Making sense of Twitter

David Laniado[1] and Peter Mika[2]

[1] DEI, Politecnico di Milano
Via Ponzio 34/5, 20133 Milan, Italy
david.laniado@elet.polimi.it
[2] Yahoo! Research
Diagonal 177, 08018 Barcelona, Spain
pmika@yahoo.inc.com

Abstract. Twitter enjoys enormous popularity as a micro-blogging service largely due to its simplicity. On the downside, there is little organization to the Twitterverse and making sense of the stream of messages passing through the system has become a significant challenge for everyone involved. As a solution, Twitter users have adopted the convention of adding a hash at the beginning of a word to turn it into a *hashtag*. Hashtags have become the means in Twitter to create threads of conversation and to build communities around particular interests.
In this paper, we take a first look at whether hashtags behave as strong identifiers, and thus whether they could serve as identifiers for the Semantic Web. We introduce some metrics that can help identify hashtags that show the desirable characteristics of strong identifiers. We look at the various ways in which hashtags are used, and show through evaluation that our metrics can be applied to detect hashtags that represent real world entities.

1 Introduction

Twitter, a service for publishing short messages has been growing nearly exponentially in the past two years. Twitter handled over 600 messages every second by January, 2010[3], and has become a cultural phenomenon in many parts of the world. This success can be attributed in a large part to the simplicity of system, and the resulting cleanliness of its web site and its APIs. The ease of publishing also means that Twitter inspires timely contributions and has become an important source of information for late-breaking news, and it is already being exploited by major search engines. While appealing to publishers, the simplicity of Twitter has its downsides for anyone consuming and processing Twitter data, especially when it comes to aggregating messages. Aggregation is a necessary first step for many applications of Twitter mining, including news and trend detection, brand management and customer service, and it is also a crucial first step in separating personal communications from public discussions.

[3] http://blog.twitter.com/2010/02/measuring-tweets.html

Within the current system, however, the aggregation functions are limited to filtering tweets by users or restricting by keywords. Even in the latter case, tweets are organized by time, and not by relevance as is common for search engines. Without formal organization, aggregating tweets that belong to the same conversation or discuss the same topic is daunting. Table 1 shows ten consecutive messages retrieved for the keyword *banana*. These messages are not only posted in different languages, but are part of different ongoing conversations and refer to very different topics (the plant, a chain store, a dance, a club, and others). Keyword search is not only imprecise in aggregation, but is also missing out on a number of messages that do not contain the particular keyword. As Twitter messages are unusually short, keyword search is likely to fail in recall. As an example, during a January, 2010 earthquake in the San Francisco Bay Area, search engines have been criticized in showing only tweets that explicitly mentioned the word *earthquake*. A second, related problem is separating personal communication and news publishing, the two main cases of Twitter usage [12]. This is a crucial function for aggregators that are interested only in the conversations that concern topics of broader interests such as news or current events.

As a community solution to these problems, Twitter users have adopted the convention of adding a hash at the beginning of a word to turn it into a *hashtag*. Hashtags are meant to be identifiers for discussions that revolve around the same topic. By including hashtags in a message, users indicate to which conversations their message is related to. When used appropriately, searching on these hashtags would return messages that belong to the same conversation (even if they don't contain the same keywords), and thereby solving the aggregation problem. Coincidentally, this is the same function that strong identifiers (URIs) play in the Semantic Web. The questions we ask then is which hashtags behave as strong identifiers (if any), and could they be mapped to concept identifiers in the Semantic Web?

In this paper, we propose a set of metrics to measure the extent to which hashtags exhibit the desirable properties of strong identifiers. Our first contribution is thus formalizing the characteristic properties of strong identifiers in terms of usage in social media systems. We give a general description of hashtag usage according to these metrics (Section 2). Using a manually collected data set, we evaluate how well our metrics can identify those hashtags that represent named entities and concepts found in Freebase, a large and broad-coverage knowledge base (Section 3). Our contribution is in measuring the quality of hashtags as identifiers and selecting the hashtags that are candidate concept identifiers, a necessary first step in mapping hashtags to Semantic Web knowledge bases and identifying hashtags that are candidates for extending knowledge bases. We discuss related work in Section 4 and point to future work in Section 5.

2 Metrics for hashtag evaluation

There is no special support for tagging in Twitter, and new tags are simply introduced by prefixing a word with the hash sign. Hashtags may be used for

Boo368	@AvenLantz OMG I WANT A BANANA HAMMOCK XD
Endivisual	Got my dress..from banana republic..uhh im wearing dis dress once..? Thx..i dont need it to be so expensive -_-”
DevvonTerrell	World_of_Lala Fuh Sure!!RT @_RosettaStone_: Real talk DevvonTerrell grandmother needs to open up a bakery. Her Banana Pudding is on. HAHA!!
makalovesbieber	RT @bieberhechos: RT si te gusta la banana de Justin (? JAJAJA no mentira.
reidnwrite	@EDHMovement Unforgettable goes SUPER hard...he slipped like banana peels for not having you know you know on the album!
jojoserquina	Chicken Tinola with bitter melon, hot long horn and banana pepper, ginger and spices http://twitgoo.com/14sosn
Vol_Sus	RT @So_Delicious: Hot Fudge-Dipped Frozen Banana Bites wa recipe for Coconut Peanut Butter Hot Fudge Sauce! http://bit.ly/aknbRe YUM!
Markaw00	Eating a banana sandwich and watching Hero.
LauraRogers13	Mom asks me if I want a banana and I start doing the banana dance...I’ve been at cheer too much!
MissRicaRica	RT @philthyrichFOD: @MissRiCaRiCa *PHILTHY RICH* Coming Home Party And Video Shoot July 4th @ Banana Joes 950 10th St Modesto http://twitpic.com/1oh6ji PLZ RT.

Table 1. A consecutive sequence of Twitter message for the query ’banana’.

personal categorization, but in the vast majority of cases the intention of those who introduce a new hashtag is to evolve it into a symbol that is used by a community of users interested in and discussing a particular topic. The goal of such a hashtag is to help search and aggregation of messages related to the same topic, a function that is similar to the role of (shared) URIs in the Semantic Web.

There are a number of desirable criteria that a hashtag should fulfill in this role, similar to how ‘cool URIs’ are differentiated from poor URIs. In the following, we formalize some of these characteristics.

1. **Frequency.** The hashtag is used by a community of users with some frequency. We measure frequency both in number of users and number of messages sent, and explore the correlations between the two ways of measuring frequency.
2. **Specificity.** The extent to which the usage of a hashtag deviates from the usage of the word without a hash.
3. **Consistency in usage.** The hashtag is used consistently by different users and in different messages to indicate a single topic or concept.
4. **Stability over time.** The hashtag should become a part of the persistent vocabulary of Twitter users, i.e. it should have sustained levels of usage and should have a stable meaning over a period time.

In the following, we formalize these notions based on a Vector Space Model (VSM) for hashtags.

2.1 A vector space model for hashtags

The basic model of Twitter can be represented by a set of tuples $S \subset M \times U \times \mathcal{P}(H) \times T$ where M is the set of all sequences of not more than 140 characters, U is the set of registered Twitter users, H is the set of hashtags and T is a set of discrete timestamps with a total order. The set of hashtags is the set of possible words that start with a hash. Hashtags form part of the message in the raw data, and we extract them using a regular expression `"#[a-zA-Z0-9_]+"`. The size limitation imposed on messages puts an upper bound on the potential length of hashtags, the number of possible hashtags as well as the number of hashtags that may appear in a single message.

In line with previous works on the analysis of folksonomy systems [5], we capture the semantics of the hashtags by their usage in the social media system. In particular, we will represent the meaning of hashtags using a Vector Space Model (VSM) [20]. VSMs are commonly used in information retrieval as a representation of documents, where each dimension corresponds to a term in the collection and each value measures the weight of that term for the document. In our case, we form virtual documents for each hashtag by considering all messages where the hashtag appears. We don't filter messages by language, but it would be possible to build language specific representations this way.[4]

Formally, each hashtag h_j can be represented by a vector $\mathbf{h_j} = w_{1,j}, w_{2,j}..w_{N,j}$ where $w_{i,j} \in W, N = |W|$ and W is the set of unique terms in all of M. The simplest method for assigning weight is to consider term frequencies, i.e. $w_{i,j}$ is the number of messages in which term i co-occurs with hashtag h_j. In order to account for the different levels of specificity of terms with respect to hashtags, and to reduce the importance of the most common words, we obtain a more accurate model by applying *tf-idf* normalization: $w_{i,j} = tf_{i,j} \cdot idf_i$ where $tf_{i,j} = \frac{w_{i,j}}{\sum_{i=0}^{N} w_{i,j}}$ is the relative frequency of term i with respect to hashtag h_j; $idf_i = \log \frac{|H|}{|\{\mathbf{h_j} : w_{i_j} > 0\}|}$ is inversely proportional to the logarithm of the relative number of hashtags which term i appears with. For reasons of efficiency, we set elements $w_{i,j}$ lower than a threshold k to zero. In particular, this allows efficient indexing of the vectors using inverted indices.

We also introduce a bigram language model for hashtags; to do this, we define as *bigram* each pair of consecutive terms in a message, and as $\mathbf{b_j}$ the vector of all bigrams coocurring with tag h_j, $b_{i,j}$ being the number of messages in which bigram i and tag h_j co-occur. We apply tf-idf normalization in the same way as we compute it for single word co-occurrence.

Finally, we represent hashtags on a social dimension by means of their user occurrence vector $\mathbf{u_j}$, where $u_{i,j}$ is the number of messages tweeted by user u_i and containing hashtag h_j.

[4] Based on previous experience, languages can be detected with good accuracy despite the short length of messages. The Twitter Search API also allows restricting tweets by language.

2.2 Frequency of usage

The **frequency of a hashtag** $h_j \in H$ in terms of the number of users and messages can be defined as

$$F_u(h_j) = |\{u : \exists (m, u, H_i, t) \in S \wedge h_j \in H_i\}| \quad (1)$$

$$F_m(h_j) = |\{m : \exists (m, u, H_i, t) \in S \wedge h_j \in H_i\}| \quad (2)$$

where H_i is the set of tags used in message i.

2.3 Specificity

While in most tagging systems tags are added as external metadata to describe the content, in Twitter tags are just words making part of the message, highlighted by means of a hash to assign them a special function. Often, the hash is added as a form of emphasis (e.g.: "I'm so #happy!"), and the user may not be aware that the word as a hashtag has a more specific or otherwise different meaning than the word itself. A hashtag can often just refer to the meaning of the corresponding word, but in some cases it can assume a very different usage. For example, the hashtag "#milan" seems to be prominently used to refer to the Italian town, while the word "Milan" is much more frequently used in the context of the football team.

It is thus interesting to observe if a hashtag has a meaning close to the one of the corresponding word without hash, that we will call a *non-tag*. As with URIs on the Semantic Web, we assume that hashtags that closely match the meaning of the corresponding non-tag will be used more frequently. On the other hand, we also expect that words that are used mostly as hashtags, or hashtags that are used with a different semantics than their non-tag, will be used more consistently, because they are re-used intentionally.

Similarly to our previous definitions, we define $\mathbf{n_j}$ as the term vector of the non-tag n_j derived from h_j by removing the hash. When building the term vector $\mathbf{n_j}$, we only consider non-tag n_j occurring in a message when the corresponding hashtag h_j is not used inside the same message. The intuition is that when a non-tag appears in a message where the corresponding hashtag has already been used, the semantics of the two are probably the same. We apply tf-idf normalization to non-tags analogously to the one described in Section 2.1 for hashtags.

We compute the **specificity of a hashtag** as the similarity between the vectorial representation of the hashtag and the corresponding non-tag. For computing similarity, we use the well-known cosine similarity of the two co-occurrence vectors [21].

$$wsim(h_j, n_j) = \frac{\mathbf{h_j} \cdot \mathbf{n_j}}{\|\mathbf{h_j}\| \, \|\mathbf{n_j}\|} \quad (3)$$

Analogously, we define $\mathbf{\bar{u}_j}$ as the model of the users of the non-tag u_j, where $\bar{u}_{i,j}$ is the number of messages in which user i used non-tag n_j. We measure *social specificity* by comparing the model of the users of hashtag h_j to the model of the users of non-tag n_j:

$$usim(h_j, n_j) = \frac{\mathbf{u_j} \cdot \mathbf{\bar{u}_j}}{\|\mathbf{u_j}\| \, \|\mathbf{\bar{u}_j}\|} \tag{4}$$

To be able to compare tags and non-tags also according to frequency, we define $\bar{F}_u(n_i)$ and $\bar{F}_m(n_i)$ the frequency of a non-tag in terms of users and messages, respectively.

2.4 Consistency of usage

An important requirement for strong identifiers on the Semantic Web is that they need to be used consistently across documents and users. As a measure of the variety of usage contexts of a hashtag, we study the *entropy* of our vectorial representations of hashtags. Entropy measures the amount of uncertainty associated with the value of a random variable, in other words how uniformly the probabilities are distributed across possible values of the variable.

We define the entropy of a hashtag h_j as:

$$H(h_j) = -\sum_{i=1}^{n} p(w_{i,j}) \log p(w_{i,j}) \tag{5}$$

Higher values of entropy point to more even distributions of probabilities, corresponding to tags being used in a variety of contexts, while lower values of entropy signifies more restricted usage of a tag.

Similarly, we measure entropy of bigrams co-occurring with a tag as

$$Hb(h_j) = -\sum_{i=1}^{n} p(b_{i,j}) \log p(b_{i,j}) \tag{6}$$

Non-tag entropy is measured like tag entropy: $\bar{H}(j) = -\sum_{i=1}^{n} p(\bar{w}_{i,j}) \log p(\bar{w}_{i,j})$

2.5 Stability over time

To study the evolution of hashtags on a temporal dimension, we chose to analyze them day by day. First of all, to be able to identify new tags emerging, we define as *new* on day d a tag not appearing in the previous k days. We will define *longevity* of a new tag $l_{d,k}(h_j)$ as the number of days in which tag h_j appears at least once, over the k days after its first occurrence on day d.

We then define $\mathbf{h_j^d}$ the vector of words appearing with tag h_j in some message on day d, and we measure similarity of a hashtag h_j on day d with respect to the previous day as

$$wsim_d(h_j) = \frac{\mathbf{h_j^d} \cdot \mathbf{h_j^{d-1}}}{\left\|\mathbf{h_j^d}\right\| \left\|\mathbf{h_j^{d-1}}\right\|} \tag{7}$$

Analogously, $\mathbf{u_j^d}$ is the vector of users who used tag h_j on day d, and $usim_d(h_j)$ is the similarity among users on day d and $d-1$.

The intuition behind these measures is that a stable tag should endure over time and its meaning should not deviate much from one day to the other.

3 Evaluation

3.1 Dataset

For this study we relied on a dataset of 539,432,680 messages, collected over the whole month of November 2009 (about 18 million per day). Slightly less than 50% of tweets are in English; to filter out messages in non-latin encoding, that we are not able to parse and study, we discarded all messages containing non-ASCII characters, reducing the size of the dataset by about 28%.

Twitter user interfaces allow for forwarding of messages; the original message is so "retweeted" with a special string "rt" at the beginning. As our study is based on the co-occurrence of words inside the same message, and massive retweeting that characterizes several tags might have a strong impact biasing the results, we decided to filter out all retweets. Retweets constitute 5.4% of messages, so the actual dimension of our dataset, after filtering, is of about 369 million messages.

To compute words co-occurring with a hashtag, we filtered out from the messages all Web links and Twitter usernames (words starting with "@"). To reduce the size of co-occurrence vectors, discarding items having a very low tf-idf, we used a threshold $k = 0.01$.

3.2 Descriptive statistics

Figure 1 shows the distribution of the number of hashtags per message; overall, only 31.5 million messages, corresponding to 8.5%, have at least one hashtag. The percentage of users using at least a hashtag is higher, around 20%. Figure 2 shows that the number of users per tag follows a power low distribution, with some outlier tags used by hundreds of thousands of users. Both the distribution of the number of messages and of distinct tags tweeted by each user also follow a heavy tailed distribution, with a few extremely active users, tweeting up to 10 thousand messages or one thousand distinct tags in a month. The total number of distinct tags encountered is over 2 million; however, only about 93 thousand, corresponding to 4.14%, appeared in more than 20 messages over the whole month: for our study, we considered only these tags, and discarded all the others.

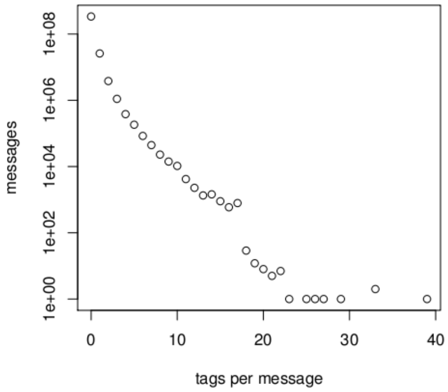

Fig. 1. Representation of the number of messages having a given number of hashtags, on a logarithmic scale.

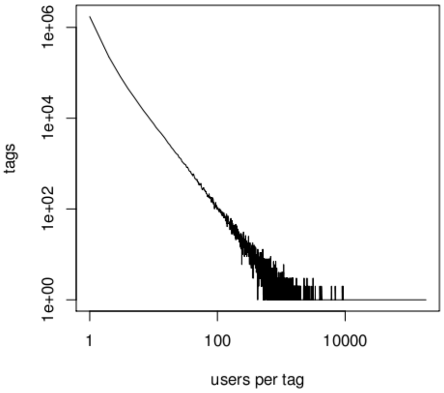

Fig. 2. Distribution of the number of users using a hashtag, on a log-log scale.

3.3 Evaluating hashtags

In this Section we will illustrate some results obtained by applying the metrics described in Section 2 to evaluate hashtags contained in our dataset.

Frequency of usage A first interesting question about hashtags is whether the corresponding non-tags also appear; about 73.5% of hashtags have the corresponding non-tag appearing at least once in our dataset. Among these, 57.8% are more frequent as hashtags than as non-tags. A "map" representing the frequency F_m of each hashtag in function of the frequency $\bar{F}_m$ of the corresponding non-tag in shown in Figure 3. The graphic exhibits a *glove* shape, which seems to point out the distinction between two kinds of tags: those corresponding to common words, that appear only sometimes preceded by a hash, and those on the "thumb", Twitter specific tags which are more often used with hash, and do usually not correspond to any commonly used word. Examples of this second kind of tags are `#tagtuesday`, `#iranelection`, `#sextips` and `#tcot` (acronym for "top conservatives on Twitter"). We obtained a very similar shape for user frequencies F_u and $\bar{F}_u$.

Specificity Figure 4 shows the similarity between tags and the corresponding non-tags, both in terms of co-occurrence vectors and of users. About a half of tags have null values of *usim*, meaning no user in common with the corresponding non-tag, while *wsim* is null for about one third of tags; while considering this second result, it must be taken into account the fact that we have cut all values of tf-idf below a threshold of 0.01.

Among tags having the highest values of *wsim* we find for example `#daylight`, almost always used in the context of "daylight savings", `#lady`, mostly referred

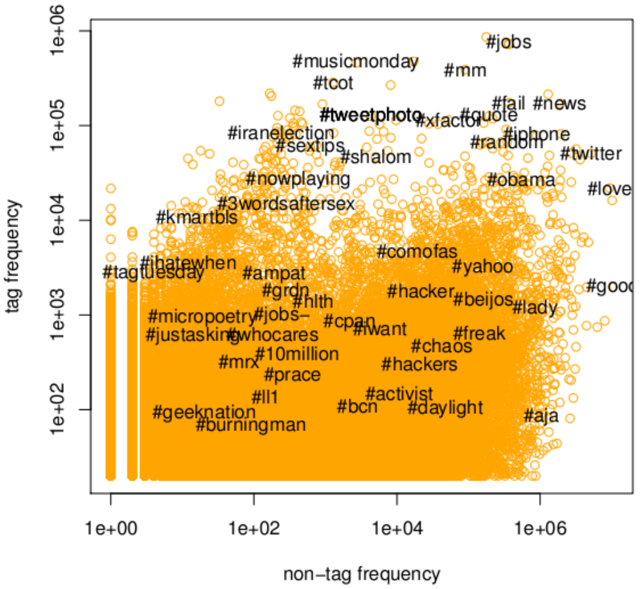

Fig. 3. Frequency of each hashtag in function of the frequency of the corresponding word with no hash.

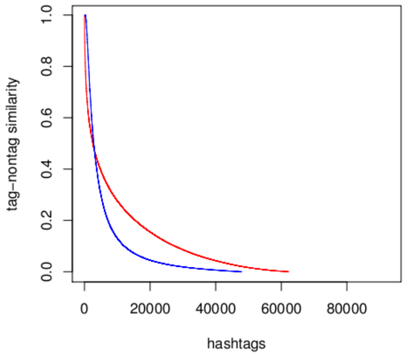

Fig. 4. Similarities *wsim* (red) and *usim* (blue), in descending order.

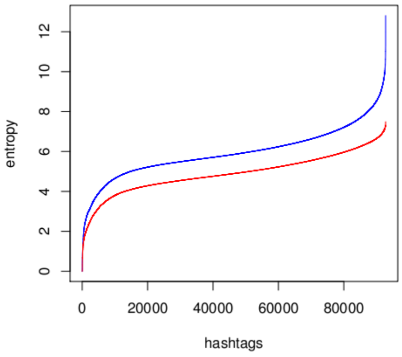

Fig. 5. Entropies *H* (red) and *Hb* (blue) of tags, in ascending order.

to the singer Lady Gaga both as a tag and as a non-tag, and `#comofaz`, which is a Portuguese slang word for "How do I do?" Among those having null or very low similarity we find tags like `#tweetphoto`, mainly found in messages generated by an application, and `#li`, that corresponds to a common word in several languages, like Portuguese, Italian and Chinese, but as a hashtag is mainly used to refer to the social network platform LinkedIn.

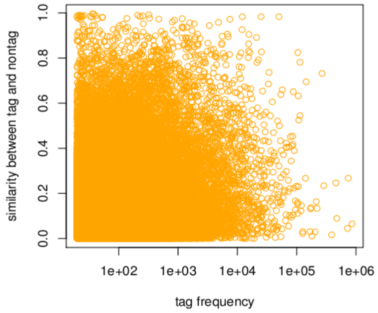

Fig. 6. Similarity between each tag and the corresponding non-tag, in function of tag frequency.

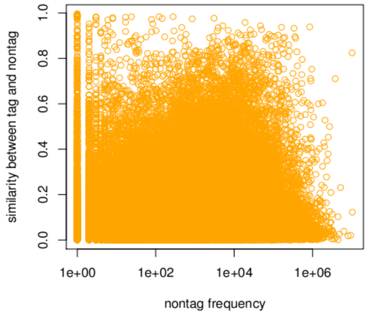

Fig. 7. Similarity between each tag and the corresponding non-tag, in function of non-tag frequency.

In Figures 6 and 7 similarity $wsim$ is plotted in function of tag and non-tag frequency, respectively. Apart from a tendency of very frequent tags to have a lower similarity, no precise relationship can be detected between $wsim$ and F_m. On the other hand, high values of similarity seem to be more likely for tags corresponding to words having a frequency in the order of a few thousands, with a peak between 1e+04 and 1e+05.

Consistency of usage In Figure 5 we plotted the entropies of tags, in descending order. Most of the tags have values of H lying in the range between 4 and 6; entropy based on bigram co-occurrence tends to be higher, with values ranging mostly between 5 and 7.

Among tags having very high entropy we find especially tags expressing sentiments, like `#whocares`, `#argh`, `#_#`, beyond some words used in a variety of contexts, like `#freak`. Tags with a very low entropy are typically generated by applications, like `#dongdongdong` (a tweeting church), `#tweetphoto` or `#iphonebabes`.

3.4 Stability over time

While until here we have studied tags as static entities for the whole period of observation, in this Section we will illustrate some results based on the observation of tags over different days.

As an example, we report some statistics observed for tags appearing on November 10th, 2009; to identify new tags we used a temporal window of $k = 9$ days. The total number of distinct hashtags observed on November 10th is over 160 thousand, about 50% of which were not appearing in any of the 9 previous days. We looked for these *new tags* in the messages from the 9 following days to evaluate their longevity l. Most of the tags have $l = 0$ and only 36 tags (about 0.045%) appear in all days until November 19th. This is an interesting indicator of how off-handedly users add hashes to words.

In this way, we have selected for each day very few new tags, that are potentially new trending topics; we can now illustrate the results obtained by applying the measures defined in Section 2.5 to two of these tags, to characterize them.

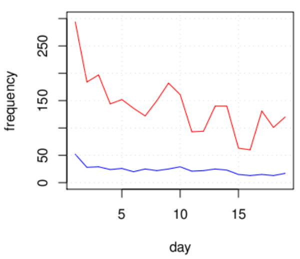

Fig. 8. Frequency F_m (red) and F_u (blue) of tag `#ampat` by day (November 12th-30th).

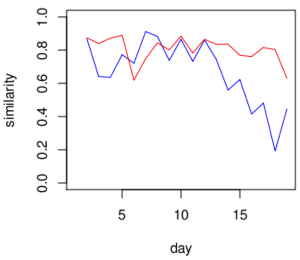

Fig. 9. Values of $wsim_d$ (red) and $usim_d$ (blue) of tag `#ampat` (November 12th-30th)

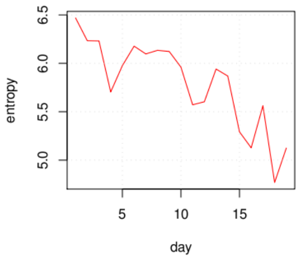

Fig. 10. Entropy of tag `#ampat` over days (November 12th-30th).

Tag `#ampat` stands for “American patriot”, and seems to have been adopted by a well defined community. Frequency of messages and users (Figure 8) exhibit a slow decreasing trend, after starting with about 300 messages in the first day, tweeted by 50 users; entropy tends to decrease in time (Figure 10) pointing out a convergence towards some context; both the meaning and the community behind the tag seem to be quite stable, though users tend to differentiate a bit in the last observed days (Figure 9).

`#kmartbls` stands for Kmart’s blue light special offers; the extremely high similarity between consecutive days in terms of co-occurrences (Figure 12), together with the very low entropy (Figure 13), is a signal of the scarce variety of information carried by the messages; these data, contrasted with the very high frequency (Figure 11), can easily bring to the conclusion that the tag has been massively promoted by some automatic application, retweeting almost identical messages from different accounts.

3.5 Manual assessment

In order to assess how well our metrics are able to indicate which hashtags represent stable concepts with a unique identity, we have performed a manual

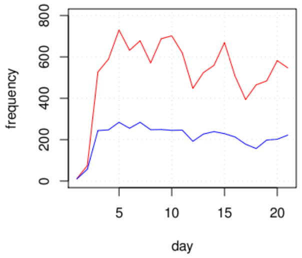

Fig. 11. Frequency F_m (red) and F_u(blue) of tag `#kmartbls` by day (November 10th-30th).

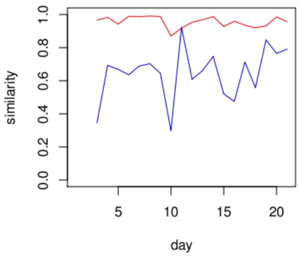

Fig. 12. Values of $wsim_d$ (red) and $usim_d$ (blue) for tag `#kmartbls` (November 10th-30th)

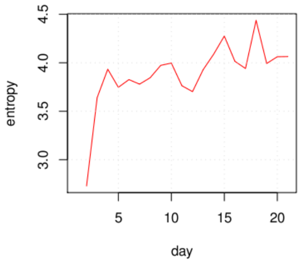

Fig. 13. Entropy of tag `#kmartbls` over days (November 10th-30th).

evaluation on a random sample of 257 hashtags, relying on 7 evaluators, experts in the field of NLP. For each tag, we collected a random sample of 100 messages with that hashtag, and asked our evaluators to answer the following questions:

1. whether they could guess the meaning of the tag just by looking at it;
2. whether the hashtag represented:
   - an event, person, organization, product, or other named entity;
   - messages generated by an application (e.g. spam);
   - messages with a common sentiment;
   - other;
   - not clear;
3. whether the tag referred to the same meaning in all messages or not.

Furthermore, the evaluators were asked to choose the closest matching concept from Freebase[5], by means of the Freebase Suggest tool[6].

In roughly 39% of cases, the messages were found to refer to a named entity; for 20% of the tags the messages were characterized by a common sentiment (e.g. `#thankfulfor`, `#grrr` or `#youknowyouareuglyif`), while 12% of the times they were recognized as generated automatically by some application (e.g. `#soundcloud`, an audio distribution platform that relies on Twitter to spread notifications about users' activities, or `#shop`, massively used by spammers). In 26% of the cases, the hashtag did not represent a named entity, a sentiment or an application, but was created for some other reason, typically to discuss a general topic (e.g. `#tv`, `#politics`, `#immigration`). The meaning of the tag remained unclear in 6.7% of the cases. Among named entities, organizations were the most common (27%), followed by products, events, persons and other entities (16%, 12%, 6%, 29%).

Slightly more than half of the tags (137) could be associated to a Freebase entry; this is higher than the number of named entities because Freebase contains

[5] `http://freebase.com`

[6] `http://code.google.com/p/freebase-suggest/`

also some general terms, like domains or common words, which are not named entities. As expected, most application and sentiment tags could not be mapped to Freebase. Only 33% of application and 14% of sentiment tags could be resolved, and many of these mappings are rough approximations of the intended meaning (e.g. the protest tag `#freegary` mapped to `gary_mckinnon`). We have also explicitly measured agreement on this task by reevaluating 31 judgments. 18 out of the 31 tags in this sample could be mapped to Freebase. The inter-annotator agreement on the task of determining if a hashtag can be mapped to Freebase is very high (Cohen's κ of 0.79). The judges agreed on the exact target in 12 out of 18 cases, and 4 of the 6 instances of disagreements were simply due to the same topic appearing in multiple hierarchies within Freebase. One of the other two cases was a close match (`technician` vs `technology` for the tag `#tech`), the other a broader match (`bacon` vs `food` for `#bacon`).

Using the whole set of judgments, we have also performed a logistic regression on the binary variable indicating whether there was a mapping to Freebase for a given hashtag. We have normalized the input variables by a linear transformation to the [0,1] interval, so that we obtain coefficients that are comparable in magnitude. Table 2 shows the coefficients of the resulting model. This model shows that tag frequency, non-tag frequency, the number of users are negatively correlated with the success of mapping to Freebase, because these frequency measures are indicators of Twitter-specific usage. Entropy is also negatively correlated, because the higher the entropy, the less consistently the tag is used. The number of non-tag users is positively correlated, because it indicates common words/sentiments. Similarities are also positively correlated, but to a smaller extent. Altogether our model achieves a 66% accuracy, a relative improvement of 25% over the baseline of choosing the majority class.

Variable	F_m	$\bar{F}_m$	F_u	$\bar{F}_u$	Hb	H	$\bar{H}$	$wsim$	$usim$	Intercept
Coefficient	-2.00	-3.45	-6.80	5.45	3.56	-3.68	0.11	0.78	0.34	-0.01

Table 2. Logistic regression coefficients of the input variables reported, for predicting output variable FBID (i.e., whether a hashtag can be mapped onto a Freebase entry).

4 Related work

After the appearance of the first social bookmarking applications, a considerable effort has been spent in the study of tag semantics. Work in this field is strongly related to ours, different in that tagging is explicit and often serves personal categorization. Classifications of tags based on their usage are proposed in [8] and [22]; an insight into the use of non subject related tags is offered in [11]. Motivations and incentives behind tagging have been investigated in [16] and [2]. In

[7] some metrics are introduced to evaluate tags, based on user behaviour. The authors of [1] evaluate the potential of folksonomies to generate semantic metadata; an assessment of delicious tag vocabulary efficiency from an information theory perspective is provided in [6]. Among the studies aiming at extracting emergent semantics from folksonomies, the work described in [24] relies on a metric of tag entropy to evaluate the ambiguity of tags.

While in our work we could represent hashtags as virtual documents, based on messages in which they appear, in traditional social tagging applications the context in which a tag can be analyzed is usually just constitued by other tags used concurrently; a tripartite model of tags, users and resources is the basis for most works [17]. In [5] some measures to compute tag relatedness are presented, and delicious tags are grounded to WordNet synsets in order to contrast semantic relations with the results of the different metrics proposed; the best semantic precision was achieved with metrics based on the cosine between each tag's context, represented as a vector of co-occurring tags. Also the study described in [4] resonates with our work for the use of information retrieval techniques to compare tags with each other. In [13] a classification of users according to their tagging behaviour is leveraged to improve the effectiveness of algorithms for emergent semantics extraction from folksonomies. The idea of integrating tags into the Semantic Web is pursued in FLOR [3], a framework for the enrichment of folksonomies with semantic information from existing ontologies. Models have been proposed to anchor tags to Semantic Web URIs, such as MOAT [19] and CommonTag[7]; NiceTag ontology allows for the representation of different kinds of tagging actions, by means of named graphs [15].

Twitter's social network and information diffusion dynamics have been studied in [10] and [12]; the authors of [14] investigate the use of Twitter during conferences, identifying classes of hashtags and finding out a prevalence of technical terms, and a general tendency to address especially people belonging to the same community. In [9] tagging behaviour in Twitter is compared with the one in delicious, and it is described as *conversational*; the authors in particular study the phenomenon of memes emerging around hashtags that are often abandoned after a short time, and introduce statistical metrics to detect them. A tripartite model of users, hashtags and messages is introduced in [23] to turn Twitter into a folksonomy, and to extract emergent semantics. Special syntaxes have been proposed to allow users express structured information inside a tweet; among these we mention twitlogic[8] and HyperTwitter[9], which allows users specify relationships among hashtags (equivalent, subtag) and express arbitrary properties; an alternative distributed platform for microblogging, based on Semantic Web principles, is described in [18].

[7] http://commontag.org

[8] http://twitlogic.fortytwo.net/

[9] http://semantictwitter.appspot.com/

5 Conclusions and future work

Since their introduction, hashtags have shown to be a popular feature of microblogging platforms as a practical solution to the problem of aggregating content in the disorganized and fragmented stream of information that characterizes these systems. However, not all hashtags are used in the same way, not all of them aggregate messages around a community or a topic, not all of them endure in time, and not all of them have an actual meaning. In this work we have addressed the issue of evaluating Twitter hashtags as strong identifiers, as a first step in order to bridge the gap between Twitter and the Semantic Web.

The first contribution of this paper stands in the formalization of the problem, and in the elaboration of a number of desired properties for a good hashtag to serve as a URI. We have proposed a Vector Space Model for hashtags, representing them as virtual documents; in parallel we have introduced the notion of *non-tag*, to be able to compare each tag with the corresponding word without hash. We have defined several metrics, based both on the messages containing a hashtag and on the community adopting it, to characterize hashtag usage on a variety of dimensions: *frequency*, *specificity*, *consistency*, and *stability* over time. We have applied these metrics to a dataset of more than half a billion messages, collected over the whole month of November 2009. Beyond qualitatively illustrating the results, showing how the metrics proposed tend to correspond to actual properties of the data, we have performed manual classification of a sample of tags. Based on these data, we have tested the results obtained with the algorithms described in the paper, showing how a combination of the proposed measures can help in the task of assessing which tags are more likely to represent valuable identifiers. These results are promising, with respect to the perspective of anchoring Twitter hashtags to Semantic Web URIs, and to detect concepts and entities valuable to be treated as new identifiers. Also spam detection tasks can benefit from the metrics we have illustrated.

This work is only a first step in the direction of the investigation of hashtag semantics, and of automatic hashtag classification. Different machine learning algorithms can be used to improve the performances; cleaner results might be obtained by taking into account the different languages of tweets. A more complete analysis may result by considering also links, usernames and emoticons, and by comprising retweet dynamics in the investigation. As a further step, we plan to study similarity between hashtags, based both on word and user co-occurrence vectors, in order to find clusters and study emergent semantics.

References

1. H. S. Al-Khalifa and H. C. Davis. Exploring the value of folksonomies for creating semantic metadata. *International Journal on Semantic Web and Information Systems*, 2007.
2. M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proc. of CHI*, 2007.

3. S. Angeletou. Semantic enrichment of folksonomy tagspaces. In *Proc. of ISWC*, 2008.
4. D. Benz, M. Grobelnik, A. Hotho, R. Jaschke, D. Mladenic, V. D. P. Servedio, S. Sizov, and M. Szomszor. Analyzing tag semantics across collaborative tagging systems. *Dagstuhl Seminar 08391 - Working Group Summary*, 2008.
5. C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In *Proc. of ISWC*, 2008.
6. E. H. Chi and T. Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *Proc. of HT*, 2008.
7. U. Farooq, T. G. Kannampallil, Y. Song, C. H. Ganoe, J. M. Carroll, and L. Giles. Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In *Proc. of GROUP*, 2007.
8. S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 2006.
9. J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in twitter. In *Proc. of HT*, 2010.
10. B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 2009.
11. M. E. Kipp. @toread and cool : Subjective, affective and associative factors in tagging. In *Proc. of CAIS/ACSI*, 2008.
12. H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proc. of WWW*, 2010.
13. C. Krner, D. Benz, M. Strohmaier, A. Hotho, and G. Stumme. Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. *Proc. of WWW*, 2010.
14. J. Letierce, A. Passant, J. Breslin, and S. Decker. Understanding how twitter is used to widely spread scientific messages. In *Proc. of WebSci*, 2010.
15. F. Limpens, A. Monnin, F. Gandon and D. Laniado. Speech acts meet tagging: NiceTag ontology. *Proc. of I-SEMANTICS*, 2010.
16. C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proc. of HT*, 2006.
17. P. Mika. Ontologies are us: A unified model of social networks and semantics. *Proc. of ISWC*, 2005.
18. A. Passant, T. Hastrup, U. Bojars, and J. Breslin. Microblogging: A semantic web and distributed approach. In *Proc. of SFSW*, 2008.
19. A. Passant and P. Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. *Proc. of LDOW*, 2008.
20. V. V. Raghavan and S. K. M. Wong. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 1986.
21. G. Salton. *Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer.* Addison–Wesley, 1989.
22. S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. tagging, communities, vocabulary, evolution. In *Proc. of CSCW*, 2006.
23. C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of SemSearch*, 2010.
24. X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *Proc. of WWW*, 2006.