Building Semantic Profiles of Social Network User Groups

Aluno: Paulo Viana Bicalho [1], Orientadora: Gisele Lobo Pappa [1]

[1]Universidade Federal de Minas Gerais (UFMG)

Abstract. *Information about users in social networks is highly valuable for understanding user preferences regarding products, brands or politics. Many works have tried to cluster or classify each user according to her common interests. This paper, in contrast, proposes an approach to characterize groups of users with a common interest, expressed through relationships in social networks. In the characterization process, users preferences are defined by the semantic topics they most discuss. In order to find semantic topics, we propose a method that successfully merges topics found through matrix factorization by performing random walks in a latent-topic transition graph. Experiments were performed in documents datasets for easier evaluation and in social network datasets. The method showed to be scalable and accurate in identifying the semantic topics inside a selected group of users.*
Keywords: users profiles, topic discovery, NMF

1. Introduction

Twitter has already proven to be a powerful source of information for monitoring trends, sentiments and topics, but to this date not a lot is known about Twitter users. The majority of studies regarding users focuses on issues such as the demographics of Twitter, community detection and measures to assess user influence. However, only a few works have focused on another very interesting issue: user profiling[Abel et al. 2011a].

A user profile may contain different elements, including user preferences (as expressed in messages), personal information (such as gender and location) or social network behavior (usage time and number of posts per day, among others). In this work we define users' profiles based on their general preferences, where a preference is defined by a *semantic topic* discussed in their posts.

In contrast with previous methods for creating user profiles for specific user accounts [Pennacchiotti and Popescu 2011, Abel et al. 2011b], here we are interested in generating profiles of *groups of users*. These groups comprise users that share at least one common interest, which may become explicit when they follow the same Twitter account or frequently discuss the same topic. Identifying such groups and profiles is of great value for companies, which can take advantage of them for targeted marketing, recommendation, or to better understand users preferences over time.

By profiling groups of users we avoid problems of privacy and data scarcity while being able to determine the underlying habits and preferences [Zheleva and Getoor 2009]. On the other hand, there are several challenges in such profiling task. The first challenge arises when determining the semantic topics that characterize the users, since they must be, simultaneously, the *smallest representative* set of *highly cohesive* and *non-fragmented* topics observed in a set of messages. The second challenge is related to the efficiency of the profiling method, since the number of posts is usually huge and increasing. The third challenge is the lack of context associated with microblog messages such as Twitter, making the topic detection, and thus profiling particularly challenging. The fourth challenge is, given a set of semantic topics, to identify the users that will compose each group, coupling with their diversity and the complexity of the topic. Finally, we may summarize

the profile of each group based on the semantic topics that arise in the group and their intensity.

Formally, considering that we extract semantic topics from users posts, any set of users U can be characterized by a weighted set of preferences $<p_1,p_2,\ldots,p_n>$, where p_x represents a semantic topic identified from the posts of the group. Although we focus just on topics as preferences in this work, other personal user characteristics may be added to the preference set later.

In order to build semantic profiles of groups of users, we propose the UPsCAle (User Profile CreAtor) framework, which is organized in three main phases: (i) it identifies semantic sub-topics of a user group using a traditional matrix factorization method; (ii) it merges semantic sub-topics into more cohesive and unique semantic topics; (iii) it maps the final set of semantic topics into users profiles. In the first phase, we benefit from the fact that the task has been extensively studied [Pons-Porrata et al. 2007], and employ a Non-negative Matrix Factorization (NMF) [Berry et al. 2007], which is capable of generating good-quality topics, despite vocabulary overlaps. However, these semantic topics may still lack cohesion[Cheng et al. 2013], and we propose a new method for phase (ii), based on Markovian theory, for merging topics in order to generate more cohesive topics. Finally, the desired profiles are then generated based on a simple strategy that has shown to perform surprisingly well.

In summary, this paper has two main contributions. The first, a new generic semantic topic identification method that can be used in any topic detection task and optimizes the three contradictory but desirable properties in topic identification simultaneously: representativeness, cohesion and non-fragmentation. The method is also scalable, efficient and handles the lack of context usually inherent to microblogs and other social media platforms. The second contribution is a simple yet effective method to define users profiles considering groups of users, represented by a weighted set of preferences that initially correspond to semantic topics.

UPsCAle was evaluated in two phases. First, we tested the topic identification method in both text and social network collections where the main topics were known. Second, we performed a case study with over 50,000 users that are Twitter followers of Barack Obama, considering more than the 700,000 messages posted during the American Elections in 2012. The results showed that the method is able to generate less fragmented and more concise topics when compared to other state of the art methods.Such topics enabled the determination of good quality profiles, as the results show.

2. Related Work

This section reviews methods for both semantic topic identification and user profiling. Concerning topic identification, there are four main efforts in the literature: (i) clustering, which includes traditional data mining algorithms applied to textual data [Aggarwal 2012]; (ii) natural-language processing (NLP) methods, well-known as the most effective for semantic analysis in several scenarios [Mihalcea and Radev 2011] but also the ones that require most effort to define properly the semantic representation in each domain; (iii) probabilistic, such as Latent Dirichlet Allocation (LDA), which employ a similar interpretation to data as linear algebra methods and (iv) non-probabilistic, which are the methods we are interested in.

Non-probabilistic methods, such as matrix factorization and sparse coding [Bai et al. 2013, Cheng et al. 2013], assume that there are few latent factors not directly observable from data that are able to represent most of the original data. In this case, each latent factor is defined as a semantic topic. As one may refer to the same semantic topic using different vocabularies, we say this type of technique actually generates fragmented or even redundant representations of semantic topics, known as semantic sub-topics.

In order to deal with the problem of redundancy, Kuhn et al. [Kuhn et al. 2007] uses a Latent Semantic Indexing (LSI) method followed by a clustering process to identify semantic topics from the source codes of a system. Their method identifies latent factors in the raw data, and then represents the original files using the new N-dimensional space defined by the N latent factors. Next, it clusters the files using a co-variance matrix represented on the top of this new space of files. The main drawback of this approach is the use of co-variance matrices, which do not scale to large volumes of data.

For short texts such as tweets, matrix factorization methods also have problems dealing with highly sparse data, a problem attacked in [Cheng et al. 2013]. Here we minimize this problem by using a set of tweets instead of a single one.

Regarding user profiling methods, they can build user profiles from different user characteristics, which include user preferences (found in messages), personal information (such as gender and location) or social network behavior (usage time, number of posts per day, among others) [Tao et al. 2012, Pennacchiotti and Popescu 2011]. Previous works have already used the text from tweets to help classifying users, and even applied linear algebra methods, such as LDA, in order to find topics that are not explicitly mentioned in the text. Pennacchiotti et al. [Pennacchiotti and Popescu 2011], for instance, proposes an architecture that builds user profiles in order to classify Twitter users with regards to particular political views (e.g., Democrats vs. Republicans). The authors use features such as the linguistic content of the tweets posted by a specific user and his/her social behavior. The information extracted is given to a machine learning algorithm that classifies users as republicans or democrats. A description of interests related to each class is also provided by an LDA-based method. However, the main problem in this case is that the authors assume the existence of training data, which is not always available for semantic analysis.

Tao et al. [Tao et al. 2012] and Abel et al. [Abel et al. 2011b], in contrast, use tweets to infer user interest profiles. Here, instead of identifying general topics, the authors first enrich the tweets by identifying and extracting a set of 39 pre-defined entities using OpenCalais[1], and also consider hashtags. [Tao et al. 2012] also uses a set of 19 pre-defined topics to identify the users preferences. The identified preferences are weighted and shown with clould tags and graphs. In [Abel et al. 2011b], in turn, as the final goal of the authors is to use the profiles to improve news recommendation, the authors also identify links in tweets and associate them with the original webpages, which in most cases refer to news. Topics are then extracted from these identified news also using the OpenCalais taxonomy. The authors compared profiles built with different combinations of the information described above, and concluded that entity-based and topic-based profiles make the semantics of the preferences much more explicit than other types of information.

From the methods mentioned above, the works of [Kuhn et al. 2007] and

[1] www.opencalais.com

[Pennacchiotti and Popescu 2011] are the most similar to ours. However, in contrast with [Kuhn et al. 2007], UPsCAle represents a scalable and robust strategy for modeling semantic topics. It is also independent of training data, and does not require a set of predefined topics to work with.

3. UPsCAle : A Framework for Creating Users Profiles

This section describes UPsCAle , the framework proposed to identify semantic profiles for a set of targeted users from Twitter. As previously described, the users profiles are a set of weighted preferences, representing a semantic topic frequently discussed in their posts D. Figure 1 illustrates the framework. Steps one to four, i.e.,text preprocessing, data modelling and extraction of latent factors, do not differ from what has been done in the literature so far. The main contribution of the framework is steps five to nine, where semantic sub-topics are merged into semantic groups and then used to describe user preferences. Finally, these user preferences are used to generate users semantic profiles.

The text preprocessing phase follows the traditional steps of text preprocessing in information retrieval, which includes the removal of special characters, stop-words, plural and genre markers, and the conversion of verbs to infinitive. After preprocessing, the data modelling phase creates a matrix, given as input to a matrix factorization method.

The data has three dimensions to be considered as part of the matrix: users, terms and posts. Our final goal is to identify the topics most discussed by the users in a given domain. Hence, one intuitive representation would be a matrix of *users* x *terms*, once topics appear from correlations between terms, and we want to associate topics with users.

However, the set of terms that represent a user will probably refer to more than one topic. As a post is usually made of one sentence expressing the user's opinion or preference in a specific matter, sets of posts will represent sets of topics. For this reason, data is represented as a matrix of *posts* $\times$ *terms*, making the correlation between different topics simpler to identify. The term associated with each position in the matrix is represented using the term frequency (TF), which showed better results than a TF X IDF representation.

After these initial steps, the resulting matrix is used to identify latent factors or semantic sub-topics (representing relations between different terms) using a Non-negative Matrix Factorization (NMF) methods, which generates semantic topics that will be associated with user preferences. From the set of user preferences, users profiles are generated. These two steps are detailed in the next sections.

3.1. Identifying latent factors

As discussed in Section 2, many works in the literature have dealt with the problem of identifying latent factors. Here we work with NMF because it does not assume that latent factors compose a space of independent variables. Further, NMF provides a more intuitive modelling of documents through these factors, defining each document as a sum of positive components. We can briefly describe NMF as follows. Given an input matrix $A \in \mathbb{R}^{m \times n}$, where each line represents a post and each column represents a term, and an integer $k < min\{m, n\}$, representing the number of desired latent factors, NMF finds two non-negative matrices $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{n \times k}$, where $A \approx WH^T$.

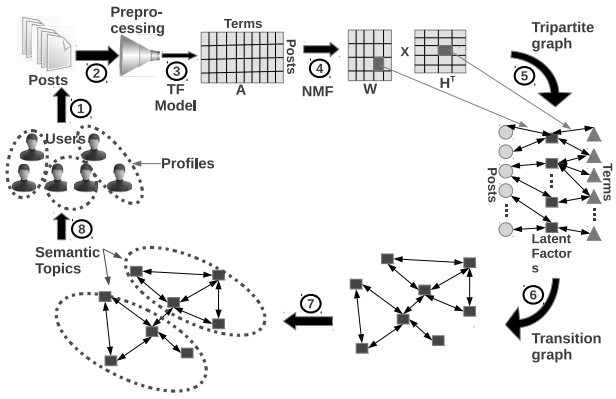


Figure 1. UPsCAle : a framework for user semantic profile identification

Defining k for NMF is not simple, since it is widely believed that NMF is a non-convex problem with a unique solution, and only local minima can be found [Lin 2007]. Here we propose a heuristic for finding k based on the variability analysis of the post collection, as the objective of this phase is to identify as many distinct relevant topics as possible.

We assume that the dimensions of highest variability define the most representative topics on each collection. Thus, by knowing the number k of topics necessary to provide x% variability is enough to identify the number of most representative topics. In this sense, we run a Principal Component Analysis (PCA) method to estimate k. The rationale behind this choice is that the PCA is well-known for returning an ordered set of linearly independent components that represent the dimensions of most variability on the data. According to the theory of Principal Component Analysis (PCA) [Johnson and Wichern 2002], the total population variance (tpv) can be described as $\sum_{i=1}^{k} \lambda_i$, where λ_i is the ith eigenvector of the co-variance matrix associated with the input matrix of posts $\times$ terms A [Johnson and Wichern 2002].

The number k of eigenvectors is chosen based on the desired tpv, or equivalently to the percentage of desired representativeness. Hence, the framework replaces k by a more intuitive parameter, which is the x% most representative topics.

3.2. Finding Semantic Topics

We have already discussed that matrix factorization methods produce semantic sub-topics, but they are not able to guarantee that the latent factors found represent **distinct**, **representative** and **cohesive** semantic topics. These three properties are important to show the topics are as independent and disjunctive as possible. According to these properties, the semantic topic identification problem can be defined as: identify the minimum number of semantic topics that represent, with high intern cohesion and low fragmentation the most representative topics in a set of posts D. Note that we are not interested in identifying all topics, but the most representative subset.

The main objective of UPsCAle is to find topics with these three properties, which is a real challenge. Representativeness refers to the frequency of the topics in D. However, it cannot be directly measured once different vocabularies might refer to the same topic. These different vocabularies have a direct impact on cohesion. Cohesion is defined as the

capacity of a topic being associated with a single subject. Finally, the topics have to be the least fragmented as possible. This is necessary because topics are actually a hierarchy of subtopics. For instance, when we talk about sports, we can talk about football or tennis. In general, the existing techniques are able to define each subtopic as a distinct topic, which generates a fragmented set of topics about the same subject. This is one of the main differences from our method to the others: identify non-fragmented semantic topics.

In order to merge semantic sub-topics, we first assume latent factors are modelled by a stochastic process, and that there is a probability of reaching a factor f' when leaving a factor f. Hence, we first represent the users, posts and terms as a tripartite graph, and transform it into a transition graph between different latent factors. We then perform a random walk between latent factors on this graph, and merge factors with high probabilities of mutually reaching each other.

3.2.1. Creating a tripartite graph

The NMF output matrices describe the relations between the latent factors found and the input terms and posts. The output $m \times k$ matrix W represents the m input posts through k latent factors. Each value W_{ij} greater than zero defines both a directed edge from post p_i to latent factor f_j and an edge in the opposite direction from f_j to p_i. Similarly, matrix H of $n \times k$ dimensions describes each term through k factors, and each cell H_{ij} greater than zero defines a relation between a term and a latent factor.

Based on this relation, we build a weighted directed tripartite graph G_t with three types of nodes T, F and D, representing the terms, latent factors and posts, respectively. The intuition behind this graph is that a term can be associated with more than one latent factor (e.g., the term *depression* might be present in the topic about health and politics with different meanings) as a post can talk about one or more latent factors (e.g., the same post can talk about sports and food "Watching the football at Albanos while eating the best pesto spaghetti ever #goPSG"). Each edge weight represents the intensity of the relationship between two nodes, and the graph weights are given by the values of W and H. As the values of the output matrices of NMF are always positive, we used the normalized values of W_{ij} and H_{ij} as edge weights. Edges leaving a term or post towards a latent factor, W_{ij} or H_{ij}, are normalized by the sum of the values in the ith line of W or H, making the sum of all leaving probabilities equals to 1. Analogously, for edges leaving from latent factors to terms/posts, we normalize their values by the sum of the values of the jth column.

3.2.2. Building the Topic Transition Graph

From the tripartite graph G_t, we want to convert it into a new graph G describing only relationships between latent factors in F. As the edge weights in G are normalized, each weight can be interpreted as a probability of leaving one node and reaching directly a different one. Following this rationale, each latent factor $f_i \in F$ has also an indirect probability of reaching another factor $f_j \in F$ through terms and posts. We want to transform these indirect links into direct relations between factors. Having a graph that represents only latent factors, we can then calculate the probabilities of a random walker

Algorithm 3.1 Merging latent factors

```
function JOINTOPICS(k, M, α)
    while k > 1 do
        P_min = getMinTransitionProbability(M)
        M' = modifyTransitionMatrix(M)
        candidatePairList = getImpact(M', k, α)
        candidatePair ← getBestImpactPair(candidateList)
        while impact (candidatePair) > 0 do
            updateTransitionMatrix(M, candidatePair)
            k − −
            candidate ← getBestImpactPair(candidateList)
    return M
function GETIMPACT(M_t, k, α)
    for i = 1 → k do
        for j = i + 1 → k do
            M'_t = mergeTopics(M_t, i, j)
            Δ_Coh = cohesionDifference(M_t, M'_t)
            Δ_Uni = uniquenessDifference(M_t, M'_t)
            impact(i,j) = α × Δ_Coh + (1.0 - α) × Δ_Un
    return impact
```

leaving a latent factor f_i and reaching a latent factor f_j. In order to do that, we transform this two-step probabilistic paths into a single edge by using Equation 1. Note that we join these probabilities by summing up their results, but other kinds of combinations could be tested.

$$P(f_i \rightarrow f_j) = \sum_{k \in T^{f_i}} P(t_k|f_i) \times P(f_j|t_k) + \sum_{k \in D^{f_i}} P(d_k|f_i) \times P(f_j|d_k) \tag{1}$$

where T^{f_i} and D^{f_i} are the sets of terms and posts with an input edge f_i. In Equation 1, the first sum represents all indirect paths from f_i to f_j passing through the terms, while the second comprises the indirect paths through documents.

The resulting graph G is represented by a stochastic transition matrix M, i.e., where the sum of each row in M is equal to 1. As we are concerned with semantic relations among latent factors, and that not semantically correlated topics may share a subset of terms, it is important to distinguish effective relations from noisy ones. A noisy relation is defined as an edge in G_i with transition probability smaller than the random transition probability in a graph of the same size. Noisy relations are identified and removed from G by setting to zero their corresponding cells in matrix M. In order to ensure that M remains stochastic, we re-normalize each of its rows. Further, in order to ensure that the random walking process conducted in the next step converges to a unique solution, we make G irreducible (all of its nodes are mutually reachable) and aperiodic (there is no integer $S > 1$ that divides the length of every cycle of the graph).

3.2.3. Merging Topics

Having the topic transition graph, the next step of our method is to merge semantic sub-topics. The idea is to merge topic pairs with a high mutual probability of reaching each other while the gains regarding *uniqueness* surpass the losses regarding *cohesion*, as described in Algorithm 3.1. The algorithm receives as inputs the number of latent factors identified in Section 3.1 (represented by k), the transition matrix M generated in the previous step and the linear weight α that determines the relative relevance of *uniqueness* gains and *cohesion* losses, once they are contradictory goals, as explained below.

The algorithm is iterative, and at each iteration a minimum transition probability P_{min} is defined between all pairs of topics a, b in M as the mean probability of a random walker go from a to b (without using self-loops) times the probability of the random walker go from a to any topic chosen at random (line 3). Next, as the graph G represented by M is not usually a clique, line 4 calculates the probabilities of nodes not directly connected reach each other, generating M'. As M is irreducible, this information is obtained by exponentiating M d times (M^d), where d is the diameter of G.

Next, in line 5, the function *getImpact* calculates, for each pair of topics, the impact that merging them will produce in M' w.r.t. *fragmentation* and *cohesion*. $\Delta_{Cohesion}$ is given by the difference between the minimum transition probability among distinct topics in G and the random transition probability in a clique with N distinct topics over the random transition probability. The higher the minimum transition probabilities observed among topics in G when compared to a random walk in a clique, the more cohesive G is. $\Delta_{Unicity}$, in turn, is defined as the difference between the mutual transition probability between two distinct topics i and j selected to be merged, and the minimum transition probability among topics G over the minimum transition probability. In this case, the higher the mutual transition probability, the higher the gains w.r.t. uniqueness in merging i and j.

Equations 2 and 3 define the $\Delta_{Cohesion}$ and $\Delta_{Unicity}$, respectively. $\overline{reach(G)}$ denotes the minimum transition probability among all pairs of distinct topics in G; N refers to the number of distinct topics of G and $reach_{j,i}(G)$ denotes the transition probability between topics i and j of G. Tests using the average and maximum transition probability for $\overline{reach(G)}$ were also tested, but the values of minimum transition probability presented the best results.

$$\Delta_{Cohesion} = \frac{\overline{reach(G)} - \frac{1}{N}}{\frac{1}{N}} \tag{2}$$

$$\Delta_{Unicity} = \frac{(\frac{reach_{,i,j}(G) + reach_{j,i}(G)}{2}) - \overline{reach(G)}}{\overline{reach(G)}} \tag{3}$$

The impact is defined by the linear combination of uniqueness gains and cohesion losses weighted by α. If the highest impact of a pair of topics caused in M is positive, the pair will be merged and M updated by replacing the two merged pairs with a single new topic and updating their the transition probabilities of their in- and out-edges. Next, a new topic with high impact in M is selected, and this process goes on until there are no remaining topics with probability greater then P_{min}, or the number of resulting topics

reaches one.

At the end of this process, the original latent factors are grouped. As users generate posts, the relations between users and semantic topics may be determined indirectly through the posts he/she created.

3.3. Creating User Profiles

This section shows how posts and users are mapped to semantic topics. Algorithm 3.1 outputs a matrix M of posts per semantic topics, where each cell shows the probability P_{ij} of a given post i belong to a topic j. Initially we set to 0 all the probabilities in M that are lower than the random probability of a post belong to a topic. From the modified M we extract the probability that a user k in U talks about a semantic topic given his/her posts according to Eq. 4. In Eq. 4 $P_{k,i}$ represents the uniform probability of a user posting i. Note that, in this process, a post can be associated with zero (in case all its probabilities of belonging to a topic are below random), one or more than one topic.

$$\sum_{i,j \in M, k \in U} P_{i,j} \times P_{k,i} \tag{4}$$

After this process, we end up with a new matrix M_u of users per topics (k x j), which has the probability of each user talking about a topic. Given this matrix, we can extract general preferences (topics) of the group of users or those more specific to subgroups of users. Here we are interested in the users general preferences, but the method can be easily adapted to work with specific ones by, for instances, clustering the users.

The general users preferences can be the generated, for example, by calculating the mean or median of the probabilities of all users talk about a topic. By ranking the users' preferences according to any of these simple metrics, we obtain the most popular preferences describing the set of users.

4. Experimental Results

The experimental evaluation carried out with UPsCAle was divided into two parts. First we evaluate the semantic topics extracted by the method proposed in Section 3.2 using two labelled datasets. Next we show a case study on a Twitter dataset that considers the followers of Obama and their messages posted during the presidential election campaign in 2012, assessing the full extension of UPsCALe. Table 1 shows the main characteristics of the three datasets considered. In order to simulated a scenario with restricted contextual information, as in social media posts, for AgNews we consider only the titles from each news in this collection to associate the documents with the 11 classes. The second labelled dataset is the Observatory, composed by a set of 6,000 tweets collected using keywords and manually classified as belonging to one out of six topics: religion, dengue fever, soccer, election, traffic, or cars.

4.1. Semantic Topic Identification

This section describes the topic identification evaluation followed by the results for topic identification in the two labelled datasets.

Table 1. Datasets used by UPsCAle to extract semantic topics

Dataset	Users	#Posts	#Topics	#Terms
AgNews	-	878,705	5	99,394
Observatory	-	5,431	6	1,956
USAElections	53,571	708,121	-	61,526

4.1.1. Evaluation Methodology

Evaluating the topics identified by UPsCAle can be as difficult as obtaining them. For this reason, this section defines a set of metrics that, given a known set of topics, evaluates the performance of the semantic topic identification phase. Consider we are interested in finding semantic topics that present high representativeness, high cohesion and low fragmentation (uniqueness). The representativeness of the topics is currently given as a parameter to the system, and used by PCA to define the number of latent factors we are looking for. In the experiments, different results of representativeness were tested and a subset of them is presented in this section. For cohesion and uniqueness, we defined the following metrics.

Cohesion, or the intra-clustering similarity, is measured using the purity of the semantic topics found [Witten and Frank 2005]. In order to calculate purity, each semantic topic is paired with the most frequent class in the documents of the topic, and then the accuracy is measured by counting the number of correctly assigned documents over the total number of documents, as showed in Eq. 5. T represents the sets of semantic topics (clusters) found, C is the set of real, known topics (or classes), and D the total number of tweets.

$$purity(T, C) = \frac{1}{D} \sum_{i=0}^{k} max_j |t_i \cap c_j|. \tag{5}$$

Fragmentation, in turn, measures the inter-clustering similarity using the MRVI (*Mean Relative Vocabulary Intersection*) [Kao et al. 2004] among pairs of topics T_i and T_j. The intuition behind this metric is that if T_j is a subset of T_i, then the vocabulary intersection between samples of T_j and T_i with the complete set T_i are similar. A global fragmentation analysis is done by plotting the distribution of MRVI among all pairs of topics $T_i, T_j \in T$. The higher the RVI the higher the probability of T_j being a fragment of T_i.

Apart from the two aforementioned metrics, we also calculate the precision and recall of the topics. The way precision and recall are calculated depend on how the documents are mapped to topics found and how topics are mapped to the known classes. Recall that a post can be associated with zero, one or more topics. Similarly, a topic can be mapped to one or more classes. Hence, when precision and recall are calculated, they consider only the documents that belong to at least one topic. For this reason, we refer to the precision and recall metrics as an "estimated" precision and recall, or e-precision and e-recall.

The evaluation measures showed here were calculated under four scenarios. The first, named 1-1_1-1, maps one document to one topic, and one topic to one class. This is the most restrictive metric, which associates the document with the most probable topic

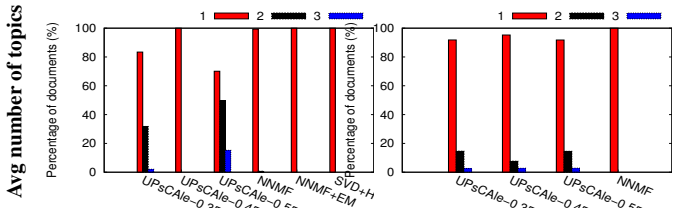


Figure 2. Topics most frequently assigned to documents for Observatory and agNews

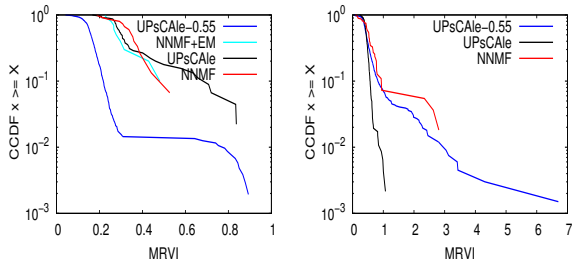
and the topic with the most frequent class among the documents in that topic. The second, 1-1_N-1, maps one document to one topic but n topics to one class. In this case, classes may refer to more than one topic. The third measure allows one document to be mapped to more than one topic, and again more than one topic may be associated with a single class (1-N_N-1). Finally, the less restrictive metric is 1-N_N-N, which maps one document to n topics, and these n topics with n classes, characterizing a multi-label class problem. For the first three cases, e-precision and e-recall are calculated using the tradition precision formula [Witten and Frank 2005]. For the fourth case, a multi-label version of these metrics is used [Tsoumakas and Katakis 2007].

In cases where the number of topics found is greater than the number of actual classes, depending on which type of topic-class mapping is being used, only the topics with the highest number of documents are mapped to the real classes, and this is reflected in the recall of the methods.

Considering these metrics, UPsCAle is compared with three other semantic topic identification techniques. The first, NMF, corresponds to the first part of UPsCAle and does not perform any merging of topics after finding the latent factors. The second applies a hierarchical average-linkage clustering algorithm to the topics extracted using Singular Value Decomposition (SVD), as done in [Kuhn et al. 2007]. A version of Matlab's hierarchical clustering algorithm was used. Finally, the third method applies the expectation-maximization algorithm to the topics created by NMF to merge topics. In this case, instead of using the PCA to find the number of latent factors needed for a given representation, we use a predefined number of latent factors defined by the formula given in [Kuhn et al. 2007]. A second version of this baseline, replacing NMF by SVD was also tested, and the results were statistically worst than those reported here. For all baselines, the mapping between topics and classes follows the same principles as for UPsCALe. NMF runs for 100 iterations in all cases.

4.1.2. Experimental Results

For the two labelled datasets, AgNews and Observatory, we report a subset of tests with different levels of topics representativeness (35%, 45% and 55%) and α (0.3 and 0.5). As the topic-document mapping plays an important role during evaluation, Figure 2 shows the number of documents to which 1, 2 and 3 topics were assigned. The analysis shows that no more than 3 topics were assigned to any document in both datasets. As data representativeness increases and reaches 55% the number of documents assigned to more than two topics increases, but all baselines considered are assigned only one topic per



(a) Observatory (b) AgNews

Figure 3. Distribution of $MRVI$ among all pairs of topics

document. Note that the values for SVD-HC and NMF+EM are not shown for agNews because they present a prohibitive computational time.

Table 2 shows the average results obtained over 10 runs of the methods. Results are compared using a two-tailed t-test with significance level 0.95. The second column shows the method considered: UPsCAle, SVD+HC, NMF+EM and NMF. For the latter, the number of latent factors to be found (k) is set as the number of known topics. In the case of UPsCAle, it is followed by the data representativity required (given as a parameter to PCA, which returns k). When the value of representativeness is missing, k is set to $(m \times n)^2$ [Kuhn et al. 2007]. The next two columns, under the name # Topics, show the number of topics returned by PCA (k) followed by the final number of topics, obtained after running the merging phase of UPsCAle (k'). Note that, for AgNews, although the initial number of given topics was 12, this value decreased to 5 after the document-topic mapping, as documents belonging to 7 different small classes did not belong to any topics. The next column, Docs with Topics, shows the fraction of documents which were assigned to at least one topic. We emphasize that the metrics of e-precision and e-recall for the four situation previously described provided in the next columns were calculated over these documents.

In general, we observe that the higher the representativeness, the higher the number of covered examples, as more topics cover more documents. Similarly, the higher the value of cohesion, the higher the number of final topics k', as the topics become more fragmented to be more cohesive. The differences between k and k' show that the merging process proposed significantly reduces the number of topics found. In average, the proposed method performed 19 iterations for Observatory and 29 to AgNews.

Now looking at the Observatory dataset. The number of documents assigned to at least one topic varied from 88 to 99% according to different representativeness. When the number of topics was previously set, this number decreased to 40%. In order to analyse precision and recall, we select one of the results and compare it against the others using the t-test. The results compared are highlighted in gray, and the symbol ▲ in a cell indicates its result is better than the highlighted, ▼ indicates it is worse and ● that there is no statistical evidence to state any difference.

Analysing the precision and recall, UPsCAle with $\alpha = 0.5$ and 0.55 representativeness was chosen as a reference, as it will cover more documents and set an equal trade-off for cohesion and uniqueness. Comparing these results with the other using $\alpha = 0.3$,

Table 2. Results obtained for Observatory and AgNews. Documents with topics represent the number of documents assigned to at least one topic. The metrics of precision and recall are calculated under these documents. The four cenarious presented considering different mapping of Doc_Topic_Topic_Class.

Cohesion (α)	Method	#Topics		Docs with Topics	Purity	1-1_1-1		1-1_N-1		1-N_N-1		1-N_N-N	
		k	k'			e-prec.	e-recall	e-prec.	e-recall	e-prec.	e-recall	e-prec.	e-recall
Observatory													
0.30	UPsCAle-0.35	115	15.6	▼0.883	▼0.641	▼0.574	▲0.380	●0.477	▲0.435	●0.437	▲0.466	●0.789	▲0.884
	UPsCAle-0.45	171	23.3	▼0.968	▼0.678	▼0.579	▲0.280	●0.461	▲0.358	●0.422	▲0.395	▼0.757	▼0.863
	UPsCAle-0.55	242	31.3	●0.995	▼0.685	▼0.554	▲0.217	●0.451	▲0.304	●0.406	▲0.342	▼0.753	●0.873
	UPsCAle	25	5.9	▼0.391	▲0.825	▼0.520	▲0.429	●0.494	▲0.450	▲0.476	▲0.455	▲0.852	●0.868
0.50	UPsCAle-0.35	115	25.1	▼0.885	▼0.696	▼0.582	▲0.283	▲0.473	▲0.364	●0.434	▲0.390	●0.769	▼0.855
	UPsCAle-0.45	171	32.9	▼0.963	▼0.698	●0.631	▲0.244	●0.461	▲0.326	●0.413	▲0.363	▼0.761	▼0.864
	UPsCAle-0.55	242	43.9	0.995	0.745	0.657	0.175	0.438	0.247	0.404	0.271	0.773	0.873
	UPsCAle	25	9.7	▼0.395	●0.747	●0.631	▲0.442	▲0.566	▲0.508	▲0.552	▲0.516	▲0.859	●0.883
-	SVD+HC	-	2,333	▲0.999	▼0.354	▲0.898	▼0.024	▲0.776	▲0.775	▲0.776	▲0.775	▲0.956	▲0.956
	NMF+EM	25	12.7	▼0.925	▼0.540	▼0.435	▲0.289	●0.390	▲0.415	●0.390	▲0.415	▲0.812	▼0.812
	NMF	6	6.0	▼0.111	▲0.780	●0.670	▲0.629	▲0.671	▲0.699	▲0.669	▲0.700	▲0.921	▲0.923
AgNews													
0.30	UPsCAle-0.35	59	13.9	▼0.011	▼0.559	▼0.181	▲0.191	▼0.200	▲0.204	●0.200	▲0.204	●0.763	▲0.836
	UPsCAle-0.45	103	19.4	▼0.085	▼0.694	▼0.242	▲0.166	●0.206	▲0.181	●0.207	▲0.186	▲0.781	▲0.806
	UPsCAle-0.55	163	31.9	●0.359	●0.803	●0.246	●0.125	●0.212	●0.137	●0.197	●0.140	●0.762	●0.797
	UPsCAle	154	29.5	●0.283	●0.796	▼0.242	▲0.143	●0.212	●0.154	●0.189	●0.154	▲0.765	▲0.796
0.50	UPsCAle-0.35	59	16.5	▼0.011	▼0.557	▼0.229	▲0.219	●0.227	▲0.238	●0.231	▲0.239	▲0.784	▲0.855
	UPsCAle-0.45	103	24.1	▼0.090	▼0.693	●0.309	▲0.183	●0.256	▲0.202	●0.248	▲0.207	▲0.776	▲0.802
	UPsCAle-0.55	163	34.4	0.360	0.810	0.292	0.117	0.252	0.132	0.235	0.137	0.751	0.783
	UPsCAle	154	31.9	●0.283	▼0.787	●0.304	▲0.139	●0.245	▲0.157	●0.228	▲0.165	▲0.770	▲0.799
-	NMF	-	11	-	-	-	-	-	-	-	-	-	-

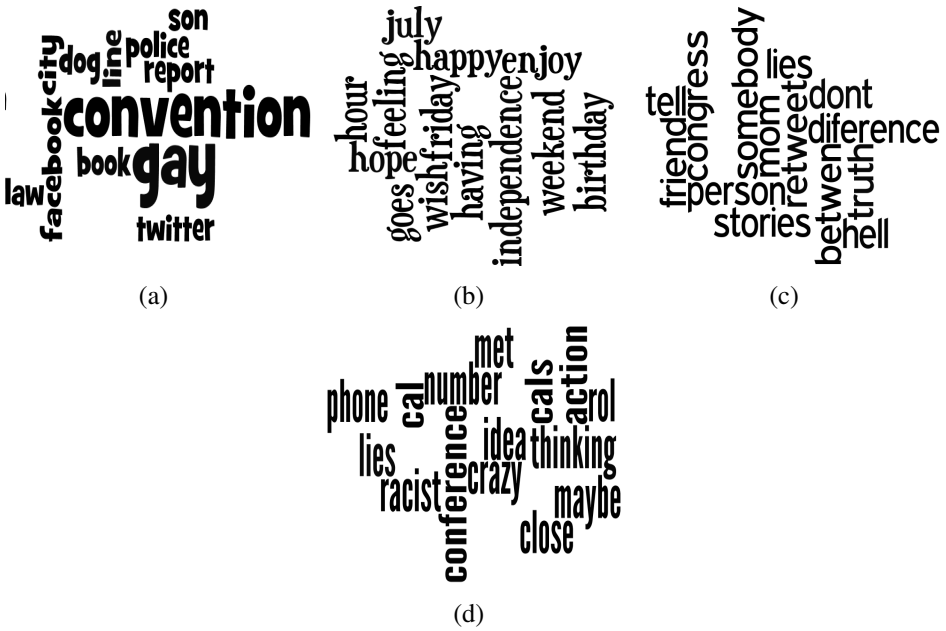


Figure 4. Main topics discussed by Obama's followers: Independence Day, Democrats Convention in Sep. 2012 and gay marriage, Congress, and Call for an Ant-Racist Action.

for all e-precisions except the most restrictive, the results found are better than the reference. Looking at the baselines, the results of SVD+HC seem far better than those obtained by UPsCAle. However, note that SVD+HC found 2,333 clusters. In the evaluation process, as only six clusters were mapped to the real classes in the 1-1_1-1 evaluation, the recall is around 2%, explaining the high precision. NMF also presented a much higher e-precision and e-recall than UPsCAle, but the column "Docs with Topics" shows that only 11% of documents were assigned to at least one topic in this case. NMF+EM is the strongest competitor, but given its final number of topics, the fairest comparison is to compare it with UPsaAle-0.35 with $\alpha = 0.3$. The latter covers 4% less documents, but presents higher purity and both e-precision and recall for all mapping, except for the less restrictive precision. Hence, tuning α and the representativeness, we can obtain better results than the current methods more efficiently, as discussed below. The much lower purity of NMF+EM might be due to the final number of topics.

With respect to agNews, we observe a completely different situation. The cost of the baselines was prohibitive, and when 11 was given to NMF as the number of latent factors to be found (i.e. the number of known-topics), only 57 documents (out of 878,705) were assigned to topics. The values of assignment for UPsCAle are also low, reaching 36% with the highest value of representativeness and 34 topics in average.

The last metric calculated over the labelled datasets, fragmentation, is showed in Figure 2. The histograms for the two datasets are completely different (recall that the lower the MRVI, the better). For Observatory, the graph shows that UPsCAle 0.55 is superior, i.e., the topics found overlap less. For agNews, UPsCAle using the predefined number of topics is superior, but this is due to the size of the topics (clusters) generated. As they are smaller, they tend to overlap less, and hence MRVI is better.

Table 3. Results obtained by UPsCAle considering Obama's followers in Twitter

Obama's Followers				
	Repres.	#Topics		Purity
		k	k'	
α= 0.5	0.35	94	40.5	0.608
	0.45	141	43.75	0.613
	0.55	204	57.67	0.629

From the results previously discussed, we can conclude that UPsCAle presents solutions statistically better than those obtained by the baselines when the value of α is tuned accordingly. The semantics of the topics is also good, but results are omitted here due to space restrictions, and presented in the case study in the next section. More importantly, the proposed framework was conceived to scale to bigger datasets. This is its main advantage over the baselines that aim to reduce the number of topics. While EM is linear in the number of documents N, the proposed algorithm depends on the number of latent factors k used to describe N and in the diameter d of the graph. Algorithm 1 has time complexity O($(k-1) \times d \times k^3$), where k-1 is the maximum number of iterations.

4.2. Obama's Followers: A Case Study

This section presents a case study for characterizing a specific group of Twitter users: Obama's followers. The dataset was collected from July to September 2012.

Recall that, for this dataset, we do not know the discussed topics. Hence, we evaluated the quality of the generated clusters by using an adaptation of the purity metric defined in Eq. 5. Instead of considering the purity of the clusters with respect to the real class, we evaluate the purity of the terms across different groups, as done in [Kao et al. 2004] with entropy. The results are shown in Table 3, and for all three data representativeness used, the results of purity are very similar.

Here we perform a more detailed analysis of how topics are merged and how these merges improve user characterization with semantic topics. For this dataset, 33 iterations of the method were performed. Figure 4 shows four of the top-ten topics most that characterize 80% of the users. The first topics, showed in Figure 4 (a), characterizes 70.5% of the users. The topic related to the selection criteria of the users, which was indirectly politics. Its main subject is the Democrats convention in September 2012, which decided to support gay marriage. Note the terms *son* and *dog* are not directly related to the topic, and were added during the merge of almost half of NMF output files. Analysing the topics in the earlier levels of the merge, we observe that all merges relate to politics. However, it is interesting to notice that the method actually produces a hierarchy of topics, which will generate single topics if the maximum number of iterations is performed. Correctly choosing where to stop in the hierarchy can give the user the topics at the semantic level he wants. Fine-tuning the stopping criteria is the next step.

The second topic which characterizes users (Figure 4 (b)) is related to the 4th of July. It actually shows terms referring to the independence and others referring to wishes of happy birthday or references to a happy and enjoyable weekend. Here UPsCAle merged 9 different topics related to best wishes in various events. The third topic is a merge of two topics referring to the congress and its bills. The fourth topic relates to calls for an anti-racist act. Note that the term *phone*, related to call, was associated to this topic in the NMF output.

5. Conclusions and Future Work

In this paper we analysed how the semantic topics can be used to extract general characteristics from the users. We propose UPsCAle, a framework that works in three phases: it identifies semantic sub-topics of a user group using a traditional matrix factorization method and then merges those sub-topics into more cohesive and unique semantic topics. In a last phase, it maps the final set of semantic topics to users profiles.

The proposed method was evaluated considering two labelled dataset and an unlabelled one. The results show that, when compared to other state of the art methods for topic identification, UPsCAle can find better results than the baselines if the the trade-off between cohesion and uniqueness is properly set. Furthermore, the method scales well to very large datasets.

As future works, the stopping criteria of UPsCAle will be carefully studied. As it generates a hierarchy of topics, identifying the level of interest to be explore is important. The method will also be enhanced with the approach proposed in [Cheng et al. 2013] to deal better with data sparsity. Finally, other user characteristics will be added to the current profile.

6. Origins and General Discussion

The previous sections are part of an article sent to the SIAM International Conference on Data Mining 2014. It was a collaborative work with two doctorate students, namely Tiago Cunha and Fernando Mourão. I have worked mostly in the first 7 steps of the framework (see figure 1), implementing and discussing ideas and hypothesis with the team.

The objectives of POC 1 were successfully achieved, and I was able to exercise the acquired knowledge throughout my undergraduate course. During the development of the framework, new problems and ideas for future work were found. The first, and most important, is that the matrix of $posts \times words$ is very sparse and this can be a problem for the NMF algorithm. Dealing with sparsity is the first goal for POC 2. We will investigate a similar approaches to the one proposed in [Cheng et al. 2013].

The second objective of POC 2 is to make UPsCAle an online method. Currently, it is considered an offline method since it uses static input data, e.g. data do not vary over time. As we want to integrate it with *Observatório da Web* we need to learn how to deal with streams of tweets and news.

References

Abel, F., Gao, Q., Houben, G., and Tao, K. (2011a). Analyzing user modeling on twitter for personalized news recommendations. In *UMAP'11*, pages 1–12.

Abel, F., Gao, Q., Houben, G., and Tao, K. (2011b). Semantic enrichment of twitter posts for user profile construction on the social web. In *The Semanic Web: Research and Applications*, pages 375–389. Springer.

Aggarwal, C. C. (2012). *Mining text data*. Springer Science+ Business Media.

Bai, L., Guo, J., Lan, Y., and Cheng, X. (2013). Group sparse topical coding: from code to topic. In *ACM WSDM*, pages 315–324.

Berry, M., Browne, M., Langville, A., Pauca, V., and Plemmons, R. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173.

Cheng, X., Guo, J., Liu, S., Wang, Y., and Yan, X. (2013). Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *SDM*, pages 749–757.

Johnson, R. A. and Wichern, D. W. (2002). *Applied multivariate statistical analysis*, volume 5. Prentice Hall.

Kao, H.-Y., Lin, S.-H., Ho, J.-M., and Chen, M.-S. (2004). Mining web informative structures and contents based on entropy analysis. *Knowledge and Data Engineering, IEEE Transactions on*, 16(1):41–55.

Kuhn, A., Ducasse, S., and Gírba, T. (2007). Semantic clustering: Identifying topics in source code. *Information and Software Technology*, 49(3):230–243.

Lin, C. (2007). On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *Neural Networks, IEEE Transactions on*, 18(6):1589–1596.

Mihalcea, R. and Radev, D. (2011). *Graph-based natural language processing and information retrieval*. Cambridge University Press.

Pennacchiotti, M. and Popescu, A.-M. (2011). Democrats, republicans and starbucks afficionados: user classification in twitter. In *ACM SIGKDD*, pages 430–438.

Pons-Porrata, A., Berlanga-Llavori, R., and Ruiz-Shulcloper, J. (2007). Topic discovery based on text mining techniques. *Information processing & management*, 43(3):752–768.

Tao, K., Abel, F., Gao, Q., and Houben, G. (2012). TUMS: twitter-based user modeling service. In *Proc. of the 8th Int. Conf on The Semantic Web*, pages 269–283.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2nd edition.

Zheleva, E. and Getoor, L. (2009). To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *WWW'09*, pages 531–540.