Sample[*] Diachronically Like-minded User Community Detection[*]

Ali Fani
University of Windsor
afani@uwindsor.ca

Hossein Fani
University of Windsor
hfani@uwindsor.ca

ABSTRACT

This is a *sample* evaluation report for your proposal. The sections and subsections are *by no means* fixed and should be indeed changed or customized according to the proposal.

KEYWORDS

More Specific Keywords, Specific Keyword, General Keyword, More General Keyword

1 EXPERIMENTAL SETUP AND EVALUATION

1.1 Dataset

In our experiments, we use a publicly available Twitter dataset collected and published[1] by Abel et al. [1]. It consists of approximately 3M tweets posted by 135,731 unique users between November 1 and December 31, 2010. In addition to its text, each tweet includes user id and timestamp. The whole two months time period is sampled on a daily basis, i.e., L = 61 days.

1.2 Setup

Our proposed approach consists of three phases to identify temporally like-minded user communities; finding topics, building user vector representations, and detecting user communities. Here, we provide the implementation details and the setup of our approach in each of these phases.

1.2.1 Finding topics. Extracting topics from tweets suffers from the sparsity problem when topic modeling methods such as LDA are used [10]. As suggested in [11, 12], we annotate each tweet with entities defined in Wikipedia to obtain better topics from Twitter with no change in the underlying topic detection methods. For instance, for a tweet such as '*NATO Leaders Seek Time on Afghan Exit Strategy - http://nyti.ms/cMMDuR*', a semantic annotator such as TagMe [3] is able to identify and extract several Wikipedia entities, namely 'NATO'[2], 'Afghan', and 'Exit_Strategy'. Using entities instead of words can lead to the reduction of noisy content within the topic detection process, because each concept implicitly represents a collection of typical terms which are collectively more meaningful than a single word or a group of less coherent words [8]. We annotated the text of each tweet with Wikipedia entities using the TAGME RESTful API[3], which resulted in 350,731 unique entities.

In order to find topics of interest in our dataset, we have applied MALLET[4] for LDA. LDA-based approaches to topic detection need *a priori* knowledge for the number of topics. The number of topics has been already investigated and set to 50 for the same tweet dataset by other researchers in [2]. We populate the points of temporal interest (PoTI) for our topic set $\mathbb{Z}$ on a daily basis, i.e., L = 61 days, and screen out values less than 0.1. The condition for homogeneity c is set such that the difference of values falls in the range $[0, 0.1)$.

1.2.2 Building user vector representation. We extended CBOW architecture in Gensim[5] to learn user embeddings as already introduced in this paper. The training phase uses a learning rate of 0.025 and in each epoch we decrease it by 0.002 for 200 epochs. We perform the experiments on different vector sizes of $d = 100, 200, ..., 500$ in an increasing order till we see no further performance gain.

1.2.3 Detecting user communities. We build temporal topic-based communities according to our proposed approach in Section **??**. We build the weighted graph G and apply the Louvain method with resolution parameter 0.1 using Pajek[6]. This leads to our temporal topic-based communities $\mathbb{P}^*$.

1.3 Baselines

We compare our work against the following baselines whose details has been already given in the related work section:

Fani et al. [2]. This approach models user's contributions toward the topics of interest through a multivariate time series. We use LDA in its topic detection step with 50 topics and build the time series for daily time intervals L= 61 days in its user modeling step. The approach uses two dimensional cross correlation to measure the similarity of a pair of users' time series. We use the implementation in MATLAB[7] for calculating time series cross-correlation. Finally, we use the Louvain method in Pajek for its community detection step as proposed by the authors.

Hu et al. [4] . This is a parametric unified probabilistic generative model for topics and communities. The number of topics is set to 50 and we perform experiments on increasing number of communities for $C = 5, 10, ..., 30$ till we see no performance gain. The number of iterations is set to 1,000. This method is a mixture model in which all users are members of all communities with a probability distribution. In our comparison, we only consider the community with the highest probability as each user's community.

Figure **??** provides an overview of the distribution of users across different communities. For Hu et al.'s work, the number of communities needs to be specified as shown ranging from 5 to 30. For our proposed approach, the number of communities is automatically determined by the graph partitioning method; however, the size of the embeddings needs to be provided, which has been set from 100 to 500. As seen in the figure, our proposed method leads to a more fair distribution of users across communities while the two baseline methods have a higher skewness in the distribution of users in

[*] Link to Github or an online repo
[1] www.wis.ewi.tudelft.nl/umap2011/
[2] en.wikipedia.org/wiki/NATO
[3] services.d4science.org/web/tagme/documentation
[4] mallet.cs.umass.edu/topics.php
[5] radimrehurek.com/gensim/models/word2vec.html
[6] vlado.fmf.uni-lj.si/pub/networks/pajek/
[7] www.mathworks.com/help/signal/ref/xcorr2.html

their identified communities. While this by itself is not a measure of community quality, as we will show later, disproportionate distribution of users in communities could lead to poor application level performance.

1.4 Evaluation Protocol and Gold Standard

On the one hand, contrary to typically small scale networks or synthetic ones, gold standard communities for real social networks are not available. So, well-defined quality measures such as Rand index, Jaccard index, or normalized mutual information (NMI) that require comparison to a gold standard are not applicable. On the other hand and in the absence of a golden standard, quality functions such as *modularity* are not helpful either since they are based on the explicit links between users, which are not applicable to our work. For instance, in the context of our work and the baselines, a perfect community detection algorithm might have a low modularity value as those users that are deemed most similar might not have explicit social connection with each other. Therefore in our work, the communities that achieve high modularity are not necessarily optimal from temporal and topical points of view [6].

Fortunately, the performance of community detection methods can be measured through observations made at the application level, as suggested in [5, 6]. In these evaluation strategies, a temporal like-minded user community detection method is considered better iff its output communities improve an underlying application. We deploy three applications: news recommendation, user prediction, and community selection. Note should be taken that we do not attempt to improve the state of the art in any of these three applications but rather to show that the application of the proposed community detection method is able to provide a stronger performance compared to the other two state of the art community detection baselines.

To this end, we first build a gold standard dataset for the said applications by collecting news articles to which a user has explicitly linked in her tweets (or retweets). We postulate that users post news article urls since they are interested in the topics of those news articles. Similar to tweets, we annotate news articles with Wikipedia entities. We build the gold standard from a set of news articles whose urls have been posted by user u at time t. We see each entry as a triple (u, a, t) consisting of the news article a, user u, and the time t. As a result, $\mathbb{G} = \{(u, a, t) : u \in \mathbb{U}, a \in \mathbb{A}, 1 \leq t \leq L = 61\}$ forms our gold standard where $\mathbb{U}$ and $\mathbb{A}$ are sets of users and news articles. The gold standard $\mathbb{G}$ consists of 25,756 triples extracted from 3,468 distinct news articles posted by 1,922 users.

1.5 User Prediction

Another application with which we evaluate our approach and the baselines is the user prediction application. Given the gold standard $\mathbb{G}$ and the like-minded user communities $\mathbb{P}^*$, this time the goal is to predict which users posted the news article a at time t. To do so, we find the closest community to the news article in terms of topics of interest at time t. Then, the members of such community would constitute our prediction list. We employ precision, recall, and f-measure to report user prediction performance. We summarize the results for these metrics in Figure 1. As shown, in terms of precision, our approach with all different dimensions except d = 100 outperforms all other baselines. In terms of recall, however, Hu et al. [4] with $C = 5$ competes with our proposed approach when $d = 400$. The reason for such high recall in Hu et al. with C=5 can be attributed to the lower number of communities in this method. The fewer the number of communities are, the higher the recall of the method would be. In other words, if we only have one community that includes all users, recall would be 1.0. As the number of communities increases from $C = 10$ to $C = 30$ in Hu et al, recall decreases which supports our explanation. Overall, F-measure shows the superiority of approach in user prediction in all its variants except for $d = 100$, which is weaker than Fani et al.

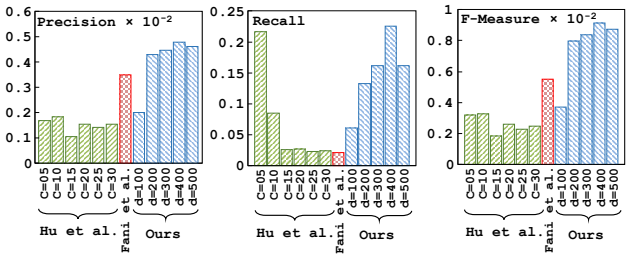

Figure 1: Comparative performance on the user prediction application.

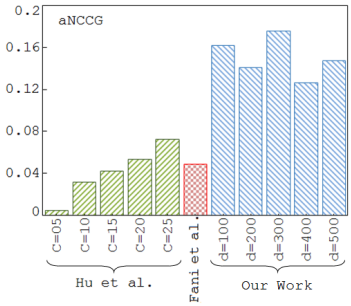

Figure 2: Performance on community selection application.

1.6 Community Selection

In the realm of cluster-based information retrieval systems, the entire collection of documents are split into clusters such that only the documents in highly related clusters to a given query are accessed. As a result, fewer documents are searched from within a large collection of documents which results in improved response time. Better clustering solutions in this context are those that can group relevant documents for previously unseen queries. This approach is referred to as *collection selection* and normalised cumulative cluster gain (NCCG) [7] is a metric used for evaluating collection selection. According to NCCG, the best clustering would be the one where all the documents related to a given query are all located in the same cluster. The worst clustering is the one where the relevant documents to an input query are scattered across many clusters. NCCG is the difference between the current clustering gain and the worst possible, formulated as follows:

$$NCCG = \frac{s - s_{min}}{1 - s_{min}} \quad (1)$$

where $s = \frac{\sum^{|g|} cumsum(g)}{n^2}$ and g is a sorted gain vector whose elements represent each cluster's gain, i.e., the number of relevant documents in a cluster, n is the total number of relevant documents and *cumsum* represents the cumulative sum of a vector. The worst possible gain s_{min} happens when the relevant documents to the query are uniformly distributed over all clusters.

However, NCCG has been criticized for being sensitive to the number of clusters and population distribution; therefore, De Vries et al. [9] have proposed an *adjusted* version of NCCG (aNCCG), i.e., NCCG's divergence from a random null base model, to alleviate such problem. We employ aNCCG to evaluate the temporal and topical coherence of the identified output communities of the different approaches in the application of community selection as follows: given a news article a at time t (the input query), we want to find the communities of those users (similar to documents related to an input query) who have mentioned the news article at that time. The output user communities are more effective iff users who mention a news article a (topical) at time t (temporal) are all located in one community instead of being distributed across several communities. We report aNCCG for our approach and the baselines in Figure 2.

As seen in the figure, our approach, for different number of dimensions, outperforms the other two baselines in terms of aNCCG. This means that, in our approach, the users who mention the same news articles in specific time intervals are placed within similar user communities, i.e., such users are distributed across fewer communities. A lower aNCCG value as exhibited by Hu et al. and Fani et al. means that these methods distribute users that have posted similar news articles at specific time intervals across a larger number of user communities, which is not desirable.

REFERENCES

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. [n.d.]. Analyzing User Modeling on Twitter for Personalized News Recommendations. In *UMAP 2011*.

[2] Hossein Fani, Ebrahim Bagheri, Fattane Zarrinkalam, Xin Zhao, and Weichang Du. 2017. Finding Diachronic Like-Minded Users. *Computational Intelligence: An International Journal* (2017).

[3] Paolo Ferragina and Ugo Scaiella. [n.d.]. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *CIKM 2010*.

[4] Zhiting Hu, Junjie Yao, and Bin Cui. [n.d.]. User Group Oriented Temporal Dynamics Exploration. In *The 28th AAAI Conference on Artificial Intelligence, 2014*.

[5] Zhiting Hu, Junjie Yao, Bin Cui, and Eric P. Xing. 2015. Community Level Diffusion Extraction. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*. 1555–1569.

[6] Farnaz Moradi, Tomas Olovsson, and Philippas Tsigas. 2012. An Evaluation of Community Detection Algorithms on Large-Scale Email Traffic. In *SEA 2012*.

[7] Richi Nayak, Christopher M. De Vries, Sangeetha Kutty, Shlomo Geva, Ludovic Denoyer, and Patrick Gallinari. [n.d.]. Overview of the INEX 2009 XML Mining Track: Clustering and Classification of XML Documents. In *Focused Retrieval and Evaluation, 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009*.

[8] Georgios Petkos, Symeon Papadopoulos, Luca Maria Aiello, Ryan Skraba, and Yiannis Kompatsiaris. [n.d.]. A soft frequent pattern mining approach for textual topic detection. In *WIMS 2014*.

[9] Christopher M. De Vries, Shlomo Geva, and Andrew Trotman. 2012. Document Clustering Evaluation: Divergence from a Random Baseline. *CoRR* abs/1208.5654 (2012).

[10] Fattane Zarrinkalam and Ebrahim Bagheri. 2017. Event identification in social networks. In *Encyclopedia with Semantic Computing and Robotic Intelligence*. Vol. 01. 1630002 [7].

[11] Fattane Zarrinkalam, Hossein Fani, Ebrahim Bagheri, and Mohsen Kahani. [n.d.]. Inferring Implicit Topical Interests on Twitter. In *ECIR 2016*.

[12] Fattane Zarrinkalam, Hossein Fani, Ebrahim Bagheri, and Mohsen Kahani. [n.d.]. Predicting Users' Future Interests on Twitter. In *ECIR 2017*.