ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-1119

Unsupervised semantic clustering of Twitter hashtags

Carlos Vicient and **Antonio Moreno**[1]

Abstract Social micro-blogging networks such as Twitter provide an enormous amount of information, and their automated and unsupervised analysis constitutes an exciting research challenge in Artificial Intelligence. This work presents a novel methodology, based on a semantic clustering of the set of hashtags, which permits to obtain automatically the topics associated to a given set of tweets. A case study on the field of Oncology shows how the main topics of interest are successfully discovered.

1 INTRODUCTION

Micro-blogging services such as Twitter constitute one of the most successful kinds of applications in the current Social Web. Every day more than 500 million tweets are sent, providing up to date information about any imaginable domain of knowledge. Each tweet is a string of up to 140 characters that usually contains text, links and hashtags (strings preceded by the # symbol with which the user tags his/her message). An important research area is the design and development of tools that allow users to analyse large unstructured repositories of user-tagged data in order to discover and extract meaningful knowledge from them.

In this work we have focused on the problem of clustering English hashtags that refer to the same topic, which is a first step to classify tweets and help to solve the problems of data visualisation, semantic information retrieval, information extraction, detection of users with similar interests, etc. Classifying freely chosen hashtags automatically in an unsupervised way is a very complex task [1]. Previous works on automated hashtag clustering (e.g. [2], [3]) mostly consider their co-occurrence or the co-occurrence of the words in the tweets containing the hashtags. Some authors (e.g. [4], [5]) have tried to classify tweets, usually employing a bag-of-words model to represent them and also using the co-occurrence between words as a similarity measure between tweets. Most works on Twitter topic detection try to classify a tweet into a general pre-defined small set of categories (e.g. [6]). The lack of a semantic treatment of the content of the tweets, including the hashtags, is the main shortcoming of all these approaches.

The contribution of this paper is twofold: on the one hand, we propose a methodology to perform an unsupervised semantic classification of a set of hashtags; on the other hand, we describe how to analyse the hierarchical classification in order to identify the classes that are really significant.

The rest of the paper is structured as follows. Section 2 explains the novel methodology of analysis, which is applied in section 3 to a corpus of tweets related to Oncology, in which encouraging results have been obtained. The final section summarizes the work and sketches future lines of work.

[1] Computer Science and Mathematics department, Universitat Rovira I Virgili, Tarragona, Catalonia (Spain) email: carlos.vicient@urv.cat

2 METHODOLOGY

Given a set of tweets, we extract the hashtags they contain. Word-breaking techniques are applied to split those that are composed by more than one word. Then, the three steps of the analysis are applied: *semantic annotation* (section 2.1), *semantic clustering* (section 2.2) and *class selection* (section 2.3).

2.1 Semantic annotation of hashtags

This stage aims to discover the link between hashtags and their meanings (in our case, WordNet concepts) in order to be able to compare later pairs of hashtags at the conceptual level using semantic similarity measures. The set of WordNet concepts potentially associated to each hashtag is calculated as follows. If the hashtag matches directly a WordNet concept, then there is a single candidate. If the hashtag is not contained in WordNet (a very common situation, due to the nature of Twitter hashtags), we use Wikipedia to try to find concept candidates, as shown in the *getWikipediaCandidates* function (Figure 1). If there is an entry for the hashtag, all its associated Wikipedia categories are retrieved. A category is proposed as an annotation candidate if the main noun of its description matches with a WordNet concept. The hashtags with a final empty list of candidate concepts are removed.

```
getWikipediaCandidates (hashtag h)
 wikiCandidates := ∅
 if existsWikiEntry(h)
  auxCategories:= getCategoriesFromWiki(h)
  forall cat in auxCategories
   mainNoun := getMainNoun(cat)
   auxCat := getWordNetConcept(mainNoun)
   if auxCat != ∅
    wikiCandidates = wikiCandidates + auxCat
 return wikiCandidates
```

Figure 1. Algorithm of *getWikipediaCandidates* function

2.2 Hashtag clustering

At this point each hashtag h has an associated list of WordNet concepts LC_h. After choosing any suitable ontology-based semantic similarity measure [7], the similarity between two hashtags *h1* and *h2* is defined as the maximum similarity between one concept in LC_{h1} and another in LC_{h2}. It may be argued that the use of the maximum pairwise similarity solves, indirectly, the problem of disambiguating the correct sense of the hashtag [8]. A symmetric semantic similarity matrix between all pairs of hashtags is taken as the input of a hierarchical clustering method, which obtains as a result a classification of the hashtags in a taxonomical hierarchy.

2.3 Selection of relevant clusters

Due to the nature of social tags, the result of traditional clustering methods contains a large proportion of noise. A method to filter the results is presented in figure 2, where *HC* is the result of the clustering, *t1* is the minimum inter-cluster homogeneity required to select a class (the average semantic distance between all its elements) and *t2* is the minimum number of elements required to select a class. The *filtering* function iteratively makes horizontal cuts in the tree, from the one that provides *maxK* classes down to the one that gives *minK* classes. HC_{kc} denotes the *c*-th class when the tree is divided into *k* classes. A class is selected if it is homogeneous and large enough, and it is not a superset of a previously selected class. The main topic of each selected class is its semantic centroid, calculated on WordNet with the method described in [9].

```
filtering (HC, minK, maxK, t1, t2)
 finalClusts := ∅
 forall k in maxK .. minK
  forall c in 1 .. k
   b := inter-cluster-homogeneity(HC_kc)
   if ((b >= t1) && (|HC_kc| >= t2)
      && (∄ e in finalClusts | e ⊆ HC_kc))
       finalClusts <- HC_kc
 return finalClusts
```

Figure 2. Selection algorithm

3 CASE STUDY

A test was conducted on a set of tweets related to Oncology, which was extracted from the Symplur website[2] and is composed of 5000 tweets (Oct2012-Jan2013) containing 1086 different hashtags. 930 of them (85.6%) were annotated, and the remaining 156 hashtags (14.4%) were removed. A manual analysis of this set showed 536 (57.6%) relevant medical hashtags, which were classified in a set *A* of 16 manually labelled categories (organs, professions, medical tests, etc.) and 394 hashtags (42.4%) that were listed as either noise or unrelated to Medicine. The Wu-Palmer semantic similarity function [7] was used in the clustering step. The selection parameters were *minK*=5, *MaxK*=200, *t1*=0.70 and *t2*=10.

A final set *B* of 31 classes was obtained. We compared each class in *B* with the 16 manually defined classes (plus a 17th class containing the 394 unclassified hashtags). For each class B_i in *B* Table 1 shows the class A_j in *A* with a higher precision (number of elements in B_i that belong to A_j) and the recall with respect to that class (the proportion of elements of A_j that appear in B_i). Each row shows on the left side the identifier of B_i, its semantic centroid and its number of elements, and on the right side the A_j class with a best match, with its associated precision and recall (the best results are shown in bold face). 15 classes in *B* (with a total number of 266 hashtags) matched one of the classes in *A*, whereas the other 16 matched the noisy 17th class. In two cases a couple of classes in *B* matched the same class in *A* ($\{B_{10}, B_{14}\}$ and $\{B_{13}, B_{15}\}$). Thus, 13 of the 16 target manual clusters were identified by the system with varying degrees of precision (5 classes 70-80%, 4 classes 60-64%, 4 classes 38-50%). The recall is much lower, mostly due to the subjectivity of the manual classification (2 classes over 60%, 4 classes 37-45%, 5 classes 14-24%).

[2] www.symplur.com. Last access: February 26th, 2014

Table 1. Results on Oncology tweet set

Id	Centroid	Size	Prec.	Rec.	Manual class
1	Woody_Plant	11	64%	41%	Substances
2	Day	10	50%	24%	Temporal
3	Therapy	20	**75%**	37%	Medical tests
4	Medicine	17	**76%**	23%	Medications
5	Cancer	46	**80%**	**62%**	Cancer
6	Court	14	43%	43%	Hospitals
7	Biotechnology	10	60%	15%	Biological
8	Health	23	43%	45%	Health Care
9	Medicine	43	60%	**63%**	Medical Fields
10	System	10	**70%**	21%	Body Parts
11	Area	11	**73%**	18%	Geographical Locations
12	Teaching	10	40%	14%	Academic, Research
13	Person	16	38%	12%	Medical Jobs
14	Center	10	40%	12%	Body Parts
15	Doctor	15	60%	18%	Medical Jobs
16 – 31	-	-	-	-	Noise

4 CONCLUSIONS AND FUTURE WORK

This unsupervised domain-independent methodology allows a semantic clustering of a set of hashtags and the identification of its most relevant topics, filtering the large proportion of noise inherent to these sets. The lines of future work include the analysis of the full content of the tweets, the use of general sets of millions of tweets and the study of the treatment of polysemic hashtags.

ACKNOWLEDGEMENTS

This work was partially supported by the Universitat Rovira i Virgili (scholarship of C. Vicient, 2010BRDI-06-06) and the Spanish Government in the project SHADE (TIN2012-34369).

REFERENCES

[1] K.Bontcheva and D.Rout, Making sense of social media streams through semantics: a survey. *Semantic Web Journal*, (2012).

[2] X.Wang, F.Wei, X.Liu, M.Zhou and M.Zhang, *Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach*. Procs. of the 20th ACM Conference on Information and Knowledge Management, 1031–1040, Glasgow, Scotland (2011).

[3] O.Tsur, A.Littman and A.Rappoport, *Scalable multi stage clustering of tagged micro-messages*. Poster at the 21st Int. World Wide Web Conference, Lyon, France (2012).

[4] S.Bhulai, P.Kampstra, L.Kooiman, G.Koole, M.Deurloo and B. Kok, *Trend visualization in Twitter: what's hot and what's not?* Procs. of the 1st Int. Conference on Data Analytics, 43–48. Barcelona (2012).

[5] P.Teufl and S.Kraxberger, *Extracting semantic knowledge from Twitter*. Procs. of the 3rd IFIP WG 8.5 Int. Conference on Electronic Participation, 48–59. Springer-Verlag, Berlin, Heidelberg (2011).

[6] K.Dela Rosa, R.Shah, B.Lin, A.Gershman and R. Frederking, *Topical clustering of tweets*. Procs. of the 3rd Workshop on Social Web Search and Mining. Beijing, China (2011).

[7] Z.Wu, M.Palmer, *Verbs semantics and lexical selection*. Procs. of the 32nd annual meeting on Association for Computational Linguistics, 133–138. Stroudsburg, PA, USA (1994)

[8] A.Tversky, *Features of similarity*. Psyc.Review. 84, 327–352 (1977).

[9] S.Martínez, A.Valls, D.Sánchez, *Semantically-grounded construction of centroids for datasets with textual attributes*. Knowledge-Based Systems. 35, 160–172 (2012).