A Study on the Influence of Semantics on the Analysis of Micro-blog Tags in the Medical Domain

Carlos Vicient, Antonio Moreno

A study on the influence of semantics on the analysis of micro-blog tags in the medical domain

Carlos Vicient, Antonio Moreno

Department of Computer Science and Mathematics. Universitat Rovira i Virgili.
Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA) research group.
Av. Països Catalans, 26. 43007. Tarragona, Catalonia (Spain)
{carlos.vicient@urv.cat, antonio.moreno@urv.cat}

Abstract. One current research topic in Knowledge Discovery is the analysis of the information provided by users in Web 2.0 social applications. In particular, some authors have devoted their attention to the analysis of micro-blogging messages in platforms like Twitter. A common shortcoming of most of the works in this field is their focus on a purely syntactical analysis. It can be argued that a proper semantic treatment of social tags should lead to more structured, meaningful and useful results that a mere syntactic-based approach. This work reports the analysis of a case study on medical tweets, in which the results of a semantic clustering process over a set of hashtags is shown to provide much better results than a clustering based on their syntactic co-occurrence.

Keywords: Semantic similarity, tags, co-occurrence, clustering, micro blogging

1 Introduction

In the last years the World Wide Web has moved from a collection of static pages that were created by experts and read by users to a very dynamic environment in which users are not only consuming information but also providing it (the so-called Social Web or Web 2.0)[1]. Many researchers in Artificial Intelligence are already working in the next step (the Social Semantic Web or Web 3.0 [2]), in which intelligent autonomous agents should be able to understand the semantics behind all this user-generated data in order to solve automatically complex processes like open information extraction, information retrieval, question answering, automated reasoning, etc. [3–5].

In the current Social Web users generate and distribute many different kinds of data in a wide variety of formats, from pure text to pictures, audios, videos, weblogs, etc. [6]. In order to facilitate the access to these data, many Web 2.0 applications allow users to attach them some kind of keywords, usually called tags, which provide information on the main topics related to a certain item. In some limited cases these tags may belong to well-defined domain taxonomies or ontologies, providing an accurate indexing of the tagged elements. However, in most cases uses may freely add any textual keywords, giving rise to unstructured folksonomies. An important research

problem is the design and development of tools that allow users to visualise, explore and interact with these enormous unstructured repositories of user-tagged data (HCI) and to discover and extract meaningful knowledge from them (KDD)[7]. Current tools usually rely on information retrieval techniques to find relevant pieces of information about a certain topic on such a huge volume of generated data and also on clustering methods to sort and classify the relevant documents according to their similarity. In this regard, in the last years some researchers have proposed ontology-based text clustering [8] or tag similarity [9, 10] methods. This classification process allows a posterior filtering process that keeps only those sets of documents that are more related to the user's interests, which may then finally be shown to the user applying advanced HCI visualization techniques. However, despite the great number of works in the field, it can be argued that there is still a lack of semantically-based classification and visualization mechanisms on Social Web data.

Some researchers are focusing on the analysis of the information provided by users in social networks. Micro-blogging services are one of the Web 2.0 kinds of applications that are attracting more attention. The most successful one is probably Twitter, that has around 200 million regular users that generate 500 million tweets per day [11]. A tweet is a free text with a maximum length of 140 characters. Users may tag a tweet using the so-called hashtags, which are strings of characters that begin with the "#" symbol.

As will be commented in the next section, many researchers have focused on the analysis of hashtags in order to perform complex knowledge-based tasks on a corpus of tweets [12, 13]. However, most of the scientific contributions have focused almost exclusively on a purely syntactic analysis of the content of tweets, including their hashtags. For instance, some works have proposed to discover the main topics in a corpus of tweets by clustering them taking into account the syntactic co-occurrence of hashtags. In this paper we want to argue that a semantic analysis of hashtags, based on the use of well-known ontology-based semantic similarity measures, should lead to more meaningful and relevant results that a merely syntactic analysis. We have tested this research hypothesis by clustering a set of medical tweets using both syntactic co-occurrence and semantic similarity measures, obtaining a more appropriate classification of the tweets in the second case, as we expected.

The rest of the paper is structured as follows. The next section provides a brief background on related work on the analysis of tweets, showing the current predominance of syntactic methods. Section 3 describes a new mechanism that permits to make a semantic analysis of the hashtags contained in a set of tweets, by associating them to the concepts of an ontology and then employing ontology-based similarity measures to compute the relatedness between hashtags. In section 4 the case of study using a set of medical tweets is presented, explaining the analysed data set and the syntactic and semantic clustering techniques applied on them. The following section presents and comments the results of the study, and the last section closes the paper with some final comments and the presentation of future lines of work.

2 Related Work

There have been many works in which sets of tweets are analysed for different purposes (e.g. visualization of clusters of similar tweets [14, 15], recommendation of hashtags [16], sentiment analysis [17, 18], detection of events or common topics [19, 20], clustering of tweets [21, 22], etc.). However, in most cases it can be observed that the treatment of the tweet components (mainly the words contained in the body of the tweet and its hashtags) is purely syntactic. Kywe et al [16] aim to recommend hashtags that could be associated to a particular tweet, by considering those hashtags employed by similar users (those that use the same hashtags) or used in similar tweets (those that contain the same words). Doan et al. [19] analyse sets of tweets in order to track the evolution of influenza, and they propose to filter those tweets that contain certain pre-defined words as hashtags. Russell et al. [20] analyse sets of tweets related to the Energy domain, and they define a notion of "semantic similarity" between tweets based on the co-occurrence of their terms. Bhulai et al. [14] also consider the co-occurrence of words within tweets to make clusters of tweets for visualization purposes. Teufl and Kraxberger [23] represent a tweet with a set of terms (nouns, adjectives, verbs and hashtags) and use the co-occurrence between them to define a weighted graph that can be analysed to obtain the "semantic pattern" associated to each tweet. Mathiesen et al. [24] build graphs of terms, where the edges are weighted by their co-occurrence, and study the communities present in these graphs. Veltri [21] also uses the co-occurrence between words to classify tweets related to the Nanotechnology domain. The lack of a semantic treatment of the content of the tweets is the main shortcoming of all these approaches.

Some authors have focused on the analysis of the co-occurrences between the hashtags that appear on a set of tweets, but they also seem to take a purely syntactic point of view. Wang et al. [17] define a sentiment analysis method based on different mechanisms of sentiment propagation in graphs that basically take into account the co-occurrence of hashtags. Ozdikis et al. [22] cluster hashtags by considering the co-occurrence between the hashtags and the words that appear in the tweets. Pöschko [15] also considers the co-occurrence between hashtags to group together related tweets for visualization purposes. Again, all these approaches consider only the co-occurrence of the strings that define the tags, without trying to understand their meaning to make a more thorough and semantically coherent treatment.

The main aim of this paper is to show, via an illustrative case of study in the medical domain, that clustering hashtags using a proper semantic similarity measure (supported by a domain ontology) provide a much more coherent result than the use of a purely syntactic co-occurrence metric.

3 Semantic Tag Similarity

As mentioned before, one of the main characteristics of social tagging is that users can freely annotate contents without any restriction in their choice of tags. Those tags usually lack any form of explicit organization and normalization, giving rise to un-

structured folksonomies. This fact produces, as a consequence, several problems when using tags in retrieval tasks and in their classification. One of them is that different tags might have been used for the same concept (synonymy), which makes it difficult to find all items relevant for a certain concept. Another one is that the same tag can have different meanings in different contexts (polysemy). Just to name another one, a purely syntactic analysis of tags is unable to show whether two tags have some kind of relatedness (e.g. "Religious Building" is more general than "Cathedral", and "Cathedral" is more related to "Abbey" than to "Restaurant"). To overcome these problems, it is necessary to have methods that can find tags that are conceptually related. In other words, a semantic tag similarity measure is needed in order to compare two tags at the conceptual level (i.e. a metric to measure how similar is the meaning of two lexically different tags).

Another characteristic of this kind of social tagging is the fact that tags may be acronyms, named entities, a composition of different terms or even invented words. This diversity presents a challenge in the discovery of the conceptual meaning of the tag. The linkage between the term and its meaning (a concept in a background knowledge structure, for instance a domain ontology) is what we call semantic annotation, and solving this issue is required to be able to define semantic similarity measures for tags. In this work WordNet [25], a semantic lexicon for the English language that models and semantically interlinks more than 100,000 concepts referred to by means of English textual labels, is used as the reference background knowledge structure into which tags are mapped. WordNet is organised in synsets, which are useful to solve problems like synonymy. Moreover, it also introduces hierarchies of concepts which can be used to identify related concepts.

The rest of this section aims to propose a methodology to link tags with its meaning (with WordNet elements) and to calculate the semantic similarity between them. Notice that tags are not necessarily contained in WordNet. So, the basic idea is to find the mapping between tags and WordNet concepts and, after that, to apply well-known semantic measures to compare two different tags associated with WordNet concepts. Thus, the presented methodology is divided in two main parts: the semantic annotation and the calculation of the semantic similarity.

3.1 Semantic Annotation

The aim of this first stage is to analyse a tag and associate it, if possible, with a WordNet concept. In case the tag is composed of multiple words, the system generalises it by dropping sequentially the leftmost terms until a matching is found (e.g., if the tag is "Gothic Cathedral" and this expression is not found in WordNet, the system will look for "Cathedral").

Algorithm 1 shows the procedure applied to calculate the similarity between two tags. The input parameters are the two tags to be compared and the semantic measure to be used to compare two WordNet elements. The function *getCandidates*, explained later, returns a list of possible concept candidates to be matched with the tag within WordNet. This list only contains the tag itself if it matches directly with a WordNet concept. However, due to the nature of the hashtags commonly used in Twitter, there

may be many cases in which hashtags are not found directly in WordNet and a more complex analysis is needed to find out the WordNet concept that should be associated to the hashtag (in fact, in some cases it may turn out to be impossible to find an appropriate matching between a hashtag and any concept). After that, if both candidate lists are not empty, the semantic similarity between the tags is computed by the *calculateSimilarity* function, which is also described below.

```
Sem_Sim(tag1, tag2, measure){
  candTag1 := getCandidates(tag1)
  candTag2 := getCandidates(tag2)
  if (candTag1 || candTag2) == null
    return 0.0
  return calculateSimilarity(candTag1, candTag2, measure)
}
```

Algorithm 1. Calculates the semantic similarity between tags "tag1" and "tag2", considering a given "measure" of semantic similarity between WordNet concepts

In order to find possible semantic annotations for a given tag, the first step is to check whether it appears directly as a WordNet concept. If the tag is found, then the annotation is direct. However, as pointed out before, there are many different kinds of tags that do not appear in WordNet directly. To overcome this problem, the proposed methodology relies on Wikipedia as an external knowledge base. This procedure can be seen in the pseudo-code of Algorithm 2.

```
getCandidates(tag){
  wTag := getWNConcept(tag)
  candidates <- wTag
  if wTag == null
    candidates := getWikipediaCandidates(tag)
  return candidates
}
```

Algorithm 2. Given a tag, returns a lists of possible WordNet concepts with which it could be annotated

Wikipedia is a well-known free online encyclopaedia which contains more than 30 million articles that have been written collaboratively by volunteers all around the world. In all the knowledge-based tasks associated to the analysis of Natural Language it has been assumed in the last years that Wikipedia is the premier and more comprehensive repository of textual knowledge, due its enormous breadth and the quality of its collaborative-based contents [26]. Its articles are not limited to the standard <concept, definition> structure of a dictionary, but they can describe just about anything one can imagine. Thus, Wikipedia entries have a much wider scope than WordNet concepts, permitting to find for example acronyms, named entities, different lexicalizations of the same concept, etc. Moreover, Wikipedia articles are loosely classified by means of a hierarchy of Wikipedia categories. Each of them

defines the essential characteristics of a certain topic, allowing readers a mechanism to browse and quickly find sets of related pages.

In the semantic annotation mechanism, when a tag has not been found in WordNet, it is looked up on Wikipedia. If there is an entry for the tag, all the associated categories are retrieved. A category is treated as a phrase and it is proposed as an annotation candidate only if its head (the main noun of the description of the category, extracted by a natural language parser) matches with a WordNet concept. This process is shown in the high level description of the function *getWikipediaCandidates* in Algorithm 3. So, eventually, all the proposed candidates are WordNet concepts. A tag will not have any associated candidate concepts only if it is not found either in WordNet or in Wikipedia, or if the categories found in Wikipedia do not match with any concept in WordNet.

```
getWikipediaCandidates (tag){
  wikiCandidates := null
  if existsWikiEntry(tag)
    auxCategories := getCategoriesFromWiki(tag)
    forall cat ∈ auxCategories
      mainNoun := getNN(cat)
      auxCat := getWNConcept(mainNoun)
      if auxCat != null
        wikiCandidates <- auxCat
  return wikiCandidates
}
```

Algorithm 3. Used to extract and process Wikipedia categories that will work as candidates to be annotated with WordNet concepts

3.2 Semantic Similarity

At this point there is a list of WordNet concepts (obtained directly or via Wikipedia categories) associated to each of the two tags to be compared. In order to establish the degree of semantic relatedness of the tags, the similarity between all the pairs of candidates (one from each tag) is calculated. This similarity is calculated with a function that computes the semantic similarity measure between WordNet concepts, which is given as parameter by the user. This function could be any of the well-known ontology-based similarity metrics described in the literature. The final similarity between the tags is the maximum of the similarities between the associated WordNet concepts. If the tags are associated to a certain domain of knowledge (for instance, the medical domain considered in the next section), it may be argued that calculating the similarities between all the pairs of candidates and taking the maximum one solves, in an indirect way, the problem of disambiguating the correct sense of the tag. The idea is that, even if the terms that we are comparing are polysemic, each of them will have a "medical" sense, and these domain-related senses will be the ones with a higher similarity. This process is shown in the pseudo-code depicted in Algorithm 4.

```
calculateSimilarity(candidates1, candidates2, measure){
  simMax := 0
  forall cat1 ∈ candidates1
    forall cat2 ∈ candidates2
      sim := measure(cat1, cat2)
      if(sim >= sim_max)
        simMax := sim
  return simMax
}
```

Algorithm 4. Maximises the similarity between all the pairs of candidates of the given lists.

4 Case of Study: Data Set of Medical Tweets

The underlying research work concerns the study of the influence of the use of ontology-based semantic similarity measures on the analysis of a corpus of tweets. In this paper we present a preliminary study, conducted over a set of medical tweets on the cancer disease. As will be argued in the next section, the results of this study confirm the intuition on the improvement of the results with respect to the usual syntactic analysis. This section describes the obtainment and treatment of the data set, whereas the next section provides a more detailed analysis of the results of the comparative study between syntactic and semantic clustering of tags.

The first step was the extraction of a dataset of medical tweets from the Symplur Website[1]. This dataset is composed of approximately 5000 tweets that are strongly related with "cancer" (according to Symplur), dated from October 31st 2012 to January 11th 2013. After a first filtering pre-process, we obtained a set of 1086 different hashtags used in this dataset. The graph on Fig. 1 depicts a distribution of the number of hashtags per tweets, where the y-axis represents the total number of tweets that contain exactly the number of hashtags represented in the x-axis. Notice that there are 2530 tweets tagged only with one hashtag and just one tweet containing 12 distinct hashtags.

The first step of the analysis is the semantic annotation, in which hashtags are linked to WordNet concepts. 409 hashtags (37.6 %) are found directly in WordNet. When Wikipedia categories are used to support the annotation process, as described in the previous section, the system is able to annotate 815 hashtags (75.0%). Therefore, this indirect annotation route permits to double the number of annotated hashtags. However, it has to be noted that 25% of the hashtags cannot be annotated even with the help of Wikipedia. An area of future work is to make a more detailed analysis of the hashtags so that this percentage of unannotated tags can be lowered; moreover, another field of future work is the thorough analysis of the quality of the annotation, both in its direct and indirect routes.

[1] www.symplur.com. Last access: May 24th, 2013.

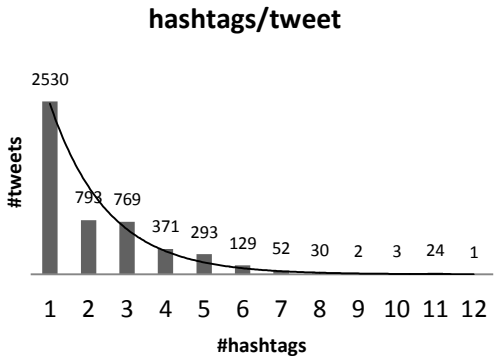

Fig. 1. Distribution of hashtags per tweet

A common way of analysing a set of tweets is to take their hashtags and partition them in a set of non-overlapping classes, so that all the hashtags in the same class (cluster) are supposed to be heavily related, whereas those belonging to different classes should be associated to different topics [15, 17, 22]. In this study we have compared two different ways of clustering the set of hashtags of the tweet corpus:

- The first clustering process is purely syntactic. The measure of similarity between two hashtags used by the clustering algorithm is based on the number of tweets in which both hashtags co-occur. The more frequently two hashtags appear, the more related they are supposed to be.
- The second clustering process is grounded on the use of a domain ontology that permits to measure the actual semantic similarity between two hashtags.

The pseudo-code of the algorithm used in this study is shown in Algorithm 5. The two input parameters of the algorithm are k (the number of classes to be generated in the clustering process) and the main topic of the extracted tweets (which in this study is "Cancer"). First, we obtain the set of tweets related to the input topic. After that, a filtering method is applied. In this step duplicates (e.g. re-tweets of the same tweet by different users) are removed, and word-breaking techniques are used in order to split hashtags composed by more than one word. The next step, described in the following subsections, is the construction of a similarity matrix between hashtags, which is used in the final step by a clustering algorithm to obtain the set of clusters of hashtags.

```
Hashtag_Clustering (k, topic){
  TW ← getTweets(topic)
  TW ← filterTweets(TW)
  Mtw ← generateSimilarityMatrix(TW)
  Ctw ← hierarchicalClustering(k, Mtw)
}
```

Algorithm 5. Algorithm used to perform the clustering of hashtags

4.1 Syntactic Clustering Based on Hashtag Co-Occurrences

In order to generate the symmetric similarity matrix used by the clustering process two steps are necessary. First, a hashtag co-occurrence matrix (as shown in Eq. 1) is constructed, where n represents the total number of hashtags, c_{ij} is the number of co-occurrences between hashtag i_{th} and hashtag j_{th} within the dataset, and c_{ii} contains the times that the hashtag i_{th} appears in the whole dataset. Notice that c_{ij} has the same value that c_{ji} , fact that implies that the order in which the hashtags appear in the set of tweets is not relevant.

$$\mathrm{C}_n = \begin{bmatrix} c_{11} & c_{12} & c_{1n} \\ \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & c_{nn} \end{bmatrix} \mid c_{ij} \in \mathbb{N} \tag{1}$$

Afterwards, the matrix has to be normalised so that all its values are between 0 and 1, since the clustering method used in this study needs either a similarity or a dissimilarity matrix as input data. To do so, each element of the matrix is normalised ($\forall i \in [1,n] \forall j \in [1,n]\ c_{ij} = NormalizedM(i,j)$) where $NormalizedM(i,j)$ is the function shown in Eq. 2. This function assigns the value 1 to each element of the diagonal of the matrix (a hashtag is equal to itself). The co-occurrence between two different hashtags is normalised by dividing it by the minimum number of their individual appearances. The rationale of this approach is that two hashtags will be considered highly similar if, in most of the tweets in which one of them appears, the other hashtag also appears. For instance, if hashtag A appears in 100 tweets, hashtag B appears in 500 tweets, and they appear together in 80 tweets, their similarity will be very high (80/100=0.8), since B appears in 80% of the tweets containing A. Other more restrictive ways of normalising the matrix values, considering for instance the maximum or the average of the individual appearances, could have also been considered.

$$\mathrm{NormalisedM(i,j)} = \begin{cases} 1, & \mathrm{i} = \mathrm{j} \\ \frac{\mathrm{c_{ij}}}{\mathrm{MIN(c_{ii}, c_{jj})}}, & \text{otherwise} \end{cases} \tag{2}$$

In Fig. 1 it was shown that more than half of the analysed tweets only contain 1 hashtag, and only around 18% of them have 4 or more hashtags. Therefore, the number of co-occurrences between hashtags is relatively small; moreover, the pairs of co-occurrent hashtags are likely to appear in different tweets. Therefore, there are many cells in the co-occurrence matrix with a 0 value (near to 86%), for all those pairs of hashtags that do not appear together in any of the input tweets (i.e. only about 14% of pairs of hashtags co-occur). With this co-occurrence based similarity measure, these pairs of hashtags are considered to be totally dissimilar.

4.2 Clustering Based on Semantic Similarities

In this section the aim is the same than in the previous section, the construction of a normalised similarity matrix between hashtags. However, in this case we want to define an ontology-based semantic similarity matrix. Let S_n (Eq. 3) be the mentioned matrix, where n represents the total number of hashtags and s_{ij} is the semantic similarity between hashtag i_{th} and hashtag j_{th} , calculated with the SEM_{sim} function (Algorithm 1), so that $\forall i \in [1,n] \forall j \in [1,n]\ c_{ij} = SEM_{sim}(i,j,measure)$.

$$S_n = \begin{bmatrix} s_{11} & s_{12} & s_{1n} \\ \cdots & \cdots & \cdots \\ s_{n1} & s_{n2} & s_{nn} \end{bmatrix} \mid s_{ij} \in [0,1] \quad (3)$$

In the experiment reported in this section Wu and Palmer's similarity measure [27] has been used to estimate the semantic alikeness between words by mapping them to WordNet concepts and computing the number of semantic links separating them. This measure gives a value between 0 and 1, so it does not require any further normalisation. As a result, terms are classified according to their semantic similarity.

It is worth noting that the S_n matrix does not have as many null values as the normalised C_n matrix. The reason is that C_n relies on direct co-occurrences among hashtags, and most of them do not tend to co-occur at all, whereas S_n is constructed with the results of the SEM_{sim} function, which is applied to WordNet elements. Thus, two very different hashtags will probably have a very low semantic similarity, but not a null one.

5 Analysis of the Results

The final part of the study consists in clustering the set of 815 annotated hashtags in two different ways, considering the C_n and S_n normalised similarity matrixes, and comparing the results. The intuition is that the semantic clustering should be more meaningful and better structured than the one based on syntactic co-occurrence. In this case, the parameter k has been fixed to 99 clusters (to obtain clusters with an average low number of elements, around 8, which can be easily managed and analysed) and the resultant clusters have been compared as follows:

1. *Centroids* are key components in many data analysis algorithms such as clustering. They basically represent a central value that minimises the distance to all the objects in the cluster. The centroid of each cluster has been calculated with the ontology-based semantically-grounded methodology presented in Martínez et al [28]. In this approach, first a set of centroid candidates is constructed, taking into account all the concepts associated to the input cluster of hashtags according to the background knowledge (i.e. WordNet). In a second step, the final centroid of each cluster is calculated as the concept candidate that minimises the average semantic distance to all concepts in the set.
2. In order to determine the internal *compactness* of each cluster (how similar they elements are) we have computed in this study the average distance to the centroid of

all the elements of the cluster which are linked to WordNet concepts. Notice that, in order to calculate this distance, any well-known ontology-based semantic similarity measure like path length, Wu-Palmer, etc. [27, 29] could have been used, since all these linked elements are in the semantic level defined by the WordNet taxonomy. Budanitsky et al [30] reported a list of those measures and evaluated their performance. In this study, as mentioned before, we have used the Wu-Palmer semantic similarity measure [27]. This measure is a path length-based measure that has the advantage of being independent of corpus statistics, uninfluenced by sparse data and easy to implement.

3. The average distance of all the elements of each cluster with respect to its centroid is calculated for both approaches (co-occurrence and semantic) and the results are compared. The lower the distance, the better the cluster (i.e. the inter-homogeneity of the cluster is stronger for lower distances).

Table 1. Average distances of clusters

		% cluster elements found in WordNet							
		10%		25%		33,3%		50%	
		Cn	Sn	Cn	Sn	Cn	Sn	Cn	Sn
match > 1	%	7,1%	**52,5%**	7,1%	**51,5%**	7,1%	**49,5%**	7,1%	**39,4%**
	avg	0,25	**0,20**	0,25	**0,20**	0,25	**0,19**	0,25	**0,18**
match > 2	%	3,0%	**35,4%**	3,0%	**35,4%**	3,0%	**33,3%**	3,0%	**27,3%**
	avg	0,31	**0,21**	0,31	**0,21**	0,31	**0,21**	0,31	**0,21**
match > 3	%	2,0%	**26,3%**	2,0%	**26,3%**	2,0%	**25,3%**	2,0%	**23,2%**
	avg	0,37	**0,23**	0,37	**0,23**	0,37	**0,23**	0,37	**0,23**

Table 1 shows the results after applying the hierarchical clustering "hclust"[2] algorithm provided by the R library separately on the C_n and S_n matrixes. Columns represent the percentage of elements of the clusters that have a direct match with WordNet concepts (e.g. the first column indicates the percentage of clusters for which at least 10% of their elements have been identified as WordNet concepts). Rows correspond to the minimum number of elements of the cluster which have been directly linked to WordNet concepts (for instance, the last row refers to those clusters that have more than 3 elements which have been identified as WordNet concepts). The combination of a row and a column restricts the possible size of a cluster. If there are *N* elements in a cluster which have been found in WordNet, and the minimum required percentage of cluster elements in WordNet is *P* (given as a number between 0 and 100), the number of elements of the cluster must be between *N* and 100**N*/*P*.

If the table results are analysed, it can be observed that in all cases the percentage of clusters that meet the requisites specified above is much bigger for the semantic analysis and, moreover, their average distances are always smaller, which means that the clusters are more compact (i.e. the inter-group homogeneity is stronger). On the other hand, the fact that the percentage of clusters obtained from C_n drops quickly to

[2] http://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html. Last access: May 24th, 2013.

very low levels (2-3%) when the minimum number of WordNet matchings is increased is due to the fact that the majority of the clusters obtained from this syntactic co-occurrence matrix are very small (from 1 to 3 elements). On the contrary, the size of the clusters obtained from the semantic similarity matrix S_n is much more homogeneous, and 23-26% of the clusters still have more than 3 elements identified as WordNet concepts. Thus, the clusters obtained in this case are much more semantically meaningful and useful than those obtained in the first case, in which there is basically a very big cluster accompanied by a large number of very small and irrelevant clusters. This fact is also shown in Fig**.** 2 which depicts the distribution of elements per cluster in a logarithmic scale.

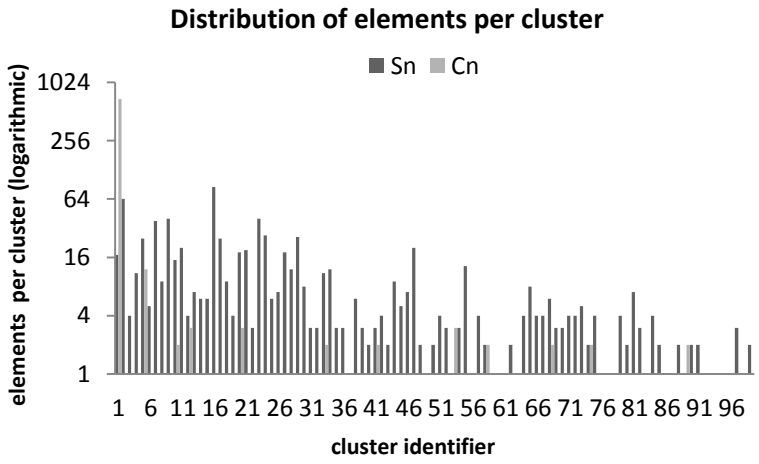

Fig. 2. Distribution of the elements per cluster in a logarithmic scale

6 Conclusions and Future Work

This study has analysed two different ways of clustering the set of hashtags that appear in a given corpus of medical tweets. A classification is based on the syntactic co-occurrence of the hashtags, whereas the other one focuses on the semantic similarity between the WordNet concepts associated to the hashtags (either directly or with the support of Wikipedia categories). The results of the case study reported in this paper seem to support the initial intuition that the common syntactic-based analysis of social tags should be replaced by more complex semantically-based treatments that provide a better structure to the knowledge extracted from Web 2.0 social applications.

There are several lines of work that can be pursued in the future. It is possible to think of more complex ways in which hashtags (especially those composed of multiple words) may be linked to WordNet concepts. The content of a tweet could be used to disambiguate the sense of the hashtag that it contains, if the hashtag corresponds to a WordNet synset with several senses (now this disambiguation is made in a more

implicit way, with the pairwise comparison of all the candidate concepts associated to the two hashtags to be compared). Concerning the similarity matrix based on co-occurrences, it could be possible to normalise it using other functions rather than the minimum (recall Eq. 2). It would also be interesting to explore similarity measures based on second-order co-occurrences (two hashtags A and B could not appear together very often in a set of tweets, but they could separately co-occur quite often with another hashtag C).

Acknowledgments

This work was partially supported by the Universitat Rovira i Virgili (pre-doctoral grant of C. Vicient, 2010BRDI-06-06) and the Spanish Government through the project DAMASK-Data Mining Algorithms with Semantic Knowledge (TIN2009-11005).

References

1. O'Reilly, T.: What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software, (2005).
2. Berners-Lee, T., Hendler, J.: The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American. 284, 34–43 (2001).
3. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Commun. ACM. 51, 68–74 (2008).
4. Fensel, D., Bussler, C., Ding, Y., Kartseva, V., Klein, M., Korotkiy, M., Omelayenko, B., Siebes, R.: Semantic web application areas. In Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB) (2002).
5. Brill, E.: Processing natural language without natural language processing. Proceedings of the 4th international conference on Computational linguistics and intelligent text processing. pp. 360–369. Springer-Verlag, Berlin, Heidelberg (2003).
6. Holzinger, A., Kickmeier-Rust, M.D., Ebner, M.: Interactive technology for enhancing distributed learning: a study on weblogs. Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology. pp. 309–312. British Computer Society, Swinton, UK, UK (2009).
7. Holzinger, A.: On Knowledge Discovery and Interactive Intelligent Visualization of Biomedical Data - Challenges in Human-Computer Interaction & Biomedical Informatics. In: Helfert, M., Francalanci, C., and Filipe, J. (eds.) DATA. SciTePress (2012).
8. Hotho, A., Staab, S., Stumme, G.: Wordnet improves Text Document Clustering. In Proc. of the SIGIR 2003 Semantic Web Workshop. pp. 541–544 (2003).
9. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems. In: Baumeister, J. and Atzmüller, M. (eds.) LWA. pp. 18–26. Department of Computer Science, University of Würzburg, Germany (2008).
10. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. Proceedings of the 7th International Conference on The Semantic Web. pp. 615–631. Springer-Verlag, Berlin, Heidelberg (2008).

11. Hold, R.: Twitter in numbers. The Telegraph, http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html, (2013).
12. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. Presented at the (2011).
13. Weng, J., Lim, E.-P., Jiang, J., He, Q.: TwitterRank: finding topic-sensitive influential twitterers. Proceedings of the third ACM international conference on Web search and data mining. pp. 261–270. ACM, New York, NY, USA (2010).
14. Bhulai, S., Kampstra, P., Kooiman, L., Koole, G., Deurloo, M., Kok, B.: Trend visualization in Twitter: what's hot and what's not? DATA ANALYTICS 2012, The First International Conference on Data Analytics,. pp. 43–48. IARIA, Barcelona (2012).
15. Pöschko, J.: Exploring Twitter Hashtags. The Computing Research Repository (CoRR). (2011).
16. Kywe, S.M., Hoang, T.-A., Lim, E.-P., Zhu, F.: On recommending hashtags in twitter networks. Proceedings of the 4th international conference on Social Informatics. pp. 337–350. Springer-Verlag, Berlin, Heidelberg (2012).
17. Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M.: Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. CIKM'11. pp. 1031–1040 (2011).
18. Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., Holzinger, A.: Opinion Mining on the Web 2.0 – Characteristics of User Generated Content and Their Impacts. Lecture Notes in Computer Science LNCS 7947. pp. 35–46 (2013).
19. Doan, S., Ohno-Machado, L., Collier, N.: Enhancing Twitter Data Analysis with Simple Semantic Filtering: Example in Tracking Influenza-Like Illnesses. Healthcare Informatics, Imaging and Systems Biology (HISB). pp. 62–71. IEEE Computer Society (2012).
20. Russell, M.G., Flora, J., Strohmaier, M., Poschko, J., Rubens, N.: Semantic Analysis of Energy-Related Conversations in Social Media: A Twitter Case Study. International Conference of Persuasive Technology (Persuasive 2011). , Columbus, OH, USA (2011).
21. Veltri, G.A.: Microblogging and nanotweets: Nanotechnology on Twitter. Public Understanding of Science. (2012).
22. Özdikiş, Ö., Şenkul, P., Oguztüzün, H.: Semantic expansion of hashtags for enhanced event detection in Twitter. The First International Workshop on Online Social Systems (WOSS) (2012).
23. Teufl, P., Kraxberger, S.: Extracting semantic knowledge from twitter. Proceedings of the Third IFIP WG 8.5 international conference on Electronic participation. pp. 48–59. Springer-Verlag, Berlin, Heidelberg (2011).
24. Mathiesen, J., Yde, P., Jensen, M.H.: Modular networks of word correlations on Twitter. Sci. Rep. 2, (2012).
25. Fellbaum, C. ed: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). MIT Press (1998).
26. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet. Web Semantics. 6, 203–217 (2008).
27. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics. pp. 133–138. Association for Computational Linguistics, Stroudsburg, PA, USA (1994).
28. Martínez, S., Valls, A., Sánchez, D.: Semantically-grounded construction of centroids for datasets with textual attributes. Knowledge-Based Systems. 35, 160–172 (2012).
29. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics. 19, 17–30 (1989).
30. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Comput. Linguist. 32, 13–47 (2006).