September 12, 2018

PERSONNEL KNOWLEDGE SEARCH SYSTEM

HP INC

Personnel Knowledge Search System

Abstract

People within large originations have difficulty finding individuals or groups who have expertise or products for any specific domain. This is largely due to the difficulty of tracking what each individual or group is working on in a dynamic environment. The proposed solution is to create a Search Engine which can query individuals based off their domain expertise. The proposed solution uses the $Doc2Vec$[3] model to produce topic vectors for every email or other textual communication by employees. The communication topic vectors provide spatial correlated topics where similar topics will be found close together, and dissimilar topics will be found far apart from each other. Converting textual communication into spatial related topic vectors allows for the use of well proven and robust correlation models such as K Nearest Neighbors (K-NN) to find appropriate similarities.

1. Introduction

With the ever-changing market, and growing organization sizes, it is difficult for individual employees to identify internal resources and contacts that can assist their project. Missing these important resources and/or contacts has devastating effects on go-to-market time, project costs, overall company agility, and more. The issue with identifying such resources and contacts is the dynamic environment of the organization, new personnel enter, others leave, projects get started and pivot regularly. Additionally, each company has different methods for tracking projects and employee roles. These tracking methods are typically not open for an average employee to search, nor are they typically very accurate as often an employee's job description and actual work deviate over time. To tackle this problem, a solution which can dynamically learn the employee's tasks and the team's projects without the aid of manually entered information is needed. Fortunately, individual's emails, texts, and other communications trough work channels can provide a strong indication on their project tasks and their domain knowledge. By training the $Doc2Vec$[3] model on a large corpus of diverse topics, a robust topic vectorization model can be created and utilized to infer the topics of each text document and search term. The search term vector can then be easily correlated with the top *K* communication topic vectors via the K Nearest Neighbors algorithm, and the most relevant contacts from the top *K* correlations can be returned to the searcher.

2. Solution

The proposed solution uses a pretrained $Doc2Vec$[3] model to infer the topic vectors of user emails, text, or search queries. The topic vectors for user emails and text can be pre-calculated and cached, thus increasing the performance, decreasing the memory overhead by only storing the vector, and greater privacy by removing the human-readable content. Each user's topic vectors will provide a computer searchable space with the assumption that individual's conversations and emails focus on their field of work. The assumption that a user's topic at work is focused on a single domain is tested and validated with the $Enron\ Email\ Corpus$[2] in the Description of Experiment section of this paper. With assumption of user topics focusing on a specific domain, an *averaging* algorithm can be applied to the user's content thus creating a single topical point. For this step, the *Centroid Mean* algorithm as used in *K-Means Clustering* was used. This algorithm gives a weighted average that focuses on desired features. The reduction of the user's knowledge area into a single vector enables a manageable memory overhead and allows comparison algorithms such as *k-nearest neighbors* to be a viable option when searching for similar content. Search terms from users would be processed and have their topic vector inferred via the same pretrained $Doc2Vec$[3] model used previously. An algorithm such as *k-nearest neighbors* can then be used to obtain the *N* users with the closest topics as the search phrase. These *N* users can be determined using the *average* user topic as derived by the *centroid mean* algorithm. Each selected user would then be ranked based off their individual email and other text vectors similarity with the search term, time the text was created, and other metrics. The top *N* users sorted by salience will have their contact information returned to the searcher.

3. Description of Experiment

The crux of this solution lies within the *Clustering* and *Centroid Mean* algorithm. To proof this system is viable, machine learning models were created to perform these functions, and validated on a dataset of real corporate emails ($Enron\ Email\ Corpus$[2]).

For the Clustering algorithm, a $Doc2Vec$[3] architecture was trained to vectorize Documents on $20\ Newsgroups$[1] Dataset. The dataset was preprocessed using stemming, lemmatizing, replacing all numbers with pound symbols (#), and converting all characters to lowercase. The trained $Doc2Vec$[3] model was then used to infer the vectors for every email in the $Enron\ Email\ Corpus$[2]. The vectorized emails produced from the pretrained $Doc2Vec$[3] model showed that users had tight clusters of emails. This indicates that the user's topics at work center around a specific domain as assumed. See fig. 2 for a visualization of the email vector-space.

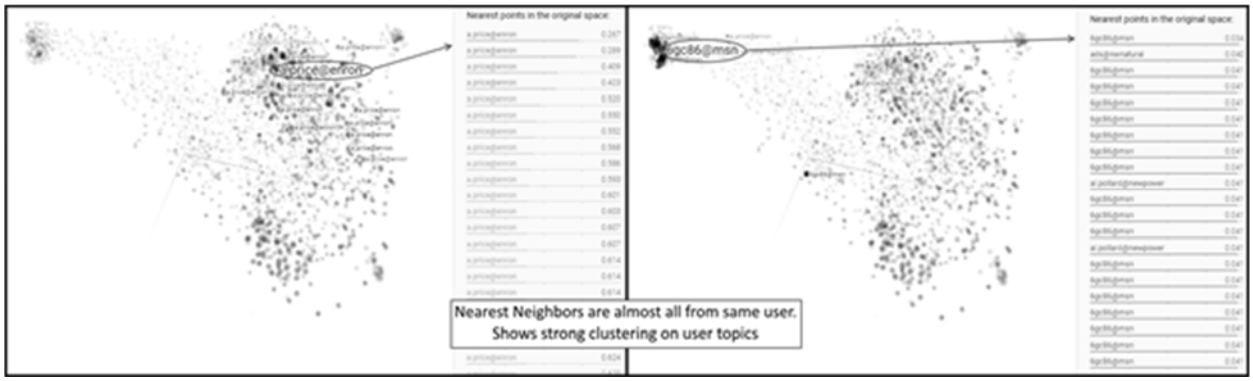

Fig. 2. Visualization of 5000 email vectors after t-SNE dimensionality reduction. This figure shows emails from various user are tightly clustered.

For the Centroid Mean algorithm, a similar algorithm to the k-means clustering was used to find the *average* user topic vector. The result from the centroid mean algorithm show distinct clusters of user topics. From the visualizations in Fig 3. it is apparent that users who talk about similar topics are appear closely in the vector-space. For this demonstration, the promotional emails from American Express and Forbes were identified as their names are apparent as to which topic they discuss.

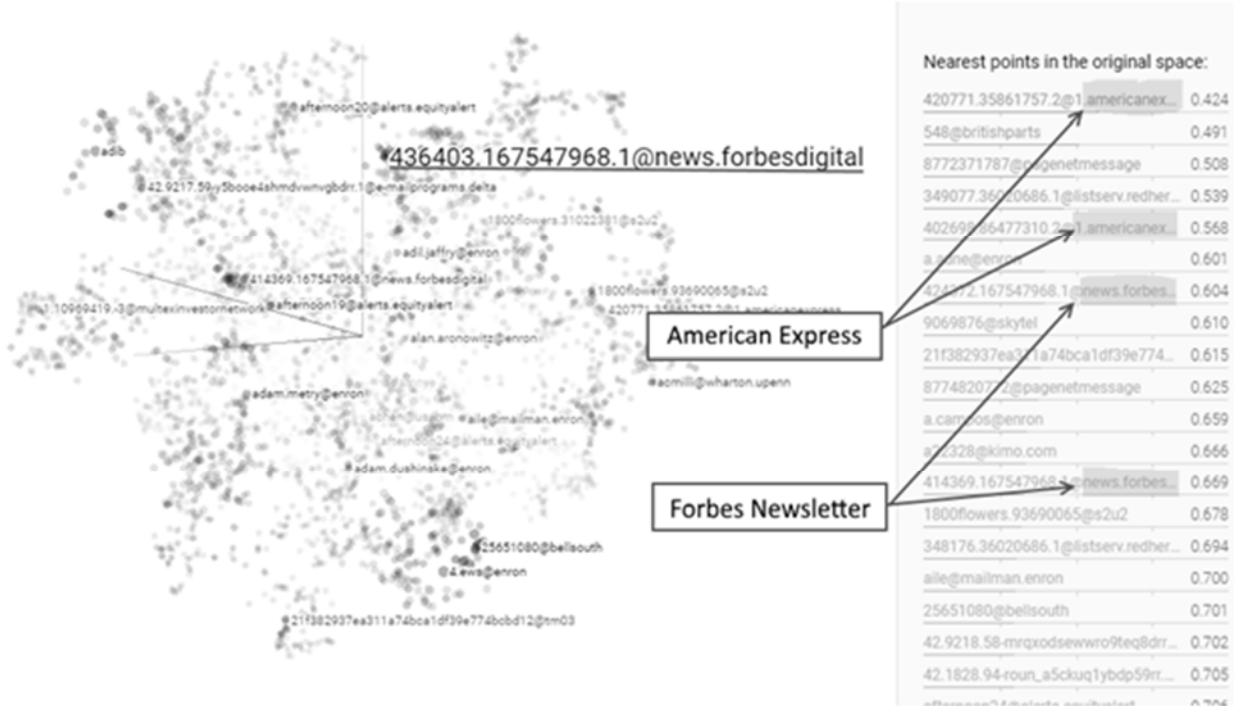

Fig 3. Visualization of 5000 user topic vectors after t-SNE dimensionality reduction. This figure shows that emails from American Express and Forbes are closely related.

4. Next Steps

For the next steps, a more robust technique to score the cluster spread of individual's corpus should be used. This will provide a measurement to indicate the focus or diversity of a person's topic and can lead to a method to compare different topic-vectorizing models capabilities. Additionally, the $Doc2Vec$[3] model should show better accuracy in generating the topic vectors if it was trained on a larger corpus of data. Lastly, the search phrase will likely not provide an accurate topic vector without preprocessing done to insure it is not in a question format. Therefore, various methods of extracting the most salient words and/or mapping them to a valid sentence to be vectorized accurately will need to be researched.

5. References

1. *20 Newsgroups data set.* T. Mitchell et al., 1999
2. *Enron Email data set.* E-Discovery & Information Governance, 2011.
3. *Distributed Representations of Sentences and Documents*, Q. Le, T. Mikolov, Google Inc., 2014

Disclosed by Lucas Randall Pettit, Mohit Gupta and Sudish Mukundan, HP Inc.