ARTICLE IN PRESS

Information Processing and Management xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Content-based characterization of online social communities

Giorgia Ramponi[1,*], Marco Brambilla, Stefano Ceri, Florian Daniel, Marco Di Giovanni[1]

Politecnico di Milano, Via Ponzio 34/5, Milano 20133, Italy

ARTICLE INFO

Keywords:
Communities characterization
Social networks analysis
Content-based data analytics
Twitter

ABSTRACT

Nowadays social networks are becoming an essential ingredient of our life, the faster way to share ideas and to influence people. 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 analyze the problem of community detection from a content analysis perspective: we argue that the content produced in social interaction is a very distinctive feature of a community, hence it can be effectively used for community detection. We analyze the problem from a textual perspective using only syntactic and semantic features, including high level latent features that we denote as *topics*.

We show that, by inspecting the content used by tweets, we can achieve very efficient classifiers and predictors of account membership within a given community. We describe the features that best constitute a vocabulary, then we provide their comparative evaluation and select the best features for the task, and finally we illustrate an application of our approach to some concrete community detection scenarios, such as Italian politics and targeted advertising.

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 (Java, Song, Finin, & Tseng, 2007; Li, Peng, Kataria, Sun, & Li, 2015; Papadopoulos, Kompatsiaris, Vakali, & Spyridonos, 2012). Digital social content can be automatically inspected, hence, social communities on Internet can be detected by algorithms (Ozer, Kim, & Davulcu, 2016; Papadopoulos et al., 2012; Sachan, Contractor, Faruquie, & Subramaniam, 2012); 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 (Fortunato, 2010) (e.g. the *friendship/follow* relationships), or take into account social activities (Sachan et al., 2012) (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

* Corresponding author.
E-mail addresses: giorgia.ramponi@polimi.it (G. Ramponi), marco.brambilla@polimi.it (M. Brambilla), stefano.ceri@polimi.it (S. Ceri), florian.daniel@polimi.it (F. Daniel), marco.digiovanni@polimi.it (M. Di Giovanni).

[1] Equal contribution.

https://doi.org/10.1016/j.ipm.2019.102133
Received 6 April 2019; Received in revised form 12 September 2019; Accepted 25 September 2019
0306-4573/

Please cite this article as: Giorgia Ramponi, et al., Information Processing and Management, https://doi.org/10.1016/j.ipm.2019.102133

different direction, and propose a **content-based approach to community detection**. We conjecture that a community can be characterized by the content that they share, 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 shared content has strong similarity properties; we can also test if a social actor is a member of a community by comparing the actor's content to the community's content. 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 Brambilla, Ceri, Della Valle, Volonterio, and Acero Salazar (2017) we defined a method for extracting emerging knowledge from social accounts based on co-occurrence of accounts with known members of a community; in Brambilla et al. (2018) 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 a systematic study of social content features that best characterized a community. Preliminary work (Ramponi, Brambilla, Ceri, Daniel, & Di Giovanni, 2019) considered fewer textual features (in particular, no latent feature) and fewer contexts of application; in this work we show that the new latent features are relevant and actually have the best performance in the new contexts.

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 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 the selection of textual features. 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 features associated to accounts, so that we can test community's strength and membership.

The research question underlying these challenges is to ascertain how much communities can be guessed by considering just the content of their social interaction. 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.

Although our approach applies to possibly large communities (e.g., the followers of politicians, as shown in Table 5), our approach is best suited to the characterization of small communities with highly specialized vocabulary, where the method performs remarkably well; problems that exhibit these features have significant applicability, discussed in Section 5 and in the conclusions.

This paper is organized as follows. In Section 2 we define some metrics used later in the paper for distance, dispersion and coherence. In Section 3 we define the syntactic and semantic features used to perform the analysis and describe the methods for extracting, while in Section 4 we select the most effective features for testing a community's strength and membership. In Section 5, we assess the power of content in two important applications, related to detection of communities in the political arena and to targeted advertising. We present related work in Section 6 and conclusions in Section 7.

2. Background

2.1. Definitions

We introduce some useful definitions in the community detection problem.

- *Community*: a community C 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.
- *Feature Vector*: we associate to every member or candidate c a *feature vector* $f_c = \langle f_{c,1}, f_{c,2}, .f_{c,n} \rangle$, whose elements are the frequencies of the textual features that we extract from a corpus consisting of the last 200 tweets of c. Thus, if for example we are considering nouns, $f_{c,i}$ is the frequency of use of the noun i in c's tweets.
- *Centroid*: given m feature vectors $\{f_1, \ldots f_m\}$ of cardinality n, we define the centroid:

$$z = \langle z_1, .., z_n \rangle$$

where:

$$z_i = \frac{1}{m}\sum_{c=1}^{m} f_{c,i}$$

2.2. Distance metrics

To evaluate the closeness of a candidate c to the centroid z we consider four distance metrics: Manhattan distance, Euclidean distance, Cosine distance, Kullback–Leibler Divergence.

2.3. 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.

2.4. Coherence metric

We can define the coherence of a text as a âæcontinuity of sensesâģ which requires arguments to be logically connected. In topic modeling, a coherent model is capable of describing a set of topics in a rigorous way. Measuring coherence is a complex task, but we refer to the work of Röder, Both, and Hinneburg (2015) which provides a systematic study on different coherence measures, and proposes C_V as the best one.

C_V is obtained by evaluating all the possible combinations of four different dimensions and picking the one that performed best on a given dataset evaluated by humans:

1. the first dimension represents the type of segmentation used to divide the word set into subsets. C_V uses a *one-one* approach, where every pair of words is selected;
2. the second dimension represents how probabilities are derived. C_V uses *Boolean Sliding Window* with window size of 10. The probability is calculated as the number of windows in which the word occurs divided by the total number of windows;
3. the third dimension is the Confirmation Measure, defining a way to compute how strong a word set supports another one. C_V uses *indirect cosine measure* to calculate cosine similarities between vectors obtained with the direct *normalized log-ratio* measure;
4. the fourth dimension concerns the aggregation of all subset scores to a single score. C_V uses the *simple average* of all the values.

The detailed description can be found in the original paper (Röder et al., 2015).

3. Content Features Description and Creation

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 text, hyperlinks and images; we focus on the text, consisting of words and hashtags, and build the syntactic or semantic features that describe a set of tweets, arbitrarily collected.

3.1. Syntactic features

Words appearing in the tweets are classified on the basis of their syntactic features and recognizing, in particular, verbs and nouns. Syntactic analysis consists essentially in associating them with their frequency in the tweet corpus.

The extraction process consists in a standard text pre-processing by deleting stop-words, tokenizing and tagging the text and retrieving the root form of the words, using the NLTK library.[2] After pre-processing, we focus on words carrying three different tags: nouns, verbs and proper nouns, which are a subset of nouns. Those sets of words are then vectorized using Term Frequency (TF) vectorizer.

3.2. 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 used two kinds of semantic features: knowledge-based features, and topic features, obtained by using topic detection techniques.

3.2.1. Extraction of semantic knowledge-based features

Knowledge-based features are extracted after text matching with a structured knowledge graph; as we do not focus on a specific

[2] http://www.nltk.org

domain of interest, we use DBpedia,[3] which is publicly available and easily accessible through APIs; it provides structured content from the information created in Wikipedia Auer et al. (2007).

In order to extract semantic features from tweets we used Dandelion,[4] a commercial software which matches a text to DBpedia entities. We then consider a term as semantically understood when it is matched to either a type or an instance, defined as follows:

- *type*: a *type* is an element of the DBpedia hierarchy; Dandelion produces matches with associated probability and we use the default threshold value (0.6).[5]
- *instance*: some words are also associated to a concept that has a page in Wikipedia; we call these concepts *instances*.

After extracting types and instances, we produce a vector by using the Term Frequency (TF) vectorizer.

3.2.2. Extraction of semantic topic features

Topic features are learned using the Latent Dirichlet Allocation process (Blei, Ng, & Jordan, 2003); the process learns the relations between words in documents and creates a fixed number of topics; each topic, in turn, is associated with a probability distribution Φ over the words that are recognized as significant for that topic.

To consolidate the use of LDA in our context, we have to decide how to set an ideal number of topics, which is a prerequisite of the method. We consider the corpus of tweets of a specific domain and divide it into a training and testing set. We build 50 different models, each one with an incremental number of topics (from 1 to 50), and for each of them we calculate the C_V coherence (Section 2.4). Then we selected the number of topics yielding to a model with the highest value of coherence. Fig. 1 shows result of our analysis for the specific corpus of tweets about chess players (discussed in the next section); in that specific corpus of Tweets, we select 7 as best number, which is also the length of the topics feature vector to be used in the analysis. In most corpuses, the best coherence value is small[6]; curves have the sharp behavior described in Fig. 1, thus the selection of the ideal number of topics is not difficult.

Given a specific tweet, LDA associates it with a probability distribution over the topics. We use this as topic features vector. For implementing the LDA model we use the Gensim library (Řehůřek & Sojka, 2010).

4. Evaluation

We can then formulate the problem of *finding the best set of features and the most effective distance metric in order to characterize community membership*. Given a community $C^* = \{c_1, ..., c_n\}$, we retrieve the tweets of these accounts and construct one feature vector for each of the six textual features discussed above. From these feature vectors, six centroids z_{type}, $z_{instance}$, z_{noun}, z_{verb}, $z_{propernoun}$ and z_{topic} are created.

We then explore which combination of textual features and distance metrics achieves 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.

The experiment 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, measure their effectiveness in ranking the top candidates, and select the features and distances associated with the best rankings.

4.1. Input data and experiment design

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. It can be seen that accuracies are highly dependent on the domain, meaning that there are communities harder to characterize because their vocabulary is not specialized enough.

The communities are formed by fashion designers, Australian writers, and chess players:

- **Fashion designers**: the research team of the Fashion In Process Lab[7], in the original experiment, collected emerging Italian brands, and we used 19 of them;
- **Australian writers**: we considered some fiction authors engaged in the Melbourne Emerging Writers Festival[8] by picking 20 accounts from the participants to the event;
- **Chess players**: we used a list of 20 top chess players and their accounts.[9]

For every Twitter account we select at most the last 200 tweets, which correspond to a single Twitter API call; exact sizes are

[3] https://wiki.dbpedia.org
[4] https://dandelion.eu
[5] The threshold is for the confidence value of the annotation extraction https://dandelion.eu/docs/api/datatxt/nex/v1/.
[6] In the domains discussed in Section 4 and 5 it ranges between 4 and 10.
[7] http://www.fashioninprocess.com
[8] http://www.emergingwritersfestival.org.au
[9] https://www.reddit.com/r/chess/comments/32t5ov/list_of_top_chess_player_journalist_twitter

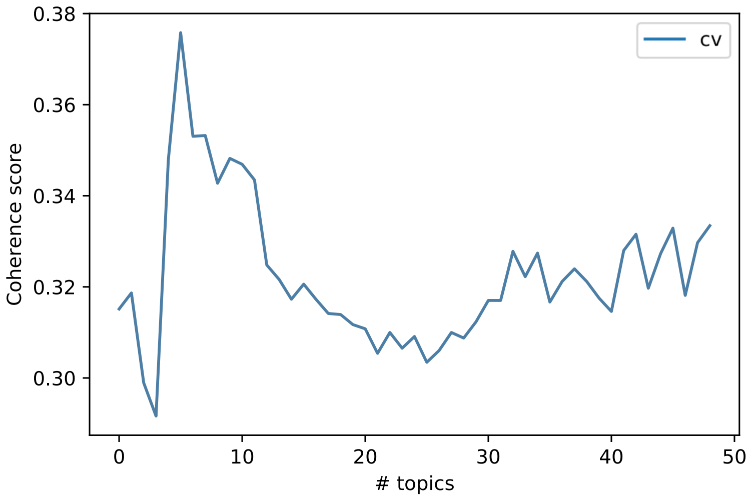

Fig. 1. Analysis of C_V coherence values for chess players.

reported in Table 1. Data have been collected on 08/02/2018.

The anonymized dataset is available at https://doi.org/10.7910/DVN/VWLEAA.

4.2. Experiment design

For every 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.

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

4.3. Comparison

Table 2 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 six syntactic and semantic features, we note that (syntactic) proper nouns and (semantic) topics and instances also provide the best precision and recall in all domains. Instances obtain comparable results to topics and proper names, but their extraction requires an interaction with a commercial software whose free use is limited in rate, so we exclude this feature from our further analysis.

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 170 candidates; and it is rather good for all domains, including Australian writers, as we find 4 members within the top ten ranked.

4.4. Dispersion indexes

We inspected the Twitter accounts of chess players, 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 writers 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

Table 1
Sizes of datasets.

	Number of users	Number of tweets
Fashion designers	19	1536
Australian writers	20	1953
Chess players	20	2262

Table 2
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. We use labels CD for cosine distance, KLD for Kullback–Leibler Divergence, l1 for Manhattan distance and l2 for Euclidean distance.

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
	Instance	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
	Topic	0.726	0.824	0.702	0.834	0.734	0.868	0.732	0.822
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
	Instance	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
	Topic	0.780	0.870	0.736	0.816	0.654	0.764	0.656	0.748
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
	Instance	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
	Topic	0.522	0.642	0.444	0.570	0.406	0.532	0.378	0.484

Section 2.3, whose values for the three communities are summarized in Table 3 (a high index is indicative of high dispersion).

4.5. Topic explanation

Topics are explained by their most recurrent words; in Table 4 we report the first 5 words explaining the first topic for each of the three domains. As we can see, in Chess players the best topic contains the word *chess* and *game*; the best topic for Fashion contains the word *love*.

4.6. Conclusion of the evaluation

After this analysis, we conclude that the best features are proper nouns and topics (associated with any distance). The former is a syntactic feature, describing terms which denote concrete aspects of reality; the latter is a latent semantic feature, representing the texts in their entirety.

The full code is available at https://github.com/DataSciencePolimi/-Characterization-of-Online-SocialCommunities. In the next section we propose two applications, showing that each selection can be the most useful for characterizing specific social communities.

5. Applications

5.1. Content-based analysis of accounts from a political perspective

One of the most interesting applications of content-based community detection is concerned with understanding 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 some politicians from the three coalitions, and we retrieved their last tweets (a single Twitter API call per user). 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

Table 3
Dispersion index for the three domains.

Features	Domain		
	AW	Fashion	Chess
NNP	0.84	0.79	0.55
instances	0.80	0.73	0.63

Table 4
Best topics that represents the three domains with their first 5 components.

Domain	Topics				
Chessplayers	raider	italy	owner	chess	playoff
Fashion	day	time	thank	get	love
AW	person	thank	time	thing	way

Table 5
Sizes of political parties datasets.

	Number of users	Number of tweets
Right parties	19	2174
Cinque Stelle	20	2295
Democratic Party	25	3452
Right parties followers	126	4948
Cinque Stelle followers	289	16,145
Democratic Party followers	306	17,201

their similarity to the centroid; we repeated this experiment 50 times, every time selecting randomly the accounts to use as community members. Data have been collected on 18/04/2018.

- 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 preferences of the followers based on the vocabulary used. We considered the followers of politicians of just one of the three coalitions, thereby excluding those followers who observe politics from a neutral perspective (e.g. journalists). Data have been collected on 06/05/2018.

The anonymized dataset is available at https://doi.org/10.7910/DVN/VWLEAA. Sizes of the datasets are reported in Table 5. Results of the first experiment are presented in Table 6. The method is extremely accurate in classifying the accounts of the elected politicians, suggesting that indeed they have a very different vocabulary.

In Table 8 we report the most frequent proper nouns for the three parties. As you can see it is not easy to interpretate this feature because proper nouns are too specifically connect with factual people, location or events occurring in Italy. Consider for instance that top mentioned proper nouns include Bologna, Milano, Calabria for Democratic Party, Friuli for the Right Party, Roma and Torino for Cinque Stelle, and these are locations where each party is either historically strong or actually at the local government.

To show the different vocabularies between parties we present most frequent nouns, that are slightly less effective than proper nouns in characterizing communities, but can be best perceived by readers based upon general knowledge. The three lists have many common terms in any conversation (e.g. day, year) or in any conversation of politicians (e.g. government, job, program, country, or law, citizen appearing in two lists out of three) and at first sight look very similar; but if one looks at terms which appear just in one list, finds Italian, tax, security in the Right Party, movement, live in Cinque Stelle and campaign, woman, club, commitmentâg in the Democratic Party; we can clearly see that the different vocabulary characterize the parties (Table 9).

Results of the second experiment, reported in Table 7, 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 Parties (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 12).

We repeat the experiment using topics as features. As we can see in Table 10 for the first analysis and in Table 11 for the second analysis, the results are not satisfying, as the method doesn't succeed in classifying political parties. A likely reason is that, while nouns are very indicative of a party, topics are not, as tweets written by politicians end up having the same topics regardless of their party.

Table 6
Prediction of parties of members of the Italian parliament using proper nouns.

	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 7
Prediction of parties of the followers of politicians using proper nouns.

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

Table 8
Most recurrent proper nouns in the vocabulary of 20 elected members of the Italian parliament, ranked by their frequency.

Democratic Party NNP	Frequencies	Right Parties NNP	Frequencies	Cinque Stelle NNP	Frequencies
Italia	0.085716	Italia	0.108296	Roma	0.069347
Bologna	0.049067	Europa	0.043148	Italia	0.042250
Roma	0.025675	Roma	0.033982	Cittá	0.026740
San	0.018526	Lazio	0.032541	Luigi	0.025314
Europa	0.014398	Liguria	0.021148	San	0.020323
Milano	0.011444	Forza	0.017809	Berlusconi	0.019966
Calabria	0.011142	San	0.014928	Piazza	0.018094
Berlusconi	0.009397	Friuli	0.014535	Torino	0.015955
Venezia	0.008994	Laura	0.014535	Augusta	0.015242
Forza	0.008390	Franco	0.013226	Sala	0.013459

Table 9
Most recurrent nouns in the vocabulary of 20 elected members of the Italian parliament, ranked by their frequency. Nouns were translated from Italian to English by the authors.

	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 10
Prediction of the parties of members of the Italian parliament using topic features.

	Right Parties	Cinque Stelle	Democratic Party
Right Parties	52%	17%	31%
Cinque Stelle	53%	24%	23%
Democratic Party	48%	26%	26%

5.2. Targeted advertising

From a commercial point of view, the most important application of community detection is targeted advertising. We assume that the advertiser already knows a community of interest, e.g. thanks to activities that the community has already performed in controlled social platforms. The advertiser's objective is to enlarge the community by finding new candidate accounts, thus potential new customers.

Among the many possible examples of applications, we consider sport events, in particular baseball or football events, where we

Table 11
Prediction of the followers of politicians of the three parties.

	Right	Cinque Stelle	Democr.
Right parties followers	52%	17%	31%
Cinque Stelle followers	16%	17%	66%
Democratic Party followers	14%	16%	70%

Table 12
Dispersion index for the followers of politicians of the three parties.

	Right	Cinque Stelle	Democr.
Dispersion index	0.34	0.58	0.48

initially know a set of accounts of players of those two sports. In such case, the advertiser's interest is to broaden the set of accounts that she can reach by adding similar accounts to the initial set. Following a pipeline similar to the one described before, we manually collected Baseball players and Football players of UCF (University of Central Florida), and randomly split them in a set of 10 accounts that represents the already known community, and a set of accounts that we expect to retrieve when mixed with random Twitter accounts. In Table 13, the sizes of the datasets are reported. Data have been collected on 22/02/2018. The anonymized dataset is available at https://doi.org/10.7910/DVN/VWLEAA. In Table 14 we compare the results obtained when using NNP and topic features, using the cosine distance.

In this case, topic features achieve the best performances in the two communities, as the community of baseball players and Football players have very distinctive interests that are different from random accounts. They generally talk about the same latent topic (sport), thus the best results are obtained by the topic-based method.

6. Related work

Community detection is a fundamental task in social network analysis (Girvan & Newman, 2002). In the following we describe related work by considering methods that use links, semantics and content.

6.1. Network 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 (Pei, Chakraborty, & Sycara, 2015), (Yang & Manandhar, 2014). 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 (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) (Blondel et al., 2008), agglomerative clustering, centrality based and clique percolation (Fortunato, 2010). Leskovec et al. compared a multitude of community discovery algorithms, and discovered the trade-offs between clustering objectives and community compactness (Leskovec, Lang, & Mahoney, 2010).

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. We remark that we cannot compare our results with these network based approaches since our method does not require that users are socially connected. The networks of the datasets investigated in this paper could even have no edges at all, resulting in meaningless networks measures, such as modularity (Newman & Girvan, 2004).

A similar approach can be found in Singh, Singh, Kumar, and Biswas (2019b) and Singh, Kumar, Singh, and Biswas (2019a) where the authors deal with Influence Maximization task by including topic information to traditional information diffusion models on networks.

6.2. Semantic methods

Another class of approaches uses the semantic content of social graphs to discover communities. Ruan, Fuhry, and Parthasarathy (2013) introduces a measure of signal strength between two nodes in the social network by using content similarity. In

Table 13
Sizes of UCF players datasets.

	Number of users	Number of tweets
Baseball players	62	5727
Football players	129	12,500

Table 14
Comparison of precision10 using NNP and Topic features in the sport domain: Baseball and Football players.

Domain	Feature	
	NNP	Topic
Baseball players	0.29	0.76
Football players	0.12	0.76

Zhou, Manavoglu, Li, Giles, and Zha (2006) 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 (Cohn & Hofmann, 2001), Community-User-Topic model (Zhou et al., 2006) and Link-PLSA-LDA (Nallapati & Cohen, 2008) 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.

6.3. Content-based methods

Other works are more similar to our approach, as they use textual similarity, without deep semantic analysis. Singh, Shakya, and Biswas (2016) proposes a method to cluster people in Twitter using words, by proposing a metric to weight the words; Mizzaro, Pavan, and Scagnetto (2015) 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 Goel, Sharma, Wang, and Yin (2013) uses a regression model. Compared to our approach, these methods require a lot of data for building an accurate model of the terms used by Twitter accounts and are more focused on similarity discovery rather than community detection.

7. Conclusions

This study provides a systematic approach to user identification and community characterization in Twitter. We provide a characterization of syntactic and semantic features that appear in a corpus of tweets, and then show which features are most suited for testing community membership and cohesiveness. Proper nouns or latent content topics perform very well if used with cosine distance or Kullback–Leibler Divergence.

In several application contexts, our method achieves a precision@10 which is 70% or above (in our designed experiment, this means that only 3 accounts are incorrect out of 10, extracted from a total of 190 candidates, mediated over 50 executions). This result is particularly remarkable if one considers that the proposed method is low-cost: it requires the extraction of the tweets of a candidate (through a single call to the standard Twitter API) and then running simple scripts (which internally call standard libraries) for extracting from this corpus the frequencies of either topics or proper nouns; as we opted for a low-cost strategy, we preferred topics to instances as representative semantic features.

Our applications show one case where a syntactic feature prevails over a semantic one (politics) but also one case where a semantic feature prevails over a syntactic one (targeted advertising for sports players). Moreover, the topic components (or even better the most frequently used nouns) hint at the typical terms used within the community, thereby providing an interesting characterization of the community from a sociological perspective.

As input, the described method requires only few examples of reference accounts considered similar by a domain expert, e.g., chess players or writers, to construct a sufficiently characterizing vocabulary. Keeping the size of the input low was one of the design goals of our work (to keep a manual search task manageable). However, we have also verified that the approach is robust with respect to larger input sizes, as shown in Tables 1, 5 and 13. Datasets of different magnitudes and belonging to communities with both low and high specialized vocabularies have been tested, and the results in terms of performance of the method are comparable.

The practical implication of our study is in extraction of targeted communities where each new candidate brings potentially high value, e.g. we used in the past a similar method to extract emerging fashion designers as candidates to participate to exhibits in a joint study with domain experts of our University [7]; the catalog of emerging designers had a high value for our colleagues. Targeted advertising as discussed in Section 5.2 is applicable to many contexts, e.g. under elections by election candidates who want to mail advertising just to potential voters of their party.

Future work includes the transfer of the proposed method to other social networks, e.g., Instagram or Facebook, inspecting the performance of the algorithm in scenarios where communities have different social ties and may be defined more or less strongly by their vocabulary. We also plan to enlarge the sizes of datasets to understand if it is possible to obtain more accurate characterizations of domains (at the same time also studying the scalability of the approach) and to study the effectiveness of the approach as an instrument to characterize polarity in online discussions.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data. The semantic web*. Springer722–735.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Brambilla, M., Ceri, S., Daniel, F., Di Giovanni, M., Mauri, A., & Ramponi, G. (2018). *Iterative knowledge extraction from social networks. Companion proceedings of the web conference 2018WWW '18*International World Wide Web Conferences Steering Committee1359–1364. https://doi.org/10.1145/3184558.3191578.

Brambilla, M., Ceri, S., Della Valle, E., Volonterio, R., & Acero Salazar, F. X. (2017). *Extracting emerging knowledge from social media. Proceedings of the 26th international conference on world wide webWWW '17*Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee795–804. https://doi.org/10.1145/3038912.3052697.

Cohn, D. A., & Hofmann, T. (2001). *The missing link-a probabilistic model of document content and hypertext connectivity. Advances in neural information processing systems*430–436.

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

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799.

Goel, A., Sharma, A., Wang, D., & Yin, Z. (2013). *Discovering similar users on twitter. 11th workshop on mining and learning with graphs*.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). *Why we twitter: Understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD2007 workshop on web mining and social network analysis*. ACM56–65.

Leskovec, J., Lang, K. J., & Mahoney, M. (2010). *Empirical comparison of algorithms for network community detection. Proceedings of the 19th international conference on world wide web*. ACM631–640.

Li, L., Peng, W., Kataria, S., Sun, T., & Li, T. (2015). Recommending users and communities in social media. *ACM Transactions on Knowledge Discovery from Data (TKDD), 10*(2), 17.

Mizzaro, S., Pavan, M., & Scagnetto, I. (2015). Content-based similarity of twitter users. In A. Hanbury, G. Kazai, A. Rauber, & N. Fuhr (Eds.). *Advances in information retrieval* (pp. 507–512). Cham: Springer International Publishing.

Nallapati, R., & Cohen, W. W. (2008). *Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. Icwsm*84–92.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 026113.

Ozer, M., Kim, N., & Davulcu, H. (2016). *Community detection in political twitter networks using nonnegative matrix factorization methods. Proceedings of the 2016 IEEE/ACM international conference on advances in social networks analysis and mining*. IEEE Press81–88.

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. (2012). Community detection in social media. *Data Mining and Knowledge Discovery, 24*(3), 515–554.

Pei, Y., Chakraborty, N., & Sycara, K. (2015). *Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. Proceedings of the 24th international conference on artificial intelligenceIJCAI'15*AAAI Press2083–2089 URL http://dl.acm.org/citation.cfm?id=2832415.2832538

Ramponi, G., Brambilla, M., Ceri, S., Daniel, F., & Di Giovanni, M. (2019). *Vocabulary-based community detection and characterization. Proceedings of the 34th ACM/SIGAPP symposium on applied computing*. ACM1043–1050.

Řehůřek, R., & Sojka, P. (2010). *Software framework for topic modelling with large corpora. Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*. Valletta, Malta: ELRA45–50 http://is.muni.cz/publication/884893/en

Röder, M., Both, A., & Hinneburg, A. (2015). *Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on web search and data miningWSDM '15*New York, NY, USA: ACM399–408. https://doi.org/10.1145/2684822.2685324.

Ruan, Y., Fuhry, D., & Parthasarathy, S. (2013). *Efficient community detection in large networks using content and links. Proceedings of the 22nd international conference on world wide web*. ACM1089–1098.

Sachan, M., Contractor, D., Faruquie, T. A., & Subramaniam, L. V. (2012). *Using content and interactions for discovering communities in social networks. Proceedings of the 21st international conference on world wide web*. ACM331–340.

Singh, K., Shakya, H. K., & Biswas, B. (2016). Clustering of people in social network based on textual similarity. *Perspectives in Science, 8*, 570–573. https://doi.org/10.1016/j.pisc.2016.06.023 Recent Trends in Engineering and Material Sciences

Singh, S. S., Kumar, A., Singh, K., & Biswas, B. (Kumar, Singh, Biswas, 2019a). C2im: Community based context-aware influence maximization in social networks. *Physica A: Statistical Mechanics and its Applications, 514*, 796–818. https://doi.org/10.1016/j.physa.2018.09.142.

Singh, S. S., Singh, K., Kumar, A., & Biswas, B. (Singh, Kumar, Biswas, 2019b). Coim: Community-based influence maximization in social networks. In A. K. Luhach, D. Singh, P.-A. Hsiung, K. B. G. Hawari, P. Lingras, & P. K. Singh (Eds.). *Advanced informatics for computing research* (pp. 440–453). Singapore: Springer Singapore.

Yang, B., & Manandhar, S. (2014). *Community discovery using social links and author-based sentiment topics. 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014)*. IEEE580–587.

Zhou, D., Manavoglu, E., Li, J., Giles, C. L., & Zha, H. (2006). *Probabilistic models for discovering e-communities. Proceedings of the 15th international conference on world wide web*. ACM173–182.