

[6] Halabi Ammar, Ahmed-Derar Islim, Mohamed-Zakaria Kurdi, A hybrid approach for indexing and retrieval of archaeological textual information, International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, /9/8, pp 527-535, 2010.

[7] Shafiq Joty, Giuseppe Carenini, Gabriel Murray, and Raymond Ng, Finding Topics in Emails: Is LDA enough? 2009 https://raihanjoty.github.io/papers/joty-carenini-ng-lda-nips-09.

[8] Veselin Stoyanovand, Claire Cardie, Topic Identification for Fine-Grained Opinion Analysis, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 817–824 Manchester, August 2008.

[9] Chin-Yew Linand, Eduard Hovy, The Automated Acquisition of Topic Signatures for Text Summarization, COLING '00 Proceedings of the 18th conference on Computational linguistics - Volume 1 Pages 495-501, Saarbrücken, Germany — July 31 - August 04, 2000.

[10] W. Holmes Finch Maria E. Hernández Finch Constance E. McIntosh Claire Braun, "The use of Topic Modeling to Analyze Open-Ended Survey Items", OpenMx XSEM with Applications to Dynamical Systems Analysis, May 22, 2017.

[11] Jinsong Su, Deyi Xiong, Yang Liu, Xianpei Han, Hongyu Lin, Junfeng Yao, Min Zhang, A Context-Aware Topic Model for Statistical Machine Translation, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th

International Joint Conference on Natural Language Processing, pages 229–238, Beijing, China, July 26-31, 2015.

[12] Lise Getoor, Christopher P. Diehl, LinkMining: A Survey, SIGKDD Explorations, 7(2), pp 3-12, 2005.

[13] Man Wang, Minghu Jiang, Text categorization of Enron email corpus based on information bottleneck and maximal entropy, ICSP2010 Proceedings, 2010.

[14] A. McCallum, X. Wang, A Corrada-Emmanuel, Topic and role discovery in social networks with experiments on Enron and academic email, Journal of Artificial Intelligence Research, Volume 30 Issue 1, September Pages 249-272, 2007 .

[15] M. Zakaria KURDI, Content-dependent vs. content-independent features for gender and age range identification in different types of texts, International Florida Artificial Intelligence Society FLAIRS-33, May 19-22, 2019, Sarasota Florida.

[16] George A. Miller, WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41, 1995.

[17] Christiane Fellbaum, WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998.

[18] M. Zakaria KURDI, Natural Language Processing and Computational Linguistics 2: Semantics, Discourse and Applications, London, ISTE-Wiley. ISBN: 978-1-848-21921-2, 2017.

[19] Gerard Salton, and Michael J. McGill, Introduction to Modern Information Retrieval, New York, McGraw-Hill Book Company, ISBN 0-07-054484-0, 1983.

[20] Furnas, G. W., T. K. Landauer, L. M. Gomez, and S. T. Dumais, “The vocabulary problem in human-system communications”, Communications of the ACM 30(11), 964-971, 1987.

[21] David Blei, Andrew Ng, Michael Jordan, and John Lafferty, (ed.). "Latent Dirichlet Allocation". Journal of Machine Learning Research. 3 (4–5): pp. 993-1022, 2003.

[22] Marina Sokolova, Guy Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing and Management 45, pp 427–437, 2009.