

# Characterizing Retweeting Behaviors in Twitter: On the use of Text vs. Concepts

Sofus A. Macskassy

Information Sciences Institute University of Southern California  
Marina del Rey, CA 90292, USA  
sofmac@isi.edu,  
WWW home page: <http://www.isi.edu/~sofmac>

**Abstract.** Twitter and other microblogs have rapidly become a significant means by which people communicate with the world and each other in near realtime. There has been a large number of studies surrounding these social media, focusing on areas such as information spread, various centrality measures, topic detection and more. However, one area which has received little attention is trying to better understand what information is being spread and why it is being spread. One recent line of work has been looking at the problem of modeling retweeting behaviors. This work has advocated mapping tweets into a conceptual space such as Wikipedia categories and reasoning about diffusion behaviors in that space. The work, however, did not show that this was in fact needed and the question is whether one can get equally good reasoning by staying at the token or word level. This paper looks at this particular question of whether one in fact improve upon reasoning by mapping into a more abstract space or whether there is a place for token-level modeling. We show that, in fact, token-level models do have their place when reasoning about whether a tweet is likely interesting based on the tweet words but that the conceptual space is better when reasoning about homophily–similarities between users. Ideally one would like a hybrid model and we show that while the hybrid model is not always the optimal, it does yield good performance. We here repeat part of an earlier retweet study on over 768K tweet and show that profiles using a combination of word-based and concept-based features work better than either of the simpler representations.

**Keywords:** information diffusion, social behaviors, social networks, text mining, user modeling

## 1 Motivation

The use of micro-blogging services, such as Twitter, has exploded exponentially in recent years. For example, currently, millions of Twitter users post millions of 140-character tweets about topics ranging from daily activities, to opinions, to links to funny pictures. Beyond the large collection of user generated text, Twitter also has a social network aspect, allowing users to publicly message one another directly, and set up a social network of people who follow one another’s tweets. This rich relational and textual setting has spurred research in a number of areas beyond traditional network analysis (e.g., [9, 7]). For instance, Twitter has been analyzed to discover breaking news [25], as a forum for analyzing media events [26], as a vehicle for information diffusion [12, 10, 11], as a mechanism for language learning [1], and even for detecting natural disasters in real-time [24].

Recent work has argued that context is critically important when one wants to delve into such details [13]. The argument went that not all links area created equal, not all

people are the same, and not all pieces of content are interesting. If one can tag people, links and content with semantically meaningful categories, then one ought to be able to generate much finer-grained behavioral and predictive models to understand the dynamics of these social media networks. To do this, the authors analyzed the text of tweets to find entities which they then looked up in Wikipedia. From here, they were able to extract out a tree of categories associated with a particular entity and these categories made up a user profile. This user profile was then used to model retweeting behaviors. The argument was using these higher-level abstract categories provided better coverage than the sparse 140-character tweets. This conjecture was not proven, however, and the question is whether pure text-based methods might in fact be just as good at modeling retweet behaviors. That is what this paper explores.

The key to our contribution lies in shedding light on whether text-based user profiles are in fact useful for predicting retweeting behaviors and whether they are competitive with the abstract profiles from prior work. Specifically, we will build a text-based user model and use it in a similar study to that of a prior study on retweet behaviors [13]. We will show that the text-based model actually outperforms the abstract user model when reasoning about the user’s affinity to a tweet but that the abstract model is better when considering homophily (how similar two users are). We explore a hybrid model to get the best of both models and show that it does perform comparably to the focused models but that it is not always the best.

The rest of the paper is outlined as follows: in Section 2 we discuss related work. We then, in Section 3 describe two approaches for building user models: the prior model for tagging content and building user profiles, and the new text-based model which is based on the Okapi BM25 information retrieval model [22]. We outline in Section 4 the four behavior-based information-propagation models from prior work which we reuse in our study here. Section 5 describes our case study on Twitter data, where we show that the new text-based models work better than the general models and outperforms the abstract models in certain cases. We finish by discussing our findings Section 6.

## 2 Related Work

Information diffusion is a topic which is receiving an increasing amount of attention in many areas, social media as well (see, e.g., [12, 6, 23, 10, 27]). Most of these endeavors, however, are focused on developing global statistics and metrics, such as understanding information cascades or learning pathways that information propagated. None of the methods are seriously considering treating individuals differently or relations differently beyond what is captured by the high-level statistics.

Community detection algorithms have received significant attention in recent years (see, e.g., [3, 17, 18, 29, 21]). The most common approaches take a graph (such as a social network) and split it into  $k$  disjoint clusters, where each cluster supposedly represents a “community” in that graph.

Chen, et al., [2] explore the problem of recommending content (tweets). They build a number of recommender approaches, one of which is “topic” based. They model the topics of a user as a bag-of-words generated from the user’s tweets (with TF/IDF weights). They then compare this feature vector modeling of the topics to a similar feature vector of an incoming tweet to determine if it should be recommended.

Another approach to analyzing Twitter that uses topics is TwitterRank, which aims to identify influential micro-bloggers [28]. This approach leverages LDA by creating a single document from all of a user’s tweets and then using LDA to discover the topics on this “document.”

Being able to semantically identify entities in content requires that we can disambiguate the entities within the content. Disambiguating entities is a (relatively) old problem in natural language processing [4] and there has been previous work on using dictionaries to aid this task (e.g., [30]). We instead leverage Wikipedia as a knowledge base. While other research has proceeded in this direction [8, 15, 16, 5, 19], a key difference to our work is that none of these approaches are leveraged to determine the topics that a user writes about, but rather are mechanisms for disambiguating entities in text.

### 3 Building User Profiles From Content

The focus of this paper is to compare and contrast two different approaches for building user profiles with the goal of modeling retweeting behaviors. The first approach is based on prior work and builds a user profile by mapping tweet content into abstract concepts as represented by Wikipedia categories [13]. The second text-based user model which we introduce in this paper is based on the Okapi BM25 information retrieval model [22]. We also discuss how we might combine these two to get the best of both worlds.

#### 3.1 Building Concept-level User Profiles

The approach we adopt here stems from prior work which profiles Twitter users based on their tweets [14]. This approach was used recently to show how such a profile performed well in modeling user’s retweeting behaviors [13]. While the approach was used specifically to profile Twitter users, it can work with any user-generated content. As our study focuses on Twitter, this approach is a good fit for us as we can leverage this to focus on the behavior models below.

The aim is to generate “topic profiles” of users based upon what they post about. We define a topic profile as a list of the common, high-level topics about which a user posts, under the premise that these are the topics of interest to a Twitter user, since s/he tweets frequently about them.

The concept-based approach to discovering a Twitter user’s topic profile is based on the idea that the topics of interests can be identified by finding the *entities* about which a user tweets, and then determining a common set of *high-level categories* that covers these entities. As a running example, consider the following real-world tweet:

```
#Arsenal winger Walcott: Becks is my England  
inspiration: http://tinyurl.com/37zyjsc
```

There are four entities of interest in this tweet: Arsenal, which refers to the Arsenal Football Club of England; Walcott, which refers to Theo Walcott, a player for Arsenal; Becks, which refers to football superstar David Beckham; and England. A category that covers these entities within the tweet might be “English Football.” To develop a concept-based topic profile for a user, we analyze all of their tweets and determine the set of common high-level categories that covers the set of tweets. This set of categories defines the topic profile. In our example, the profile may include “English Football,” “World Cup,” etc.

In order to map entities into high-level topics, and following other prior work in this space, we here use Wikipedia as a knowledge base. Wikipedia provides encyclopedic knowledge about entities which we leverage to disambiguate their mentions in the tweets. Once disambiguated, we use the “folksonomy”<sup>1</sup> defined by Wikipedia’s user-defined categories to map entities to the categories that will define the topic profile.

<sup>1</sup> A folksonomy is a crowd-sourced taxonomy

The general approach consists of two steps and is shown in Figure 1. In step 1 (“Discover Categories”), we discover the entities in the tweets, disambiguate them, and then retrieve the sub-tree of categories from the folksonomy that contains the disambiguated entity. In step 2 (“Discover Profile”), we analyze all of the subtrees for all of the discovered entities in a users set of tweets, and determine the set of categories that defines that user’s topic profile (e.g., the topics of interest).

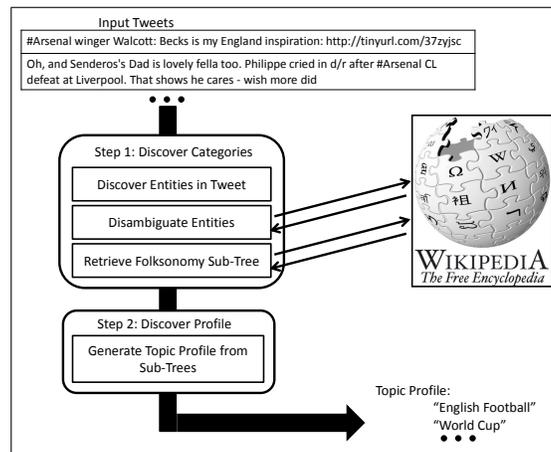


Fig. 1. Topic Profiles from User-generated Content

### Discovering Categories for Tweets

The first step in discovering the categories for tweets involves discovering the entity mentions in the tweets themselves. Generally, the task of discovering entities is called “named entity recognition” (NER). While much work on NER first parses sentences and finds phrases that include proper nouns, tweets are ungrammatical and noisy, and we therefore cannot guarantee parses for our data. Our approach is to look for capitalized, non-stopwords as possible named entities. This ensures high recall (we retrieve many possible entities) while conforming to the difficulty of our data.

Once we have discovered the entities in a tweet, we next disambiguate them by finding the page in Wikipedia about that entity. The way we identify the specific Wikipedia page for an entity is to search for the entity (looking for the entity either in the text of a page or in the title). Wikipedia may return a set of candidates that match the entity. To deal with such a disambiguation problem, we leverage the “local context” of the tweet. Specifically, we treat the text of the tweet (excluding the entity term to disambiguate) as the context for that entity. If we are using the example tweet, and our current entity to disambiguate is “Arsenal,” then the local context is {winger, Walcott, Becks, ...}. We then select the Wikipedia page that contains the most of these context words. Remember that we are interested in the higher concepts which are relevant to the entity in question. We retrieve these as a category tree based on the folksonomic category tree which the identified entity page is situated in.

This is done by following the categories which can be found at the bottom of most Wikipedia pages. Each such category has a name, and links to its category page. That category page in turn contains a list of entities that belong to that category, along with another set of categories that generalize the current one (e.g., parent categories). We empirically chose to go 2 levels deep, as at this point the categories were sufficiently general and vague, and with a branching factor averaging around 20, this already provides a large number of categories.

**Generating User Profiles from Category Trees** The output of the previous step is a set of sub-trees rooted on the categories for each of the disambiguated entities in each of the tweets. The goal of this step is to take in this forest of sub-trees and generate a user-specific topic-of-interest profile. We represent this profile as a high-dimensional vector

of category ids and their weight. The weight of unseen categories is 0. The weight of category  $c$  for user  $u$  is computed as follows:  $w(c, u) = \sum_{t \in \mathbf{t}_{c,u}} b^{-d_{c,t}}$ , where  $\mathbf{t}_{c,u}$  is the set of tweets from user  $u$  which contain category  $c$ ,  $d_{c,t}$  is the depth of category  $c$  in the sub-tree it was observed in, and  $b$  is a branching factor (we use  $b = 20$ .)

### 3.2 Building Text-based Profiles

While the category-profile above has many fine points, it is also complex and requires non-trivial computations. The key question we ask in this paper is whether such a profile is necessary. Perhaps a simpler word-based profile will work just as well in modeling between behavior. To this end, we define a word-based profile which uses only the words that show up in a tweet. The approach we chose to build a word-based profile leverages the well-known Okapi BM25 ranking function from information retrieval [22]. We chose this particular function because it works well with short text.

As with the category-based profile, the word-based profile is represented as a vector. However instead of a vector of category ids and their weights, this profile consists of a vector of words and their weights. As above, if a word is not used by the user then its weight is 0 and can be ignored for that user.

The weight for word  $w$  for user  $u$  is computed as follows:

$$w(w, u) = IDF(w) \cdot \frac{f(w, u) \cdot (k_1 + 1)}{f(w, u) + k_1 \cdot (1 - b + b \cdot \frac{|D_u|}{avgD})}$$

$$IDF(w) = \log \frac{N_u - n(w) + 0.5}{n(w) + 0.5},$$

where  $f(w, u)$  is the number of times  $u$  has used word  $w$ ,  $k_1$  and  $b$  are both free variables which we set to 1.8 and 0.75 respectively based on external sources.  $D_u$  is the number of words used by user  $u$  and  $avgD$  is the average number of words used by users. In computing  $IDF(w)$ , we define  $N_u$  as the number of users and  $n(w)$  the number of users using word  $w$ .

Because tweets are generally short, we use both stop-listing and stemming [20] to reduce the dimensionality of the words to help with the otherwise inherent sparsity introduced in the micro-blogging environment. Also, following standard practice, we ignore or prune words which are used by fewer than 5 users.

### 3.3 Building Hybrid Profiles

It may be that the category-based profile is good in a particular situation whereas the word-based profile is good in another. In fact, we will show that this is the case. Since neither is always the better profile to use, it might be possible to combine the two in a principled manner to achieve a consistently good performance.

We explored a few different hybrid profiles, including using each of the category- and word-based profiles separately and going with the profile which yielded the most confidence, naïvely combining the two profile vectors, to generate a vector in a combined space or averaging the scores from using each of the profiles.

Empirically, averaging the scores yielded the best performance among the hybrid profiles and that is the one we will report on below.

## 4 Information-propagation Behavior Models

The study which prompted the work described in this paper used category-based profiles to explore four different retweeting behavior models [13]. We perform the same experiments in this paper to ensure a valid comparison with the original work, and we therefore briefly describe the four models from the original study.

The general problem explored is that of modeling retweet behaviors in Twitter users. As a user processes a stream of tweets, at some point a decision is made to retweet one of the observed tweets. We explore four different models for selecting these tweets.

#### 4.1 General Model (general)

The general model assumes that a user will randomly retweet any tweet previously seen, but with a much higher likelihood of retweeting a tweet just seen than one seen longer ago. Based on prior work, use a powerlaw to represent the general retweeting model:

$$P_{\text{gm}}(x) = \alpha * \text{time}(x)^{-\beta},$$

where  $P_{\text{gm}}(x)$  is the likelihood that  $x$  will be retweeted and  $\text{time}(x)$  is defined as the number of minutes passed since  $x$  was original tweeted.

#### 4.2 Recent Communication Model (recent)

The second model uses the network and recency effect, where a user is be more likely to retweet someone s/he has recently been in “contact” with either through a retweet or through a direct message (by using the @user construct in Twitter.)

We modify the general propagation model to especially consider tweets by someone the user has recently been in contact with. This model is defined as:

$$P_{\text{recent}}(x) = P_{\text{gm}}(x) * [\alpha * P(x|I(x)) + (1 - \alpha) * P(x|\neg I(x))],$$

where  $\alpha$  is a parameter we estimate for the likelihood that a user will retweet someone which the user has been in contact with within the last 24 hours, and  $I(x)$  ( $\neg I(x)$ ) represent the fact that  $x$  was tweeted by someone (*not*) in the set of recent contacts.

#### 4.3 On-topic Model (topic)

This models whether a person is more likely to retweet a tweet which is aligned with the user’s topic-of-interest profile. Remember that a user’s profile and a tweet both are represented as a high-dimensional vector of weighted Wikipedia categories (or words).

We define the *similarity* between a tweet and a user’s profile as the *cosine distance* of the two vectors. By observing what is retweeted, we can generate the underlying empirical distribution of  $P_{\text{ts}}(x|sim_T(x, u))$ , where  $sim_T(x, u)$  is the similarity between a user’s profile and that of the tweet (regardless of whether those profiles are category-based, word-based or a hybrid). The topic-based model is then defined as:

$$P_{\text{topic}}(x) = P_{\text{gm}}(x) * P_{\text{ts}}(x|sim_T(x, u)).$$

As in the prior study, the empirical model  $P_{\text{ts}}(x|sim_T(x, u))$  comes from the data, we may find that there are certain levels of similarity where a user is more likely to retweet.

#### 4.4 Homophily Model

The final retweet model we use is based on profiles of users. It may be that a user is more likely to retweet another user if they share similar profiles, regardless of the content of the tweet. We define similarity as above and represent the profiles as vectors in high-dimensional space as described above.

By observing what is retweeted, we generate the underlying empirical distribution of  $P_{\text{ps}}(x|sim_H(x, u))$ , where  $sim_H(x, u)$  is the similarity between a user’s profile and that of the profile of the user who sent the original tweet. The homophily-based model is then defined as:

$$P_{\text{homophily}}(x) = P_{\text{gm}}(x) * P_{\text{ps}}(x|sim_H(x, u)).$$

As above,  $sim_P(x, u)$  is an empirical model which comes from the data.

## 5 Case Study

We now turn to our case study. We focus here on a Twitter data set which we have processed using the approach above to generate user profiles. We explore which of our four models best fit the observed retweeting behaviors in the data.

### 5.1 Data

We here consider a data set of tweets collected between 9/20/2010 and 10/20/2010 based on monitoring 2,400 Twitterers in the Middle East. We identified these individuals using a snowball sampling method where we started from a seed set of ~125 Twitterers who self-reported (in their profile) to reside in the Middle East. From there, we expanded the set of users to monitor whenever we saw a retweet or a mention (the `user` construct), adding only users who self-reportedly were in the same region. After a short period of time, we had reached ~2,400 Twitterers which turned out to be a fairly stable set of users and we have kept this set since then.

The full tweet dataset (from 9/20/2010 through 10/20/2010) contains over 768,000 tweets. We first down-selected to users who had at least three tweets and three retweets. From here, we ended up with 482K tweets, 43% of which had both categories and unpruned words and 84% of which had non-pruned words. Of these 482K tweets, 103K of them were retweets where 70% of them had both categories and words and 94% contained non-pruned words. Of these, 16K were retweets where the person retweeted also had tweeted and retweeted at least 3 times.

### 5.2 Experimental Methodology

The purpose of our study is to explore which of our models yield the better performing predictive model. Prior work suggests that the homophily-based model worked the best with a category-based profile. We here explore whether this and other prior findings hold true when put up against a word-based profile.

We identify the most likely model for each retweet by calculating the respective probabilities (e.g.,  $P_{\text{gm}}(x)$ ) and choosing the model with the highest probability. We sum up over all retweets how often each model was the most likely.

We also compute, for each user, the overall best model for that particular user. We do so by computing, for each model, the overall likelihood of seeing all retweets for a given user. We then compute, for each model, how many users were best explained by that model.

### 5.3 Fitting the Models

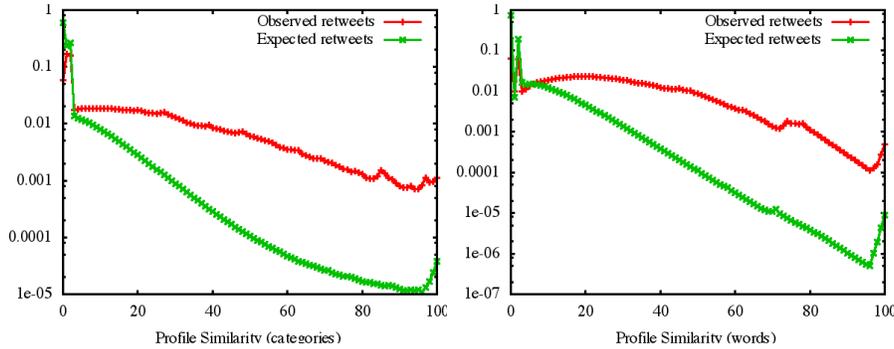
We first fit our models to the data. While there may be statistical issues with fitting and evaluating the models on the same data if one wanted to use the models for predictive behavior, we are here particularly interested in which models best fit (and hence explain) the data.

We start by fitting the general model to the distribution of the minutes between a retweet and the original tweet. This distribution follows a powerlaw distribution as we see in Figure 2 and when we fit our general model to this distribution, we get the following powerlaw distribution:

$$P_{\text{gm}}(x) = 0.2 * \text{time}(x)^{-1.15}.$$

We next fit the “recency” model. Roughly 37% of all retweets were a retweet of someone who the Twitterer had communicated with (through a retweet or a mention) within the last 24 hours. Hence, the recency model is instantiated as:

$$P_{\text{recent}}(x) = P_{\text{gm}}(x) * [0.37 * P(x|I(x)) + 0.63 * P(x|\neg I(x))].$$

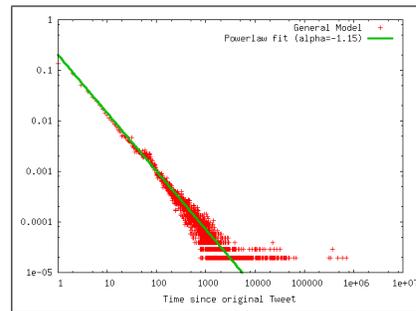


**Fig. 3.** Observed vs. expected distributions of how well the profile of an original Twitter user matches with the user doing the retweeting. The left shows the distribution when using category-based profiles, the right the word-based profiles. At worst the probability goes down to  $1e - 05$  whereas the word-based profiles go down to  $1e - 07$ . They are otherwise very similar qualitatively and we see there is a big lift in observed distributions. Note the log scale on the  $y$ -axis.

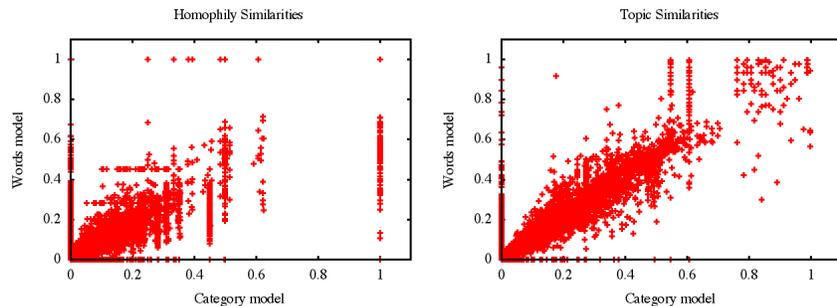
To fit the Topic model and Homophily model, we generated the distributions to compute  $P_{ts}(x|sim_T(x, u))$  and  $P_{ps}(x|sim_H(x, u))$ . To do this, we computed, for each retweet, the similarity between the tweet profile and the user profile (of the user doing the retweet) as well as the similarity between the profiles of the user doing the retweeting and the user posting the original tweet. Our similarity measure (the cosine distance) lies in the range  $[0 : 1]$ , which we discretize into 101 bins (0 through 100). We then computed, for each bin, the ratio of retweets which fell into that bin (separately for the profile-similarity and the topic-similarity). These ratios then make up the empirical distribution which is the fit for our two models. We also computed the “expected” ratios as a baseline comparison. This was done by computing the average similarity between any tweet and a user (for the expected fit of the topic-model with no observed retweeting behavior) as well as the average similarity between users (for the expected fit of the profile-model again with no observed retweeting behavior).

Figure 3 shows the empirical distributions for the homophily-models when using the category-based profiles as well as the word-based profiles. The word-based and hybrid-based profiles have almost identical distributions to that of the word-based profiles and are not shown here. Note that the  $y$ -axis is a log-scale.

We show, for both distributions, the observed retweet fit and the expected fit. We can see that although the observed distributions have high probability mass at 0 similarity, the expected models had much higher masses at 0 and much lower probability mass as the similarity increased, whereas the observed models showed that there was a strong signal in the similarity. Also note that the models generally decrease as we get closer to



**Fig. 2.** General Model:  $y$ -axis is the ratio of retweets, and the  $x$ -axis is the number of minutes between a retweet and the original tweet. As can be seen, this approximates a powerlaw distribution with a slope of  $-1.15$ .



**Fig. 4.** How well do different profiles correlate? Specifically, for the Homophily (on left) and Topic (on right) models, how much correlation is there between predictions using the category-based profile vs. the words-based profile?

| Model            | Profile Type |            |      |            |        |            |
|------------------|--------------|------------|------|------------|--------|------------|
|                  | Category     |            | Word |            | Hybrid |            |
|                  | Wins         | Pct        | Wins | Pct        | Wins   | Pct        |
| General          | 1446         | 9%         | 946  | 6%         | 1005   | 6%         |
| Recency          | 4918         | 32%        | 4976 | 32%        | 5229   | 34%        |
| Topic            | 3183         | 20%        | 3834 | 25%        | 3390   | 22%        |
| <b>Homophily</b> | 7037         | <b>45%</b> | 6486 | <b>42%</b> | 6591   | <b>42%</b> |

**Table 1.** How often was each model the most likely explanation for a retweet. As we can see, the homophily-based model was the clear winner regardless of which profile was used. We also see that the hybrid profile gets the best performance of the homophily-model.

a similarity of 1 and then increase. This is because there are just few tweets and profiles that are at the 90 – 99% level of similarity. We see the increase at the end for both the observed retweets as well as for the expected retweets. We therefore compute the probability as the logodds of the empirical distribution and the expected distribution.

Prior studies have shown that the category-based models (topic and homophily) fit the data better than the general model [13]. We will ask whether these models also perform better than the word-based models which are much simpler to compute. In other words, can we justify through improved performance the work needed to map text into higher concepts such as the Wikipedia categories.

#### 5.4 Results

To better understand the difference between the category- and word-based profiles, we look at whether their respective prediction scores correlate. In particular, we wanted to understand why there is little difference and we also wanted to validate that there were tweets that had low score in one profile but not the other (i.e., there were informative words, but none of them mapped into Wikipedia). Figure 4 show the correlation between scores of the category- and word-based profiles on the Homophily and Topic models. Not surprisingly, there is a strong correlation between scores, but we also see a fair amount of outliers on either side where one profile picks up signals where the other does not. This would suggest that the two types of profiles probably would have similar performance but perhaps a hybrid model would do better.

We first explore how often each model best explained each of the 16K retweets in our data set. Table 1 shows, for each of the four models, and for each of the three types

| Model            | Profile Type |            |      |            |        |            |
|------------------|--------------|------------|------|------------|--------|------------|
|                  | Category     |            | Word |            | Hybrid |            |
|                  | Wins         | Pct        | Wins | Pct        | Wins   | Pct        |
| General          | 203          | 12%        | 145  | 9%         | 145    | 9%         |
| Recency          | 234          | 14%        | 222  | 13%        | 232    | 14%        |
| Topic            | 518          | 31%        | 597  | 36%        | 529    | 32%        |
| <b>Homophily</b> | 1116         | <b>67%</b> | 1076 | <b>65%</b> | 1143   | <b>69%</b> |

**Table 2.** How many users were best “explained” by each model as the most likely explanation for that user’s overall behavior. As we can see, the homophily-based model is significantly better than the other models, with the topic model a far second after it.

| Model            | Profile Type |            |            | # Models | Profile Type |      |        |
|------------------|--------------|------------|------------|----------|--------------|------|--------|
|                  | Category     | Word       | Hybrid     |          | Category     | Word | Hybrid |
| General          | 12%          | 8%         | 8%         | <b>1</b> | 447          | 461  | 468    |
| Recency          | 13%          | 12%        | 13%        | <b>2</b> | 493          | 532  | 516    |
| Topic            | 27%          | 32%        | 27%        | <b>3</b> | 406          | 414  | 419    |
| <b>Homophily</b> | <b>48%</b>   | <b>48%</b> | <b>51%</b> | <b>4</b> | 319          | 258  | 262    |

**Table 3.** On average, how often was each model used by each user? How did this differ across different types of profiles? **Table 4.** How many models were needed to best fit a particular user?

of profiles, how often it was the best explanation for an observed retweet. As shown in prior work, the homophily-based model significantly outperforms the other models, regardless of the type of profile being used. However, the main question we address in this paper is whether there is a difference in the comparative performance of the three types of profiles. Overall, we see a slight increase in the topic models when we use the word or hybrid profiles; we also see that the general model is used less.

These numbers may be dominated by the users who have the most tweets. We therefore analyzed each user and identified the “best” model per user. Table 2 shows for how many users each model was deemed the “best fit”. As we can see, the qualitative behavior is roughly the same although the homophily model is now by far the best fitting model, where the hybrid profile yields the overall best performance.

Taking this analysis deeper, we look at what models best describe particular users. For each retweet, we identify the model which best explains that retweet and count these up over users. For a user we then get the number of times each model was the best. Table 3 shows, on average over all users, how often each model was picked. We see a qualitatively same distribution as before, where the Homophily model using the hybrid profile was picked the most. An alternate view of the same data is shown in Table 4, where we compute for each user how many different models were picked over all that user’s retweets. The table shows how many users were best explained by 1, 2, 3 or all 4 models. Interestingly we see that using categories resulted in more people having to use 4 models, whereas using words and a hybrid profile resulted in the richer models being used more often and fewer users needed to use the general model.

All of these results complement the earlier work where we have shown that the Homophily-based model is by far the better model but that its performance does rely on how similarity is computed. The consistent result shows that word-based and category-based profiles yield good results, where the word-based profile in general has the best

performance for the Topic-model but the hybrid profile in general gets the best performance out of the models.

## 6 Discussion

We have in this paper taken a close look at what drives retweet processes in Twitter. We studied a set of Twitter users over a period of a month and sought to explain the individual information diffusion behaviors, as represented by retweets, in this domain. The key question we asked in this paper was whether the underlying representation of profiles would have an impact on how well our retweet models fit the observed data.

The work builds upon a recent study which argued that representing user profiles in an abstract space of Wikipedia categories was necessary in order to be able to get useful similarity measures. The argument was that text from microblogs would be too sparse to generate good profiles. This paper showed that this is in fact not necessarily the case. We showed that building user profiles from the text and using them in the retweet models yielded performance comparable to that of using the Wikipedia categories. However, we also found that the ratio of tweets where categories could be extracted was less than 50%, showing the limitation of relying completely on this mapping.

We further showed that a model using a hybrid profile of both categories and text worked even better although they still did not explain all retweets observed. This suggests that there is still plenty of work to be done.

### Acknowledgments

This research was supported in part by DARPA grant No. W911NF-12-1-0034. The views and conclusions herein do not represent those of DARPA or the U. S. Government.

### References

1. Borau, K., Ullrich, C., Feng, J., Shen, R.: Microblogging for language learning: Using twitter to train communicative and cultural competence. In: Proceedings of the International Conference on Advances in Web Based Learning (2009)
2. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: Proceedings of the international conference on Human factors in computing systems (2010)
3. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. *Physical Review E* 70 (2004), 066111
4. Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (1999)
5. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of EMNLP-CoNLL (2007)
6. Gomez-Rodriguez, M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2010)
7. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proceedings of the first workshop on Online social networks (2008)
8. Kulkarni, S., Singh, A., Ramakrishnan, G., Chakrabarti, S.: Collective annotation of wikipedia entities in web text. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (2009)
9. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: *In Proc. of the International Conference on World wide web* (2010)

10. Lerman, K., Ghosh, R.: Information contagion: an empirical study of spread of news on digg and twitter social networks. In: Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM) (May 2010)
11. Lerman, K., Hogg, T.: Using a model of social dynamics to predict popularity of news. In: Proceedings of 19th International World Wide Web Conference (WWW) (2010)
12. Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., Hurst, M.: Cascading behavior in large blog graphs. In: SIAM International Conference on Data Mining (SDM) (2007)
13. Macskassy, S.A., Michelson, M.: Sofus a. macskassy (2011). why do people retweet? anti-homophily wins the day! In: International Conference on Weblogs and Social Media (ICWSM) (2011)
14. Michelson, M., Macskassy, S.A.: Discovering users' topics of interest on twitter: A first look. In: Proceedings of the Workshop on Analytics for Noisy, Unstructured Text Data (AND). Toronto, Canada (2010)
15. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (2007)
16. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM conference on Information and knowledge management (2008)
17. Muff, S., Rao, F., Cafilisch, A.: Local modularity measure for network clusterizations. *Physical Review E* 72(056107) (2005)
18. Newman, M.: Modularity and community structure in networks. In: Proceedings of the National Academy of Sciences. pp. 8577–8582 (2005)
19. Passant, A., Hastrup, T., Bojars, U., Breslin, J.: Microblogging: A semantic and distributed approach. In: Proceedings of Workshop on Scripting for the Semantic Web (2008)
20. Porter, M.F.: An algorithm for suffix stripping. *Program* 14(3), 130–137 (July 1980)
21. Porter, M.A., Onnela, J.P., Mucha, P.J.: Communities in networks. *Notices of the AMS* 56(9), 1082–1097, 1164–1166 (2009)
22. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval* 3(4), 333–389 (Apr 2009)
23. Sadikov, E., Medina, M., Leskovec, J., Garcia-Molina, H.: Correcting for missing data in information cascades. In: ACM International Conference on Web Search and Data Mining (WSDM) (2011)
24. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the International Conference on World Wide Web (2010)
25. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: Twitterstand: news in tweets. In: Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (2009)
26. Shamma, D.A., Kennedy, L., Churchill, E.F.: Tweet the debates: understanding community annotation of uncollected sources. In: Proceedings of the first SIGMM workshop on Social media (2009)
27. Suh, G., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: Second IEEE International Conference on Social Computing (SocialCom). pp. 177–184 (2010)
28. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining (2010)
29. White, S., Smyth, P.: A spectral clustering approach to finding communities in graph. In: Proceedings of the SIAM International Conference on Data Mining (SDM) (2005)
30. Yarowsky, D.: Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In: Proceedings of the 14th conference on Computational linguistics (1992)