Characterising Emergent Semantics in Twitter Lists

Andrés García-Silva[1], Jeon-Hyung Kang[2], Kristina Lerman[2],
and Oscar Corcho[1]

[1] Ontology Engineering Group,
Facultad de Informática, Universidad Politécnica de Madrid, Spain
{hgarcia,ocorcho}@fi.upm.es
[2] Information Sciences Institute,
University of Southern California, USA
{jeonhyuk,lerman}@isi.edu

Abstract. Twitter lists organise Twitter users into multiple, often overlapping, sets. We believe that these lists capture some form of emergent semantics, which may be useful to characterise. In this paper we describe an approach for such characterisation, which consists of deriving semantic relations between lists and users by analyzing the co-occurrence of keywords in list names. We use the vector space model and Latent Dirichlet Allocation to obtain similar keywords according to co-occurrence patterns. These results are then compared to similarity measures relying on WordNet and to existing Linked Data sets. Results show that co-occurrence of keywords based on members of the lists produce more synonyms and more correlated results to that of WordNet similarity measures.

1 Introduction

The active involvement of users in the generation of content on the Web has led to the creation of a massive amount of information resources that need to be organized so that they can be better retrieved and managed. Different strategies have been used to overcome this information overload problem, including the use of tags to annotate resources in folksonomies, and the use of lists or collections to organize them. The bottom-up nature of these user-generated classification systems, as opposed to systems maintained by a small group of experts, have made them interesting sources for acquiring knowledge. In this paper we conduct a novel analysis of the semantics of emergent relations obtained from Twitter lists, which are created by users to organize others they want to follow.

Twitter is a microbbloging platform where users can post short messages known as tweets. Twitter was started in 2006 and has experienced a continuous growth since then, currently reaching 100 million users[1]. In this social network users can follow other users so that they can receive their tweets. Twitter users

[1] http://blog.twitter.com/2011/09/one-hundred-million-voices.html

E. Simperl et al. (Eds.): ESWC 2012, LNCS 7295, pp. 530–544, 2012.

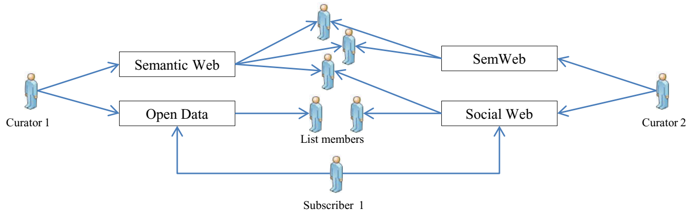

Fig. 1. Diagram showing different user roles in twitter lists. Boxes indicate list names.

are allowed to classify people into lists (see figure 1). The creator of the list is known as the curator. List names are freely chosen by the curator and consist of keywords. Users other than the curator can then subscribe to receive tweets from the listed users. Similarly to what happens with folksonomies [7,19], the classification system formed by connections between curators, subscribers, listed users, and list names, can be considered as a useful resource for knowledge extraction. In this work we analyze term co-occurrence patterns in these lists to identify semantic relations between all these elements. Co-occurrence may happen due to the simultaneous use of keywords in different lists created by curators, or in lists followed by subscribers, or in lists under which users are listed.

For instance, table 1 summarizes the lists under which an active and well known researcher in the Semantic Web field has been listed. The first column presents the most frequent keywords used by curators of these lists, while the second column shows keywords according to the number of subscribers. We can see that *semantic_ web* and *semweb* are frequently used to classify this user, which suggests a strong relationship between both keywords. In fact, these keywords can be considered as synonyms since they refer to same concept. Though less frequent, other keywords such as *semantic*, *tech* and *web_ science* are also related to this context. The other keywords according to the use given by subscribers (*e.g.*, *connections*) are more general and less informative for our purposes.

We consider that Twitter Lists represent a potentially rich source for harvesting knowledge, since they connect curators, members, subscribers and terms. In this paper we explore which of such connections lead to emergent semantics and produce most related terms. We analyze terms using the vector space model [24] and a topic modeling method, the Latent Dirichlet Allocation [5]. Then we use metrics based on the WordNet synset structure [10,26,16] to measure the semantic similarity between keywords. In addition, we ground keywords to Linked Open Data and present the relations found between them. This type of analysis lays the foundation for the design of procedures to extract knowledge from Twitter lists. For instance, ontology development can benefit of the emerging vocabulary that can be obtained from these user generated sources.

In the following we present the models used to obtain relation between keywords from Twitter lists. In section 3 we introduce the similarity metrics based on WordNet, and we describe the technique used to gather relations from linked data. Next we present, in section 4, the results of our study. Finally we describe the related work in section 5, and present the conclusions in section 6.

Table 1. Most frequent keywords found in list names where the user has been listed

Curators		Subscribers	
semantic_web	39	semantic_web	570
semweb	22	semweb	100
semantic	7	who-my-friends-talk-to	93
tech	7	connections	82
web_science	5	rock_stars	55

2 Obtaining Relations between Keywords from Lists

We use the vector space model [24] to represent list keywords and their relationships with curators, members and subscribers. Each keyword is represented by three vectors of different dimension according to the type of relation represented. The use of vectors allows calculating similarity between them using standard measures such as the angle cosine.

Twitter lists can be defined as a tuple $TL = (C, M, S, L, K, R_l, R_k)$ where C, M, S, L, and K are sets of curators, members (of lists), subscribers, list names, and keywords respectively, $R_l \subseteq C \times L \times M$ defines the relation between curators, lists names, and members, and $R_k \subseteq L \times K$ represents keywords appearing in a list name. A list ϕ is defined as $(c, l, M_{c,l})$ where $M_{c,l} = \{m \in M | (c, l, m) \in R_l\}$. A subscription to a list can be represented then by $(s, c, l, M_{c,l})$. To represent keywords we use the following vectors:

- For the use of a keyword k according to curators we define $k_{curator}$ as a vector in $\Re^{|C|}$ where entries in the vector $w_c = |\{(c, l, M_{c,l}) | (l, k) \in R_k\}|$ correspond to the number of lists created by the curator c that contain the keyword k.

- For the use of a keyword k according to members we use a vector k_{member} in $\Re^{|M|}$ where entries in the vector $w_m = |\{(c, l, m) \in R_l | (l, k) \in R_k\}|$ correspond to the number of lists containing the keyword k under which the member m has been listed.

- For the use of a keyword k according to subscribers we utilize a vector $k_{subscriber}$ in $\Re^{|S|}$ where entries in the vector $w_s = |\{(s, c, l, M_{c,l}) | (l, k) \in R_k\}|$ correspond to the number of times that s has subscribed to a list containing the keyword k.

In the vector space model we can measure the similarity between keywords calculating the cosine of the angle for the corresponding vectors in the same dimension. For two vectors k_i and k_j the similarity is $sim(k_i, k_j) = \frac{k_i \cdot k_j}{||k_i|| \cdot ||k_j||}$.

We also use Latent Dirichlet Allocation (LDA) [5] to obtain similar keywords. LDA is an unsupervised technique where documents are represented by a set of topics and each topic consists of a group of words. LDA topic model is an improvement over *bag of words* approaches including the vector space model, since LDA does not require documents to share words to be judged similar. As long as they share similar words (that appear together with same words in other documents) they will be judged similar. Thus documents are viewed as a mixture of probabilistic topics that are represented as a T dimensional random variable θ. For each document, the topic distribution θ has a Dirichlet prior $p(\theta|\alpha) \sim Dir(\alpha)$. In generative story, each document is generated by first picking a topic distribution θ from the Dirichlet prior and then use each document's topic distribution to sample latent topic variables z_i. LDA makes the assumption that each word is generated from one topic where z_i is a latent variable indicating the hidden topic assignment for word w_i. The probability of choosing a word w_i under topic z_i, $p(w_i|z_i, \beta)$, depends on different documents.

We use the bag of words model to represent documents as input for LDA. For our study keywords are documents and words are the different users according to their role in the list structure. To represent keywords we use the following sets:

- For a keyword k according to curators we use the set $k_{bagCurator} = \{c \in C|(c, l, m) \in R_l \wedge (l, k) \in R_k\}$ representing the curators that have created a list containing the keyword k.

- For a keyword k according to members we use a set $k_{bagMember} = \{m \in M|(c, l, m) \in R_l \wedge (l, k) \in R_k\}$ corresponding to the users who have been classified under lists containing the keyword k.

- For a keyword k according to subscribers we use a set $k_{bagSubscriber} = \{s \in S|(s, c, l, M_{c,l}) \wedge (l, k) \in R_k\}$, that is the set of users that follow a list containing the keyword k.

LDA is then executed for all the keywords in the same representation schema (*i.e.*, based on curators, members, or subscribers) generating a topic distribution θ for each document. We can compute similarity between two keywords k_i and k_j in the same representation schema by measuring the angle cosine of their corresponding topic distributions θ_i and θ_j.

3 Characterising Relations between Keywords

We investigate the relevance of the relations between keywords obtained from twitter lists using state of the art similarity measures based on WordNet. In addition, given the limited scope of WordNet we complement our study using knowledge bases published as linked data.

3.1 Similarity Measures Based on WordNet

To validate the relations found from keyword co-occurrence analysis in Twitter lists, we use similarity measures that tap into WordNet [10]. WordNet is a lexical database where synonyms are grouped on synsets, with each synset expressing a concept. Synsets are linked according to semantic relations that depend on the synsets part-of-speech category. Nouns and verbs are arranged in a hierarchy defined by a super-subordinate relation (is-a) known as hyperonymy. In addition, there are meronymy relations (part-of) for nouns, troponym relations (specific way of) for verbs, antonym relations for adjectives, and synonym relations for adverbs. WordNet consists of four sub nets, one for each part of speech category.

A natural measure of similarity between words is the length of the path connecting the corresponding synsets [22,16]. The shorter the path the higher the similarity. This length is usually calculated in the noun and verb is-a hierarchy according to the number of synsets in the path connecting the two words. In the case of two synonyms, both words belong to the same synset and thus the path length is 1. A path length of 2 indicates an is-a relation. For a path length of 3 there are two possibilities: (*i*) both words are under the same hypernym known as *common subsumer*, and therefore the words are siblings, and (*ii*) both words are connected through an in-between synset defining an indirect is-a relation. Starting with 4 the interpretation of the path length is harder.

However, the weakness of using path length as a similarity measure in WordNet is that it does not take into account the level of specificity of synsets in the hierarchy. For instance, *measure* and *communication* have a path length of 3 and share *abstraction* as a common subsumer. Despite low path length, this relation may not correspond to the human concept of similarity due to the high level of abstraction of the concepts involved.

Abstract synsets appear in the top of the hierarchy, while more specific ones are placed at the bottom. Thus, Wu and Palmer [26] propose a similarity measure which includes the depth of the synsets and of the least common subsumer (see equation 1). The least common subsumer *lcs* is the deepest hypernym that subsumes both synsets, and depth is the length of the path from the root to the synset. This similarity range between 0 and 1, the larger the value the greater the similarity between the terms. For terms *measure* and *communication*, both synsets have depth 4, and the depth of the *lcs abstraction* is 3; therefore, their similarity is 0.75.

$$wp(synset_1, synset_2) = 2 * depth(lcs)/(depth(synset_1) + depth(synset_2) \quad (1)$$

Jiang and Conrath [16] propose a distance measure that combines hierarchical and distributional information. Their formula includes features such as local network density (*i.e.*, children per synset), synset depth, weight according to the link type, and information content IC of synsets and of the least common subsumer. The information content of a synset is calculated as the inverse log of its probability of occurrence in the WordNet hierarchy. This probability is based on the frequency of words subsumed by the synset. As the probability of a synset increases, its information content decreases. Jiang and Conrath distance can be computed using equation 2 when only the information content is used. A shorter distance means a stronger semantic relation. The IC of *measure* and *communication* is 2.95 and 3.07 respectively while *abstraction* has a IC of 0.78, thus their semantic distance is 4.46.

$$jc(synset_1, synset_2) = IC(synset_1) + IC(synset_2) - 2 * IC(lcs) \quad (2)$$

We use, in section 4, the path length, Wu and Palmer similarity, and Jiang and Conrath distance to study the semantics of the relations extracted from Twitter lists using the vector space model and LDA.

3.2 Linked Data to Identify Relation Types

WordNet-based analysis is rather limited, since WordNet contains a small number of relations between synsets. To overcome this limitation and improve the detection of relationships, we use general purpose knowledge bases such as DBpedia [4], OpenCyc,[2] and UMBEL[3], which provide a wealth of well-defined relations between concepts and instances. DBpedia contains knowledge from Wikipedia for close to 3.5 million resources and more than 600 relations. OpenCyc is a general purpose knowledge base with nearly 500K concepts around 15K types of relations. UMBEL is an ontology with 28,000 concepts and 38 relations. These knowledge bases are published as linked data [3] in RDF and with links between them: DBpedia resources, and classes are connected to OpenCyc concepts using *owl:sameAs*, and to UMBEL concepts using *umbel#correspondsTo*.

Our aim is to bind keywords extracted from list names to semantic resources in these knowledge bases so that we can identify which kind of relations appear between them. To do so we harness the high degree of interconnection in the linked data cloud offered by DBpedia. We first ground keywords to DBpedia [12], and then we browse the linked data set for relations connecting the keywords.

After connecting keywords to DBpedia resources we query the linked data set to search for relations between pairs of resources. We use a similar approach to [14] where SPARQL queries are used to search for relations linking two resources r_s and r_t. We define the path length L as the number of objects found in the path linking r_s with r_t. For $L = 2$ we look for a $relation_i$ linking r_s with r_t. As we do not know the direction of $relation_i$, we search in both directions: 1) r_s $relation_i$ r_t, and 2) r_t $relation_i$ r_s. For $L = 3$ we look for a path containing two relationships and an intermediate resource *node* such as: r_s $relation_i$ $node$, and $node$ $relation_j$ r_t. Note that each relationship may have two directions and hence the number of possible paths is $2^2 = 4$. For $L = 4$ we have three relationship placeholders and the number of possible paths is $2^3 = 8$. In general, for a path length L we have $n = \sum_{l=2}^{L} 2^{(l-1)}$ possible paths that can be traversed by issuing the same number of SPARQL queries[4] on the linked data set.

For instance, let us find the relation between the keywords *Anthropology* and *Sociology.* First both keywords are grounded to the respective DBpedia resources, in this case *dbpr:Anthropology* and *dbpr:Sociology.* Figure 2 shows linked data relating these DBpedia resources. To retrieve this information, we pose the query shown in Listing 1.1.[5] The result is the triples making up the path between

[2] OpenCyc home page: http://sw.opencyc.org/

[3] UMBEL home page: http://www.umbel.org/

[4] Note that for large L values the queries can last long time in large data sets.

[5] Property paths, in SPARQL 1.1 specification, allow simplifying these queries.

the resources. In our case we discard the initial *owl:sameAs* relation between DBpedia and OpenCyc resources, and keep the assertion that Anthropology and Sociology are Social Sciences.

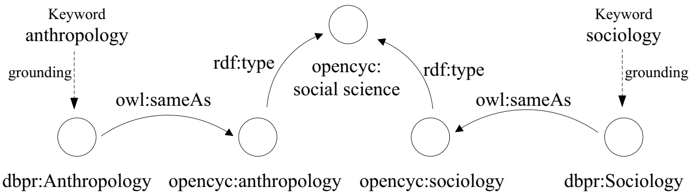

Fig. 2. Linked data showing the relation between the anthropology and sociology

```
SELECT *
WHERE{<dbpr:Anthropology> ?relation1 ?node1. ?node1 ?relation2 ?node2.
     <dbpr:Sociology> ?relation4 ?node3. ?node3 ?relation3 ?node2.}
```

Listing 1.1. SPARQL query for finding relations between two DBpedia resources

4 Experiment Description

Data Set: Twitter offers an Application Programming Interface (API) for data collection. We collected a snowball sample of users and lists as follows. Starting with two initial seed users, we collected all the lists they subscribed to or are members of. There were 260 such lists. Next, we expanded the user layer based on current lists by collecting all other users who are members of or subscribers to these lists. This yielded an additional set of 2573 users. In the next iteration, we expanded the list layers by collecting all lists that these users subscribe to or are members of. In the last step, we collected 297,521 lists under which 2,171,140 users were classified. The lists were created by 215,599 distinct curators, and 616,662 users subscribe to them[6]. From list names we extracted, by approximate matching of the names with dictionary entries, 5932 unique keywords; 55% of them were found in WordNet. The dictionary was created from article titles and redirection pages in Wikipedia.

Obtaining Relations from Lists: For each keyword we created the vectors and the bags of words for each of the three user-based representations defined in section 2. We calculated cosine similarity in the corresponding user-based vector space. We also run the LDA algorithm over the bags of words and calculated the cosine similarity between the topic distribution produced for each document. We kept the 5 most similar terms for each keyword according to the Vector-space and LDA-based similarities.

[6] The data set can be found here: http://goo.gl/vCYyD

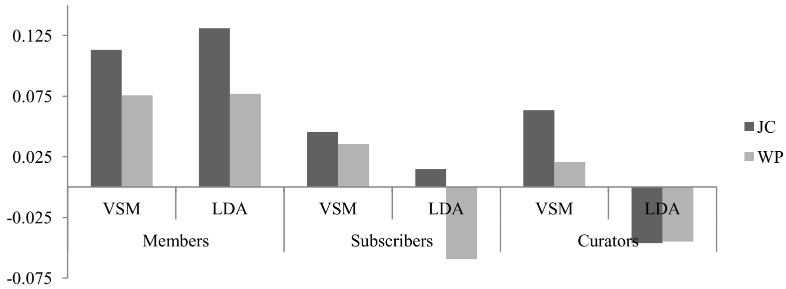

Fig. 3. Coefficient of correlation between Vector-space and LDA similarity with respect to WordNet measures

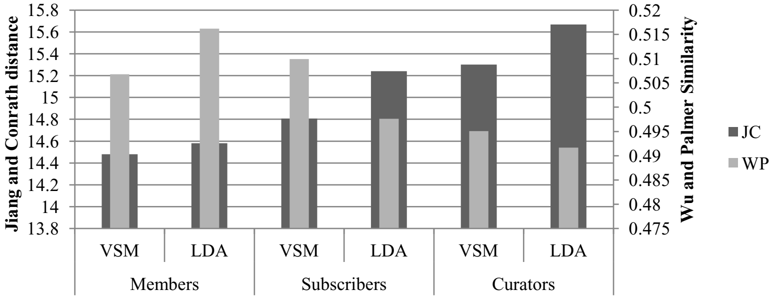

Fig. 4. Average Jiang and Conrath distance and Wu and Palmer similarity

WordNet Analysis: For each pair of similar keywords we calculated their similarity according to Jiang and Conrath (JC) and Wu and Palmer (WP) formulas. To gain an initial insight about these measures we calculate the correlation between them (see Figure 3). We use the Pearson's coefficient of correlations which divides the covariance of the two variables by the product of their standard deviations.

In general these results show that Vector-space and LDA similarity based on members produce the most similar results to that of WordNet measures. Vector-space similarity based on subscribers and curators also produces correlated results, although significantly lower. LDA similarity based on subscribers results is correlated to JC distance but not to WP similarity. Finally LDA based on curators produces results that are not correlated to WordNet similarities.

Correlation results can be partially explained by measuring the average of JC distance and WP similarity[7] (see figure 4). Vector-space and LDA similarities based on Members have the shortest JC distance, and two of the top tree WP similarity values. Vector-space similarity based on subscribers has also a short JC distance, and a high WP similarity. For the rest of similarities JC distances are longer and WP similarity lower.

[7] The averages were calculated over relations for which both terms were in WordNet.

To identify the type of relations found by Vector-space and LDA similarities we calculate, as shown in table 2, the path length of the corresponding relations in WordNet. To guarantee a base similarity, we use a threshold of 0.1; similarities under this value were discarded. Note that in WordNet different part of speech categories have distinct hierarchies and hence the path length can be calculated only for terms in the same category. According to the path length, the similarity based on members produce the highest number of synonyms (path length=1), reaching a 10.87% of the relations found in WordNet for the case of LDA similarity. In this case, the LDA model analyzes co-occurrence of groups of members across different keywords to identify related keywords. Unlike the vector space model, which requires exact members to be present in similar keywords, LDA allows synonyms, i.e., different members that tend to co-occur with the same sets of keywords, to contribute to keyword similarity.

Table 2. Path length in WordNet for similar Keywords according to Vector-space and LDA models

Path Length	Members		Subscribers		Curators	
	VSM	LDA	VSM	LDA	VSM	LDA
1	**8.58%**	**10.87%**	3.97%	3.24%	1.24%	0.50%
2	**3.42%**	**3.08%**	1.93%	0.47%	0.70%	0.00%
3	2.37%	**3.77%**	2.96%	2.06%	2.38%	**4.03%**
>3	67.61%	65.50%	67.27%	67.56%	77.83%	75.81%

Similarity based on subscribers and curators produce a significative lower number of synonyms. Likewise, similarity based on members produces the highest number of direct is-a relations (path length=2). LDA similarity based on curators produce the highest number of keywords directly related by a common superclass or an indirect is-a relation (path length=3).

Given that the majority of relations found in WordNet have a path length greater than or equal to 3, we decided to categorize them according to whether the relation is based on a common subsumer or whether it is based on linked is-a relations. In average 97.65% of the relations with a path length ≥ 3 involve a common subsumer.

As it was argued before, the depth of the least common subsumer influences the relevance of a relation. A manual inspection of the WordNet hierarchy shows that synsets being at a distance greater than or equal to 5 from the root may be considered as more specific. Figure 5 shows the percentage of relations according to the depth of the least common subsumer in the WordNet hierarchy. For a depth of the LCS greater than or equal to 5 and to 6 the Vector-space similarity based on subscribers produces the highest percentage of relations (39.19% and 20.62% for each case) followed by the Vector-space similarity based on members (37.07% and 17.96%). Starting from a depth of the LCS greater than or equal to 7 until 9 the LDA and Vector-space similarity based on members gathers the highest percentage of relations.

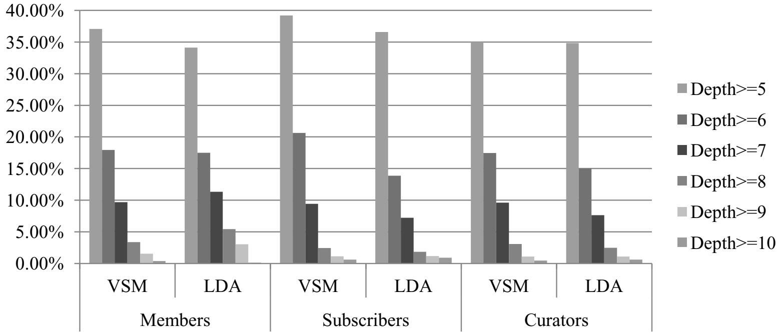

Fig. 5. Relations according to the depth of the least common subsumer LCS

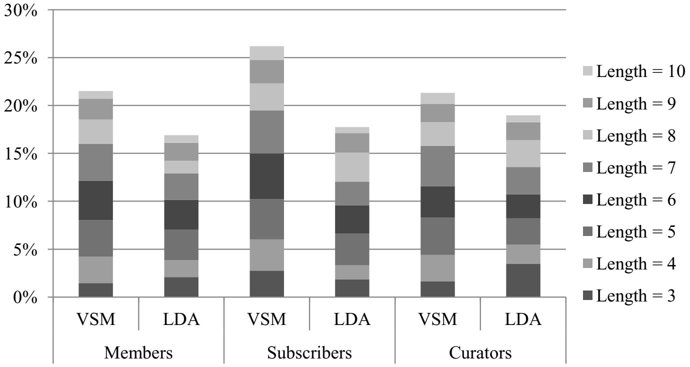

Fig. 6. Relations according to the path length for those cases where the least common subsumer has depth greater or equal to 5

In addition to the depth of the LCS, the other variable to explore is the length of the path setting up the relation. The stacked columns in figure 6 show the cumulative percentage of relations found by Vector-space and LDA models according to the path length of the relation in WordNet, with a depth of the least common subsumer greater than or equal to 5. From the chart we can state that Vector-space similarity based on subscribers produces the highest percentage of relations (26.19%) with a path length $\leq$ 10. This measure also produces the highest percentage of relations for path lengths ranging from 9 to 4. The Vector-space similarity based on members produces the second highest percentage of relations for path lengths from 10 to 6.

In summary, we have shown that similarity models based on members produce the results that are most directly related to the results of similarity measures based on WordNet. These models find more synonyms and direct relations is-a when compared to the models based on subscribers and curators. These results suggest that some users are classified under different lists named with synonyms or with keywords representing a concept in a distinct level of specificity. We also discovered that the majority of relations found by any model have a path length ≥ 3 and involve a common subsumer. Vector-space model based on subscribers produces the highest number of relations that can be considered specific (depth of LCS ≥ 5 or 6). However, for more specific relations ($7 \leq$ depth of LCS $\leq$ 9) similarity models based on members produce a higher number. In addition we considered the path length, for those relations containing a LCS placed in a depth ≥ 5 in the hierarchy, as a variable influencing the relevance of a relation. Vector-space model based on subscriber finds the highest number of relations with $4 \leq$ length ≤ 10. In general similarity models based on curators produce a lower number of relations. We think this may be due to the scarcity of lists per curator. In our dataset each curator has created 1.38 lists in average.

Linked Data Analysis: Our approach found DBpedia resources for 63.77% of the keywords extracted from Twitter Lists. In average for the 41.74% of relations we found the related keywords in DBpedia. For each relation found by Vector-space or LDA similarity we query the linked data set looking for patterns between the related keywords. Figure 7 shows the results according to the path length of the relations found in the linked data set. These results are similar to the ones produced by WordNet similarity measures. That is, similarity based on Members produce the highest number of synonyms and direct relations though in this case Vector-space similarity produces more synonyms than LDA. Vector-space similarity based on subscribers has the highest number of relations of length 3, followed by Vector-space and LDA similarity based on members.

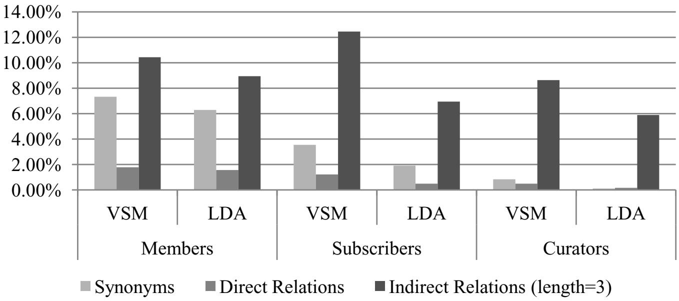

Fig. 7. Relations identified from linked data queries

Given that the Vector-space model based on members found the majority of direct relations, we present, in table 3, the relations identified in the linked data set. *Broad term* and *subClassOf* are among the most frequent relations. This means that members of lists are usually classified in lists named with keywords representing a concept with a different level of specificity. Other relations that are difficult to elicit from traditional lexicons are also obtained, such as *developer*, *genre* or *largest city*.

Table 3. Direct relations established by the Vector-space model based on members

Relation type		Example of keywords	
Broader Term	26%	life-science	biotech
subClassOf	26%	authors	writers
developer	11%	google	google_apps
genre	11%	funland	comedy
largest city	6%	houston	texas

Table 4. Indirect relations of length 3 found in the linked data set for the relations established by the Vector-space model based on subscribers

	$r_s \overset{relation_1}{\rightarrow} object \overset{relation_2}{\leftarrow} r_t$		
Relations			**Example**
type	type	67.35%	nokia → company ← intel
subClassOf	subClassOf	30.61%	philanthropy → activities ← fundraising
	$r_s \overset{relation_1}{\leftarrow} object \overset{relation_2}{\rightarrow} r_t$		
Relations			**Example**
genre	genre	12.43%	theater ← Aesthetica → film
genre	occupation	10.27%	fiction ← Adam Maxwell → writer
occupation	occupation	8.11%	poet ← Alina Tugend → writer
product	product	7.57%	clothes ← ChenOne → fashion
product	industry	9.73%	blogs ← UserLand Software → internet
occupation	known for	5.41%	author ← Adeline Yen Mah → writing
known for	known for	3.78%	skeptics ← Rebecca Watson → atheist
main interest	main interest	3.24%	politics ← Aristotle → government

In addition we also investigate the type of relations of length 3 elicited using the Vector-space model based on subscribers. The most common patterns found in the linked data set were $r_s \overset{relation_1}{\rightarrow} object \overset{relation_2}{\leftarrow} r_t$, and $r_s \overset{relation_1}{\leftarrow} object \overset{relation_2}{\rightarrow} r_t$ with 54.73% and 43.49% of the relations respectively. Table 4 shows the obtained relations according to each pattern.

With respect to the first pattern, 97.96% of the related keywords can be considered siblings since they are associated via *typeOf* or *subClassOf* relations with a common class. That is, some subscribers follow lists that share a common super concept. On the other hand, the second pattern shows a wider range of relations. Keywords are related since they are *genres*, *occupations*, *products*, *industries*, or *main interest* that appear together in the description of an individual in the linked data set.

5 Related Work

Twitter has been investigated from different perspectives including network characteristics, user behaviors, and tweet semantics among others. Twitter network properties, geographical features, and users have been studied in [15,17]. In [15] authors use the HITS algorithm to identify hubs and authorities from the network structure, while in [17] authors categorise users according to their behaviors. To identify the tweet semantics some proposals [2,1,23,6] annotate them with semantic entities using available services such as Zemanta, Open Calais, and DBpedia Spotlight [21]. In [2] tweets are linked to news articles and are enriched with semantic annotations to create user profiles. These semantic annotations of tweets have been used in a faceted search approach [1]. In [23] tweets and their semantic annotations are represented according to existing vocabularies such as FOAF, Dublin Core, and SIOC, and are used to map tweets to websites of conferences and events. In [6] authors use the semantic entities identified in Tweets to obtain the concepts associated with user profiles. In addition some classifiers have been proposed in [8] to extract players and events from sport tweets. Twitter allows the use of hashtags as a way to keep conversation around certain topics. In [18] authors have studied hashtags as candidate identifiers of concepts.

With respect to Twitter Lists, they have been used to distinguish elite users, such as celebrities, media, organizations, and bloggers [25]. In this work authors provide an analysis on the information flow of Twitter, and show dueling importance of mass media and opinion leaders. In addition, in [9] lists have been used as a source for discovering latent characteristics of users.

In the broader context of the Web 2.0 the emerging semantics of folksonomies have been studied under the assumption that it is possible to obtain a vocabulary from these classification systems. In folksonomies the set of tags around resources tends to converge [13] and users in the same social groups are more likely to use the same set of tags [20]. The semantics of the emerging relations between tags have been studied in [7,19]. A survey of the state of the art on this matter can be found in [11].

6 Conclusions

In this paper we have described different models to elicit semantic relations from Twittter lists. These models represent keyword co-occurrence in lists based on three user roles: curators, subscribers and members. We measure similarity between keywords using the vector-space model and a topic based model known as LDA. Then we use Wordnet similarity measures including Wu and Palmer, and Jiang and Conrath distance, to compare the results of the vector-space and LDA models.

Results show that applying vector-space and LDA metrics based on members produce the most correlated results to those of WordNet-based metrics. We found that these measures produce relations with the shortest Jiang and Conrath distance and high Wu and Palmer similarities. In addition, we categorize the relations found by each model according to the path length in WordNet. Models based on members produce the highest number of synonyms and of direct is-a relations. However, most of the relations have a path length ≥ 3 and have a common subsumer. We analyze these relations using the depth of the LCS and the path length as variables that help to identify the relevance of relations. This analysis shows that the vector-space model based on subscribers finds the highest number of relations when relevance is defined by a depth of LCS ≥ 5, and the path length of relations is between 10 and 4.

We also investigate the type of relations found by each of the models using general knowledge bases published as linked data. We categorize the relations elicited by each model according to the path length in the linked data set. These results confirm that the models based on members produce the highest number of synonyms and direct relations. In addition, we find that direct relations obtained from models based on members are mostly *Broader Term* and *subclassOf*. Finally, we study the type of relations obtained from the vector-space model based on subscribers with a path length of 3 and find that mostly they represent sibling keywords sharing a common class, and subjects that are related through an individual.

Acknowledgments. This work is supported by the project myBigData, and the FPI grant (BES-2008-007622) of the Spanish Ministry of Science and Innovation.

References

1. Abel, F., Celik, I., Houben, G.-J., Siehndel, P.: Leveraging the Semantics of Tweets for Adaptive Faceted Search on Twitter. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 1–17. Springer, Heidelberg (2011)
2. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 201. LNCS, vol. 6644, Part II, pp. 375–389. Springer, Heidelberg (2011)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, IJSWIS (2009)
4. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Journal of Web Semantic 7(3), 154–165 (2009)
5. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
6. Cano, A.E., Tucker, S., Ciravegna, F.: Follow me: Capturing entity-based semantics emerging from personal awareness streams. In: Making Sense of Microposts (#MSM 2011), pp. 33–44 (2011)
7. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
8. Choudhury, S., Breslin, J.: Extracting semantic entities and events from sports tweets. In: Proceedings of the ESWC 2011 Workshop on 'Making Sense of Microposts'. CEUR Workshop Proceedings, vol. 718 (May 2011)
9. Dongwoo Kim, Y.J.: Analysis of Twitter lists as a potential source for discovering latent characteristics of users. In: Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems (CHI 2010), Atlanta, CA, USA (2010)

10. Fellbaum, C.: WordNet and wordnets, 2nd edn., pp. 665–670. Elsevier, Oxford (2005)
11. García-Silva, A., Corcho, O., Alani, H., Gómez-Pérez, A.: Review of the state of the art: discovering and associating semantics to tags in folksonomies. The Knowledge Engineering Review 27(01), 57–85 (2012)
12. García-Silva, A., Szomszor, M., Alani, H., Corcho, O.: Preliminary results in tag disambiguation using dbpedia. In: Knowledge Capture (K-Cap 2009)-Workshop on Collective Knowledge Capturing and Representation-CKCaR (2009)
13. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (2006)
14. Heim, P., Lohmann, S., Stegemann, T.: Interactive Relationship Discovery via the Semantic Web. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part I. LNCS, vol. 6088, pp. 303–317. Springer, Heidelberg (2010)
15. Java, A., Song, X., Finin, T., Tseng, B.: Why We Twitter: An Analysis of a Microblogging Community. In: Zhang, H., Spiliopoulou, M., Mobasher, B., Giles, C.L., McCallum, A., Nasraoui, O., Srivastava, J., Yen, J. (eds.) WebKDD 2007. LNCS, vol. 5439, pp. 118–138. Springer, Heidelberg (2009)
16. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. CoRR, cmp-lg/9709008 (1997)
17. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proceedings of the First Workshop on Online Social Networks, WOSN 2008, pp. 19–24. ACM, New York (2008)
18. Laniado, D., Mika, P.: Making Sense of Twitter. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 470–485. Springer, Heidelberg (2010)
19. Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., Gerd, S.: Evaluating similarity measures for emergent semantics of social tagging. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, pp. 641–650. ACM, New York (2009)
20. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, pp. 31–40. ACM Press (2006)
21. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: Shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems, I-Semantics (2011)
22. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet: Similarity - measuring the relatedness of concepts. In: AAAI, pp. 1024–1025. AAAI Press / The MIT Press (2004)
23. Rowe, M., Stankovic, M.: Mapping tweets to conference talks: A goldmine for semantics. In: Social Data on the Web Workshop, International Semantic Web Conference (2010)
24. Salton, G., Mcgill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York (1986)
25. Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J.: Who says what to whom on twitter. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, pp. 705–714. ACM, New York (2011)
26. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proc. of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 133–138 (1994)