Vocabulary-based Community Detection and Characterization

Giorgia Ramponi
Politecnico di Milano
Milan, Italy
giorgia.ramponi@polimi.it

Marco Brambilla
Politecnico di Milano
Milan, Italy
marco.brambilla@polimi.it

Stefano Ceri
Politecnico di Milano
Milan, Italy
stefano.ceri@polimi.it

Florian Daniel
Politecnico di Milano
Milan, Italy
florian.daniel@polimi.it

Marco Di Giovanni
Politecnico di Milano
Milan, Italy
marco.digiovanni@polimi.it

ABSTRACT

With the increase of digital interaction, social networks are becoming an essential ingredient of our life, by progressively becoming the dominant media, e.g. in influencing political choices. Interaction within social networks tends to take place within communities, sets of social accounts which share friendships, ideas, interests and passions; detecting digital communities is of increasing relevance, from a social and economical point of view.

In this paper, we argue that the vocabulary of terms used in social interaction is a very distinctive feature of a community, hence it can be effectively used for community detection. We show that, by inspecting the vocabulary used by tweets, we can achieve very efficient classifiers and predictors of account membership within a given community. We describe the syntactic and semantic features that best constitute a vocabulary, then we provide their comparative evaluation and select the best features for the task, and finally we illustrate several applications of our approach to concrete community detection scenarios.

KEYWORDS

Social analytics, community detection, content-based data analytics

ACM Reference Format:
Giorgia Ramponi, Marco Brambilla, Stefano Ceri, Florian Daniel, and Marco Di Giovanni. 2019. Vocabulary-based Community Detection and Characterization. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), April 8–12, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3297280.3297384

1 INTRODUCTION

Defining the essence of a community is difficult: in the English dictionary, a community is the *condition of having certain attitudes and interest in common.* The concept of community is general and goes beyond social networks and Internet, but finding communities in the digital world is very relevant, as it has a huge number of social implications and potential commercial exploitations [11, 13, 17]. Digital social content can be automatically inspected, hence, social communities on Internet can be detected by algorithms [16, 17, 20]; this process comes with very interesting challenges from a social analysis perspective, as well as interesting computational problems.

Social networks can be considered as big graphs of linked nodes; most methods for community detection use as initial input the arcs among actors [8] (e.g. the *friendship/follow* relationships), or take into account social activities [20] (e.g., the *likes* or *comments*). These methods build weighted graphs representing social interactions and then look for subgraphs with certain properties (e.g., the sparsity/density of subgraphs), typically corresponding to subsets of highly interacting users.

In this paper, we explore a different direction, and propose a **content-based approach to community detection**. We conjecture that a community can be characterized by its own *vocabulary*, as it is a very strong distinctive property. With this approach, we define simple methods for community detection: given a set of social actors, we argue that they form a community if their vocabulary has strong similarity properties; we can also test if a social actor is a member of a community by comparing the actor's vocabulary to the community's vocabulary. As we will see, content-based analysis can be performed bottom-up, with very few actors forming an initial community, and thus it is less computationally demanding than link-based analysis.

This work is part of a general effort towards the use of social accounts for extracting semantic knowledge; in particular, in [5] we defined a method for extracting emerging knowledge from social accounts based on co-occurrence of accounts with known members of a community; in [4] we observed that very few accounts are sufficient to generate a community and we explored how such community grows in space and time as effect of iterative applications of the method. In this work, we concentrate on characterizing the distinctive features of a community; we demonstrate that the vocabulary of terms used by the community yields to an effective characterization of the community cohesiveness.

To better define our approach, we consider Twitter as social network and we study the communities of Twitter accounts; with this method, every Twitter account is associated with several tweets, and we consider the vocabulary of terms used in their tweets We then define the following problems: (a) Given a community of n twitter accounts, define the *strength* of the community, measuring how the community is well characterized by the shared vocabulary

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04…$15.00
https://doi.org/10.1145/3297280.3297384

of its members. (b) Given other accounts, define *membership criteria* for deciding if they are also part of the community. Solving these problems requires addressing two challenges.

- The first challenge is vocabulary characterization. As Twitter typically uses short sentences and has its own given jargon, we must choose among syntactic or semantic elements of the *Twitter jargon*.
- The second challenge is measuring the distance between vocabularies associated to accounts, so that we can test community's strength and membership.

We will consider a variety of options for both challenges, but we will eventually see that simple choices work remarkably well in practical contexts, suggesting that this approach has a wide applicability.

The paper is organized as follows. In Section 2, we characterize a community of Twitter accounts in terms of candidate features to form a vocabulary and of candidate distances between accounts. In Section 3, we select the most effective features for testing community's strength and membership. In Section 4, we assess the power of vocabularies in two important applications, related to the political arena and to targeted advertising; we also show how our previous work on semantic knowledge extraction integrates with our current work. Section 5 presents related work, and Section 6 concludes the paper.

2 CONTENT-BASED COMMUNITY CHARACTERIZATION

The method consists in using textual features provided by tweets to define a feature vector for every member of the community, then to compute the centroid of the feature vectors; once we have the centroid, we can use the distances to the centroid for assessing the community's cohesiveness and for measuring the participation of each candidate to the community.

2.1 Definitions

We introduce some definitions that are useful for defining the community detection method.

- *Community*: a community is a set of Twitter accounts that have some characteristics in common;
- *Member*: a Twitter account of the community;
- *Candidate*: a Twitter account that could be included in the community.

We formalize the community detection problem as follows:

Given a community, defined as a set of members $C =< m_1, ..., m_n >$, define a distance δ and a threshold t such that given a candidate c_i if $\delta(c_i, C) < t$ then $c_i \in C$ otherwise $c_i \notin C$.

2.2 Features

A tweet is a public message of at most 280 characters, shared by each Twitter account with all other Twitter accounts. Tweets are composed of words and hashtags (next to hyperlinks and images, which we do not further analyze in our work). We extract from tweets either syntactic or semantic features.

2.2.1 Semantic Features. The meaning of each word in a language is formed of a set of abstract characteristics known as semantic features. Every language is associated with a hierarchical structure representing semantic features, typically words are at the leafs of these hierarchies and semantics is assigned by traversing the hierarchy. When we consider semantic features, we go beyond the word itself, by extracting its meaning. In our work we use DBPedia as concept hierarchy; DBpedia extracts structured content from the information created in Wikipedia [1]. We then consider two features. To extract instances and types from tweets we use Dandelion[1], a commercial software which matches a text to either instances or types of DBpedia. We then consider two features:

- *type*: a *type* is an element of the DBpedia hierarchy; a word in a text is mapped to a type in DBpedia.
- *instance*: some words are also associated to a concept that has a page in Wikipedia; we call these concepts *instances*.

2.2.2 Syntactic Features. Words appearing in the tweets can also be classified on the basis of their syntactic features, by dividing words into verbs and nouns. We used the NLTK library to extract syntactic features from tweets: with NLTK we delete stop-words, tokenize, tag text, and retrieve the root form of the words[2]. We also considered as distinguished feature the proper nouns (NNP), a subset of nouns.

2.3 Centroid

We associate to every candidate accounts c a *feature vector* $f_c :< f_{c,1}, f_{c,2}, .., f_{c,n} >$, whose elements are the frequency the textual feature f that we extract from tweets of c. So if for example we are considering nouns, $f_{c,i}$ is the frequency of use of the noun i in c's tweets. From m feature vectors $\{f_1, ...f_m\}$ of cardinality n, we define the centroid:

$$z =< z_1, .., z_n >$$

where:

$$z_i = \frac{\sum_{j \in m} f_{j,i}}{m}$$

2.4 Distances

To evaluate the closeness of a candidate c to the centroid z we consider five distances:

- *Manhattan distance* ($l1$) between two vectors. It is the sum of the lengths of the projections of the line segment between the points into the coordinate axes. More formally:

$$l1(f_c, f_s) = \sum_{i \in [1,n]} \|f_{c,i} - f_{s,i}\|$$

- *Euclidean distance* ($l2$) between two vectors f_c, f_s is the length of the path connecting them, formally:

$$l2(f_c, f_s) = \sqrt{\sum_{i \in [1,n]} \|f_{c,i} - f_{s,i}\|^2}$$

- *cosine distance* (cd) between two vectors is a measure that calculates the cosine of the angle between them. Formally the cosine distance between two vectors f_c and f_s is defined as:

$$cd(f_c, f_s) = 1 - \frac{f_{c,i} \cdot f_{s,i}}{\|f_{s,i}\|_2 \|f_{c,i}\|_2}$$

[1] *https://dandelion.eu*
[2] *http://www.nltk.org*

- *Kullback-Leibler Divergence* (KLD), also called relative entropy, is a measure of how one probability distribution diverges from a second. So if we consider two vectors f_s, f_c as two probability distributions (and we can do it because we normalize the frequencies) the KL-divergence between f_c, f_s is defined as:

$$KLD(f_c, f_s) = \sum_{i \in [1,n]} f_{c,i} \log \frac{f_{c,i}}{f_{s,i}}$$

2.5 Dispersion Index

It measures the cohesion of a community. We consider the ratio D_c/D_T, where:

- D_c is the average distance of the members of the community to the community centroid, that should be small;
- D_T is the average distance of the members of the community to the centroid of the vocabulary used by all Twitter accounts, that should be big.

We expect a dispersion index between 0 and 1, where a smaller dispersion index is associated to communities with stronger cohesion.

2.6 Problem Formulation

We formulate the problem of *finding the best set of features and the most effective distance in order to characterize community membership, by using the distance from the community centroid.* More formally, given a community $C^* = c_1, ..., c_n$, we retrieve the tweets of these accounts and construct five feature vectors for every textual feature, relative to the syntactic and semantic features discussed above. From these feature vectors, five centroids $z_{type}, z_{instance}, z_{noun}, z_{verb}, z_{propernoun}$ are created. We next explore which combination of textual features and distances achieve the best result in predicting that a candidate account c_i is a member of the community and that the community is strongly or weekly characterized.

3 CONTENT-BASED COMMUNITY DETECTION

In this section, we comparatively evaluate the above features and distances in order to select the combination of them that better characterizes a community. We perform this task by solving a community detection problem which is artificially built by starting from known community members and separating them into two sets, one of which is merged with randomly selected accounts. We then use the alternative features and distances to measure their effectiveness in ranking the candidates, and compare the rankings. We show that the simple method based on selecting one type of feature has better performance compared to methods which combine several features together or that use latent semantic analysis for combining features. This result is particularly valuable, because it allows us to associate a community with a well-defined vocabulary, made up of few genuine terms.

3.1 Input Data

We consider three initial communities of twenty well-characterized professionals, each member of a specific domain as defined by domain experts, that constitute our gold standard. The communities are formed by chess players, fashion designers, and Australian writers. For every such community, we consider ten Twitter accounts as community members; we then consider a set of candidates constituted by the other ten members and by 160 random accounts. We repeated each extraction 50 times, and averaged the performance indexes.

3.2 Single Feature Types

For every choice of domain, feature and distance, we compute the centroid of the ten community members and we rank the candidates in terms of distance from the centroid. We consider *precision@10* and *recall@20* as relevant performance indicators; the goal is to retrieve the known ten members of the community within the top-ranked candidates.

Table 1 shows the results of our experiments. By comparing the four alternatives for distances, we note that KLD and cosine distance provide the best results in terms of precision and recall in all the domains, therefore we next focus on them. By then concentrating on the five syntactic and semantic features, we note that (syntactic) proper nouns and (semantic) instances also provide the best precision and recall in all domains.

Therefore, we consider the four combinations of KLD/cosine distances and proper nouns and instances features as the baseline of our method. Fig. 1 shows precision/recall diagrams for the four choices in the three application domains.

By comparing the domains, we note that precision and recall are generally higher for Chess Players, intermediate for Fashion Designers, and lower for Australian Writers. In particular, precision is extremely good for Chess Players, where all methods find the first 6 members as top ranked among all 210 candidates; and it is rather good for all domains, including Australian writers, as we find 4 members within the top ten ranked.

We inspected the twitter accounts, and we found that chess players tweet almost exclusively about chess, hence their vocabulary is narrower and most focused; fashion designers talk a lot about fashion but they also talk about several other close topics; and Australian writrers intertwine tweets about writing with tweets about many other topics, including personal experiences. This empirical consideration is quantified by using the **dispersion index** measuring the internal coherence of a community, defined in Section 3, whose values for the three communities are summarized in Table 2.

3.3 Combinations of Feature Types

We considered the combination of several features into mixed feature vectors (i.e., vectors combining several syntactic and semantic features); results were generally lower in terms of precision and recall. The row ALL represents the use of a normalized feature vector where each feature category is weighted 25% of the total weight, and precision - recall are generally lower. We tried other combinations of feature types, but they do not improve over the use of a single feature type.

3.4 Latent Semantic Analysis

We also considered **latent semantic analysis (LSA)**, the state-of-the-art technique used to analyze relationships between a set of documents and the terms they contain by producing a set of

Domain	Feature	$cd_{precision}$	cd_{recall}	$KLD_{precision}$	KLD_{recall}	$l1_{precision}$	$l1_{recall}$	$l2_{precision}$	$l2_{recall}$
Chess	NNP	**0.800**	**0.905**	**0.770**	**0.870**	0.800	0.885	0.140	0.270
	Noun	0.270	0.335	0.690	0.825	0.660	0.795	0.165	0.215
	Verb	0.155	0.235	0.130	0.330	0.200	0.350	0.135	0.200
	instances	**0.835**	**0.875**	**0.775**	**0.860**	0.750	0.810	0.320	0.385
	type	0.385	0.430	0.700	0.785	0.420	0.560	0.360	0.410
	all features	0.006	0.008	0.013	0.015	0.017	0.013	0.005	0.007
	LSA	0.200	0.510	0.240	0.470	0.200	0.510	0.130	0.340
Fashion	NNP	**0.510**	**0.695**	**0.560**	**0.745**	0.625	0.690	0.001	0.040
	Noun	0.180	0.345	0.485	0.610	0.710	0.770	0.075	0.150
	Verb	0.010	0.030	0.100	0.105	0.070	0.105	0.010	0.015
	instances	**0.695**	**0.765**	**0.595**	**0.765**	0.705	0.750	0.001	0.015
	type	0.120	0.250	0.165	0.195	0.235	0.315	0.125	0.240
	all features	0.006	0.007	0.012	0.013	0.011	0.011	0.005	0.005
	LSA	0.310	0.410	0.290	0.460	0.310	0.410	0.450	0.560
AW	NNP	**0.245**	**0.435**	**0.265**	**0.385**	0.310	0.450	0.030	0.030
	Noun	0.095	0.130	0.075	0.220	0.200	0.415	0.110	0.170
	Verb	0.120	0.190	0.005	0.155	0.085	0.190	0.115	0.165
	instances	**0.390**	**0.515**	**0.335**	**0.560**	0.245	0.415	0.075	0.115
	type	0.110	0.245	0.095	0.190	0.165	0.250	0.110	0.230
	all features	0.001	0.001	0.009	0.010	0.007	0.009	0.001	0.001
	LSA	0.040	0.070	0.040	0.060	0.040	0.070	0.050	0.110

Table 1: Exhaustive analysis showing the precision@10 and recall@20 for experiments built by combining in all possible ways four choices of distances and seven choices of features in three domains.

	AW	Fashion	Chess
NNP	0.84	0.79	0.55
instances	0.80	0.73	0.63

Table 2: Dispersion index for the three domains.

abstract concepts related to the documents and terms [7, 22]. With LSA, the input data is represented as a matrix in which each row corresponds to a word and each column corresponds to a document, and each matrix value contains the frequency of the word for the document; LSA consists in applying a singular value decomposition (SVD) to the matrix. We considered as documents the tweets of a specific account, and the words are all words which appear at least one time in one tweet.

Figure 2 compares the precision@10 of the various parametric runs of LSA. We used the SVD algorithm with 11 different parameters for the space size (from 10 to 100 with increment 10). In all application domains and for all the parameter settings, the precision of LSA is lower than the precision achieved just by proper nouns or instances.

These experiments were confirmed in many other domains (see also Section 4) and convinced us that a simple method, based on the use of proper nouns and cosine distance or KLD, is preferred to other choices; we prefer using proper nouns because they can be detected by a very efficient open source library, whereas extracting semantic instances requires a comparison with DBPedia, for which we currently use a computationally expensive commercial software.

4 APPLICATIONS

4.1 Content-based Analysis of Accounts from a Political Perspective

One of the most interesting applications of vocabulary-based community detection is concerned with political preferences. Politics is most influenced by the use of social media, as many politicians deliver their comments using Twitter. We therefore asked ourselves if the use of vocabulary could be suggestive of political preferences. At the March 2018 elections in Italy, three coalitions participated to the competition: the Right parties, Cinque Stelle, and the Democratic Party. We considered twenty elected politicians from the three coalitions, and we retrieved their tweets; we then performed the following experiments:

- We used as before a limited number of accounts as community members and we classified the remaining accounts on the basis of their similarity to the centroid; we repeated this experiment 50 times, every time selecting randomly the accounts to use as community members.
- We then repeated the test by using the followers. In this case, as we assume that the follower of a politician prefers the politician's party, we developed a predictor of the political

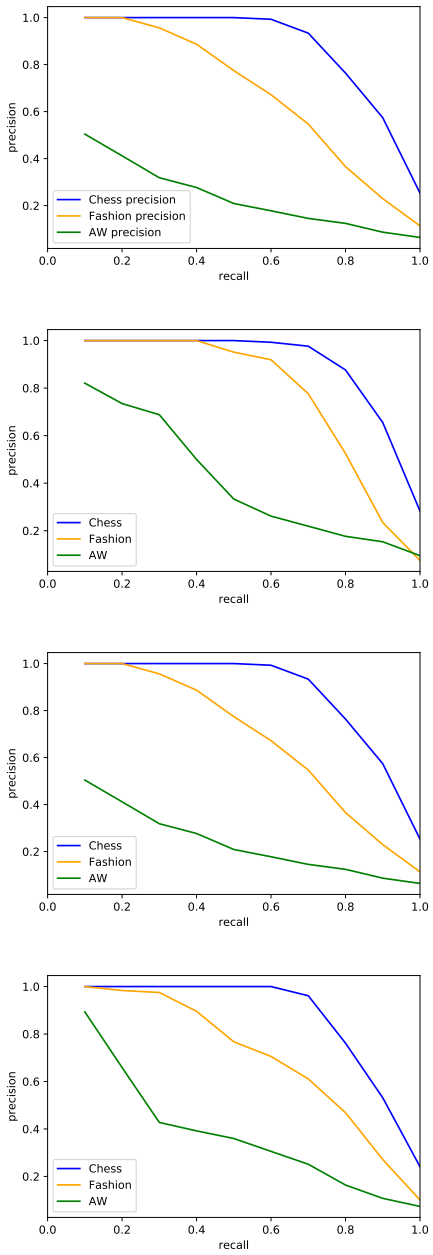

Figure 1: Precision and recall diagrams for the four combinations of selected features (NNP, instances) and distances (cosine, KDL) in the three applications.

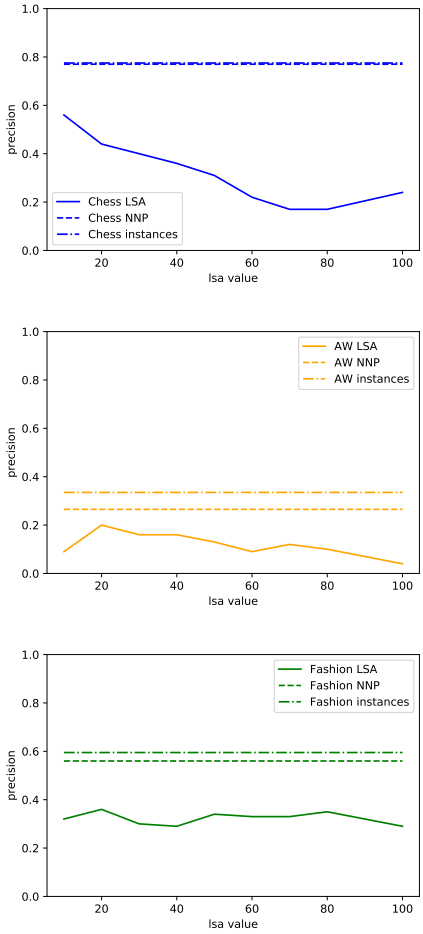

Figure 2: Precision@10 for different LDA parametrizations in comparison with precision@10 achieved by using just proper nouns or instances in the three domains.

	Right Parties	Cinque Stelle	Democratic Party
Right Parties	99.68%	0.0%	0.32%
Cinque Stelle	0.00%	100.00%	0.00%
Democratic Party	0.00%	0.00%	100.00%

Table 3: Prediction results of the prediction of the parties of members of the Italian parliament.

preferences of the followers based on the vocabulary used. We only considered the followers of politicians of just one coalition, thereby excluding those followers who observe politics from a neutral perspective (e.g. journalists).

Results of the first experiment are presented in the table 3. The method is extremely accurate in classifying the accounts of the elected politicians, suggesting that indeed they have a very different vocabulary.

Although we use proper nouns in the vocabulary, it is interesting to show the most frequent nouns which belong to the vocabulary of the three classes of politicians. We select from the three parties the most frequent 100 words, and we report in table 4 the twenty most frequent words. They reflect our expectations, with words such as *security* or *italian* in the vocabulary of the Right Parties, *euro*,

movement or *live* in the vocabulary of Cinque Stelle, and *woman*, *commitment*, *family* in the vocabulary of Democratic Party.

Results of the second experiment, reported in Table 5, are rather surprising and have an interesting sociological interpretation. We note that the method correctly predicts the followers of the Democratic Party (100% accuracy) and of Right Parties (96% accuracy). For what concerns Cinque Stelle, however, the predictor only achieved 40% accuracy, while it classified the followers as politically closer to the Democratic Party (60%) and not to the Right Paries (0%). This is an indication that the followers of Cinque Stelle do not have a distinctive vocabulary, and have stronger similarity to the Democratic Party than to the Right Parties. These results are confirmed by the dispersion indexes, which show stronger dispersion for Cinque Stelle (see Table 6).

4.2 Targeted Advertising

The most important application of community detection from a commercial point of view is targeted advertising. For this purpose, we assume that the advertiser already knows a community of interest, e.g. thanks to activities that the community has already performed upon controlled social platforms. The advertiser's objective is to enlarge the community by finding new candidate accounts. For this purpose, we use as start-point a method consisting in extracting candidate accounts on the basis of their co-occurrence with community members; these accounts can be ranked on the basis of the distance to the community centroids, thereby creating an extended community and evaluating its dispersion. Advertising will be addressed to the new members of the community.

Among the many possible examples of applications, we consider the Roma Jazz Festival, an event which occurred in August 2018; our objective is to find the Twitter accounts which are most suited for targeted advertising. We initially select 20 followers of the event, which we take as members of a community of the Roma Jazz Festival fans. We then consider as candidates all the Twitter accounts which are mentioned in the tweets of the community members; the rationale of the method is that followers of Jazz events most likely talk about accounts who are also interested in Jazz events, as accounts tends to share their interests. We build the community's vocabulary and then rank each candidate according to its distance from the vocabulary's centroid, thereby detecting relevant accounts for targeted advertising. Table 7 shows the dispersion index of the community which is created by accepting the first n candidates, with n varying from 10 to 100; the dispersion index slowly increases while the community increases in size. Table 7 also shows that the community membership problem defined in Section 2.1 is highly influenced by the considered threshold; in targeted advertising, such threshold typically depends on the available resources and on the cost of advertising.

5 RELATED WORK

Community detection is a fundamental task in social network analysis [9]. In the following we describe related work by considering methods that use links, semantics and content.

5.1 Networks Clustering

The majority of approaches to community detection use social links (followers, retweets and user mentions) in order to detect communities as clusters of strongly (or densely) connected subgraphs [18], [23]. Community detection in large graphs is a wide research topic, applied to many domains such as sociology, biology and finance. The methods used to detect community structures in graphs are based on modularity optimization [2] [3], agglomerative clustering, centrality based and clique percolation [8]. Leskovec et al. compared a multitude of community discovery algorithms, and discovered the trade-offs between clustering objectives and community compactness [12].

In general, all methods which take into account are computationally expensive in data acquisition, because in order to reconstruct significant sub-graphs it is necessary to make many queries to the Twitter API. Moreover, they cannot investigate on the similarity of users who are not linked by social links.

5.2 Semantic Methods

Another class of approaches uses the semantic content of social graphs to discover communities. [19] introduces a measure of signal strength between two nodes in the social network by using content similarity. In [24] the authors propose the CUT (Community-User-Topic) model for discovering communities using the semantic content of the social graph. Communities are modeled as random mixtures over users who in turn have a topical distribution (interest) associated with them.

Other works use generative probabilistic modeling which considers both contents and links as being dependent on one or more latent variables, and then estimates the conditional distributions to find community assignments. PLSA-PHITS [6], Community-User-Topic model [24] and Link-PLSA-LDA [15] are representatives in this category. For instance, link-PLSA-LDA finds latent topics in text and citations and assumes different generative processes on citing documents, cited documents as well as citations themselves. Text generation follows the LDA approach, and link creation between citing and cited documents is controlled by topic-specific multinomial distributions.

In these approaches, content similarity between users play a fundamental role, thereby underlining the relevance of content in community detection. These approaches have the same drawbacks in the data acquisition cost that was reported above.

5.3 Content-based Methods

Other works are more similar to our approach, as they use textual similarity, without deep semantic analysis. [21] proposes a method to cluster people in Twitter using words, by proposing a metric to weight the words; [14] proposes a method for computing user similarity based on a network representing the semantic relationship between the words occurring in the same tweet and the related topic. Other methods discover user similarities based on content similarities; the method presented in [10] uses a regression model. Compared to our approach, these methods require a lot of training data for building an accurate model of the terms used by Twitter accounts and are more focused on similarity discovery rather than community detection.

	Right Parties Nouns	Frequencies	Cinque Stelle Nouns	Frequencies	Democratic Party Nouns	Frequencies
0	government	0.020525	citizen	0.012416	job	0.014083
1	job	0.010293	job	0.010520	year	0.013420
2	year	0.010284	year	0.009318	government	0.012428
3	country	0.010215	law	0.009112	law	0.010318
4	right party	0.008931	government	0.008677	country	0.008362
5	brother	0.008686	star	0.008464	thing	0.007921
6	italian	0.008632	movement	0.007976	campaign	0.006723
7	president	0.008092	live	0.007611	day	0.006648
8	vote	0.007544	away	0.006767	person	0.006546
9	feature	0.007517	chamber	0.006494	citizen	0.005896
10	region	0.006502	country	0.006303	president	0.005836
12	tax	0.005896	program	0.005984	favour	0.005707
13	program	0.005862	president	0.005657	vote	0.005454
14	thing	0.005737	number	0.005653	woman	0.005443
15	citizen	0.005704	million	0.005204	club	0.005034
16	politics	0.005693	thing	0.005199	commitment	0.004850
17	security	0.005420	video	0.004862	hour	0.004712
18	day	0.005316	euro	0.004806	politics	0.004536
19	person	0.005312	city	0.004771	family	0.004435
20	state	0.005169	proposal	0.004529	program	0.004333

Table 4: Most recurrent nouns in the vocabulary of 20 elected members of the Italian parliament, ranked by their frequency.

	Right	Cinque Stelle	Democr.
Right parties followers	96%	0	4%
Cinque Stelle followers	0	40%	60%
Democratic Party followers	0	0	100%

Table 5: Prediction of political preferences of the followers of politicians of the three parties.

	Right	Cinque Stelle	Democr.
dispersion index	0.34	0.58	0.48

Table 6: Dispersion index for the followers of politicians of the three parties.

	dispersion index
10	0.553
20	0.554
30	0.555
40	0.557
50	0.559
60	0.560
70	0.561
80	0.563
90	0.565
100	0.566

Table 7: Dispersion index of the firs n candidates.

6 CONCLUSIONS

This study provides a systematic approach to the characterization of the vocabulary used within a community of Twitter accounts, which acts as a community fingerprint. We provide a characterization of syntactic and semantic features that contribute to the vocabulary, and then show which features are most suited for testing community membership and cohesiveness. The use of the vocabulary for community detection is very efficient, as it requires only direct access to each Twitter account rather than much more expensive access to Twitter interactions. Moreover, the vocabulary hints to the the typical topics discussed within the community, thereby providing an interesting characterization of the community from a sociological perspective.

Future work includes the transfer of the proposed method to other social networks, e.g. on Facebook using accounts and posts, to test the approach on different platforms.

ACKNOWLEDGEMENT

This work was partially supported by the ERC Advanced Grant 693174, Data-Driven Genomic Computing.

REFERENCES

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference (ISWC'07/ASWC'07)*. Springer-Verlag, Berlin, Heidelberg, 722–735.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), P10008.

[3] Vincent D Blondel, Jean loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks.

[4] Marco Brambilla, Stefano Ceri, Florian Daniel, Marco Di Giovanni, Andrea Mauri, and Giorgia Ramponi. 2018. Iterative Knowledge Extraction from Social Networks.

In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1359–1364. https://doi.org/10.1145/3184558.3191578

[5] Marco Brambilla, Stefano Ceri, Emanuele Della Valle, Riccardo Volonterio, and Felix Xavier Acero Salazar. 2017. Extracting Emerging Knowledge from Social Media. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 795–804. https://doi.org/10.1145/3038912.3052697

[6] David Cohn and Thomas Hofmann. [n. d.]. *The Missing Link-A Probabilistic Model of Document Content and Hypertext Connectivity*. Technical Report.

[7] Susan T. Dumais. [n. d.]. Latent semantic analysis. *Annual Review of Information Science and Technology* 38, 1 ([n. d.]), 188–230. https://doi.org/10.1002/aris.1440380105

[8] Santo Fortunato. 2010. Community detection in graphs. *Physics Reports* 486, 3 (2010), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

[9] M. Girvan and M. E. J. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99, 12 (2002), 7821–7826. https://doi.org/10.1073/pnas.122653799 arXiv:http://www.pnas.org/content/99/12/7821.full.pdf

[10] Ashish Goel, Aneesh Sharma, Dong Wang, and Zhijun Yin. [n. d.]. *Discovering Similar Users on Twitter*. Technical Report.

[11] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07)*. ACM, New York, NY, USA, 56–65. https://doi.org/10.1145/1348549.1348556

[12] Jure Leskovec, Kevin J Lang, and Michael W Mahoney. 2010. *Empirical Comparison of Algorithms for Network Community Detection*. Technical Report. arXiv:arXiv:1004.3539v1

[13] Lei Li, Wei Peng, Saurabh Kataria, Tong Sun, and Tao Li. 2015. Recommending Users and Communities in Social Media. *ACM Trans. Knowl. Discov. Data* 10, 2, Article 17 (Oct. 2015), 27 pages. https://doi.org/10.1145/2757282

[14] Stefano Mizzaro, Marco Pavan, and Ivan Scagnetto. 2015. Content-Based Similarity of Twitter Users. In *Advances in Information Retrieval*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.). Springer International Publishing, Cham, 507–512.

[15] Ramesh Nallapati and William W. Cohen. 2008. Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs. In *ICWSM*.

[16] Mert Ozer, Nyunsu Kim, and Hasan Davulcu. 2016. Community Detection in Political Twitter Networks using Nonnegative Matrix Factorization Methods. (2016). https://doi.org/10.1109/ASONAM.2016.7752217 arXiv:1608.01771

[17] Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali, and Ploutarchos Spyridonos. 2012. Community detection in Social Media. *Data Mining and Knowledge Discovery* 24, 3 (01 May 2012), 515–554. https://doi.org/10.1007/s10618-011-0224-z

[18] Yulong Pei, Nilanjan Chakraborty, and Katia Sycara. 2015. Nonnegative Matrix Tri-factorization with Graph Regularization for Community Detection in Social Networks. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, 2083–2089. http://dl.acm.org/citation.cfm?id=2832415.2832538

[19] Yiye Ruan, David Fuhry, and Srinivasan Parthasarathy. 2013. Efficient Community Detection in Large Networks Using Content and Links. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 1089–1098. https://doi.org/10.1145/2488388.2488483

[20] Mrinmaya Sachan, Danish Contractor, Tanveer A. Faruquie, and L. Venkata Subramaniam. 2012. Using Content and Interactions for Discovering Communities in Social Networks. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 331–340. https://doi.org/10.1145/2187836.2187882

[21] Kuldeep Singh, Harish Kumar Shakya, and Bhaskar Biswas. 2016. Clustering of people in social network based on textual similarity. *Perspectives in Science* 8 (2016), 570 – 573. https://doi.org/10.1016/j.pisc.2016.06.023 Recent Trends in Engineering and Material Sciences.

[22] Jun Wang, Jiaxu Peng, and Ou Liu. 2015. A classification approach for less popular webpages based on latent semantic analysis and rough set model. *Expert Systems with Applications* 42, 1 (2015), 642 – 648. https://doi.org/10.1016/j.eswa.2014.08.013

[23] B. Yang and S. Manandhar. 2014. Community discovery using social links and author-based sentiment topics. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. 580–587. https://doi.org/10.1109/ASONAM.2014.6921645

[24] Ding Zhou, Eren Manavoglu, Jia Li, C. Lee Giles, and Hongyuan Zha. 2006. Probabilistic Models for Discovering e-Communities. In *Proceedings of the 15th International Conference on World Wide Web (WWW '06)*. ACM, New York, NY, USA, 173–182. https://doi.org/10.1145/1135777.1135807