Discovering Users' Topics of Interest on Twitter: A First Look

Matthew Michelson
Fetch Technologies
841 Apollo St., Ste 400
El Segundo, CA 90245
mmichelson@fetch.com

Sofus A. Macskassy
Fetch Technologies
841 Apollo St., Ste 400
El Segundo, CA 90245
sofmac@fetch.com

ABSTRACT

Twitter, a micro-blogging service, provides users with a framework for writing brief, often-noisy postings about their lives. These posts are called "Tweets." In this paper we present early results on discovering Twitter users' topics of interest by examining the entities they mention in their Tweets. Our approach leverages a knowledge base to disambiguate and categorize the entities in the Tweets. We then develop a "topic profile," which characterizes users' topics of interest, by discerning which categories appear frequently and cover the entities. We demonstrate that even in this early work we are able to successfully discover the main topics of interest for the users in our study.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*

General Terms

Algorithms, Experimentation

1. INTRODUCTION

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 messages, called "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., [12, 10]). For instance, Twitter has been analyzed to discover breaking news [19], as a forum for analyzing media events [20], as a mechanism for language learning [2], and even for detecting natural disasters in real-time [18].

In this paper, we focus on discovering the topics of interest for a particular Twitter user. This allows for clustering and search of Twitter users based upon their interests. Specifically, we aim to generate "topic profiles" of Twitter users based upon what they Tweet about. A topic profile is 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. This study centers on Twitter because the text is noisy, ambiguous, and large, making it a rich data set to analyze. Further, if we can begin to address the problem of discovering user interests using the difficult Twitter data, we should be able to perform even better with cleaner data, such as that from news and blogs.

The real-world nature of Tweets means they are noisy and complex, making our problem difficult. Tweets are intentionally short (limited to just 140-characters) which forces users to be creative in how they constrain the text while preserving meaning. As with text messages in general, this leads to noise. Users rely on common acronyms (e.g., "d/r" means "dressing room" in sports), disambiguation via context ("Arsenal" in a Tweet is the football team and not the car, because other players are mentioned in the Tweet), combinations of the two ("Hawks" means "Chicago Blackhawks," if the Tweet mentions "Chicago"), and other constraining mechanisms. However, Tweets can also be information rich, because users tend to pack substantial meaning into the short space.

Our approach to discovering a Twitter user's topic profile hinges upon 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." Therefore, to develop a 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.

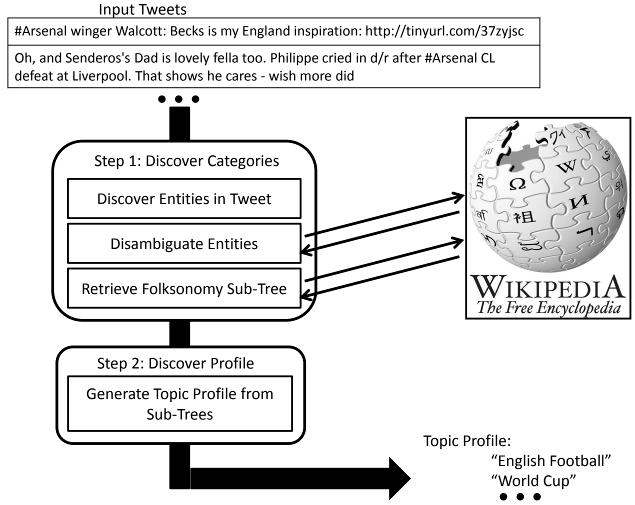
AND'10, October 26, 2010, Toronto, Ontario, Canada.

Given the huge number of Tweets and Twitter users, discovering topic profiles needs to be done automatically. However, because Tweets are noisy and ambiguous, such automatic analysis is fraught with difficulties. First, their noisy nature makes finding the entities within the Tweets quite challenging, and makes their references noisy. Second, even if the entities are found in the Tweets, they are often ambiguously described, relying on the context of the Tweet and knowledge about the poster to disambiguate the entities. For instance, having read a number of Tweets from the example user, we know he often writes about English football, therefore the entity Arsenal likely refers to the English football club. Further, the combination of Arsenal with Walcott, England, and Becks, clues us in that each of these mentions refers to the football reference.

Finally, disambiguation is compounded by the fact that the text of Tweets is short, and while there are many Tweets in the aggregate, no single user generates a huge volume (millions) of them. Therefore, we cannot treat a single user's Tweets as a usable corpus, since it would be too small. This motivates using some sort of outside knowledge base to overcome the lack of mentions and to aid in disambiguation. Further, if the knowledge base is relational, such as an ontology, we can use this relational structure to determine how to categorize the Tweets based on the discovered entities.

In this paper we address these challenges by leveraging 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"[1] defined by Wikipedia's user-defined categories to map entities to the categories that will define the topic profile.

Our overall approach, which we call "Twopics," breaks into two high level steps. In the first step, 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. This is the "Discover Categories" step. In the second step, 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). This is the "Discover Profile" step. This process is shown in Figure 1.



Input Tweets
#Arsenal winger Walcott: Becks is my England inspiration: http://tinyurl.com/37zyjsc
Oh, and Senderos's Dad is lovely fella too. Philippe cried in d/r after #Arsenal CL defeat at Liverpool. That shows he cares - wish more did
Step 1: Discover Categories
Discover Entities in Tweet
Disambiguate Entities
Retrieve Folksonomy Sub-Tree
WIKIPEDIA
The Free Encyclopedia
Step 2: Discover Profile
Generate Topic Profile from Sub-Trees
Topic Profile:
"English Football"
"World Cup"

Figure 1: Twopics: Entity-Based Topic Profiles

2. GENERATING TOPIC PROFILES

Twopics breaks into two high-level steps. In the first step, we find the entities in each Tweet, disambiguate them, and retrieve the sub-tree of the folksonomy's categories that contains the disambiguated entity. Since the output of this step is a set of categories for the Tweets, we call it the "Discover Categories" step, as in Figure 1. In the second step, we generate a topic profile for the user based on the discovered categories contained in the sub-trees. We call this the "Discover Profile" step. Here we describe each of those steps in more detail.

2.1 Discovering Categories for Tweets

The first step in discovering the categories for Tweets involves discovering the entity mentions in the Tweets themselves. This can be challenging because Tweets are non-grammatical (precluding parsing) and sometimes all capitalized (or all lowercased). Generally, the task of discovering entities is called "named entity recognition" (NER). Much work on NER first parses sentences and finds phrases that include proper nouns. However, Tweets are ungrammatical and noisy, and we therefore cannot guarantee parses for our data. Therefore, we use all capitalized, non-stopwords as possible named entities. This ensures high recall (e.g., we retrieve many possible entities) while conforming to the difficulty of our data (e.g., its not grammatical for parsing in many cases). Using our running example Tweet, the entities discovered in this manner are underlined:[2]

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

Once we have discovered the entities in a Tweet, we next disambiguate them by leveraging Wikipedia as a knowledge-base. Specifically, we first query Wikipedia for each discovered entity. Wikipedia then returns a set of candidates that match the entity (either including the entity in the text of a page or in the title). Figure 2 shows a subset of categories returned by Wikipedia for the example Tweet, with the correct candidate entity underlined. As the figure shows, for some entities, the disambiguation requires deciding between a large number of possibilities.

To deal with the 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, ...}. Again, note that we exclude stopwords from the context. More formally, we define the Tweet's local context, C_T, for an entity, E_T, as:

$$C_T \Leftrightarrow \{(t_T \in T_T)/E_T\}$$

where T_T is the set of terms in the Tweet.

We then compare the local context around an entity to the text of each candidate entity's page from Wikipedia. We define each candidate entity from Wikipedia as $e_i \in E$

[1] A folksonomy is a crowd-sourced taxonomy

[2] Note, we ignore the # sign which is specific for creating within Twitter search links.

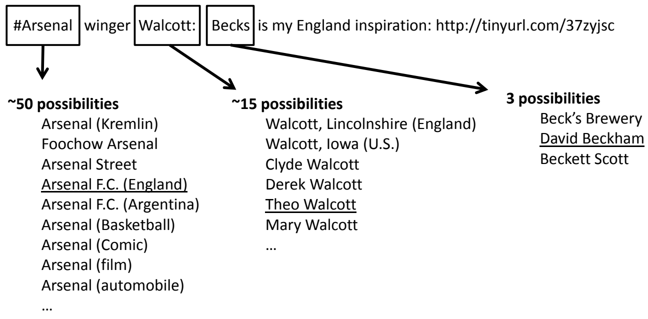


Figure 2: Candidate Entity Matches from Wikipedia

(the set of candidates), and similarly define the context for the page of each candidate entity as:

$$C_{e_i} \Leftrightarrow \{(t_{e_i} \in T_{e_i})/e_i\}$$

Because the texts of Tweets are short, and there are few of them, it makes it difficult to employ corpus based disambiguation methods (e.g., [7]). Therefore, we leverage the knowledge base (Wikipedia) as a mechanism to do this term association for us. We select the entity from Wikipedia that has the best overlap with the entity in the Tweet, given the local context. Our goal is to choose the Wikipedia entity, e_i, from the set of entity candidates E that maximizes the overlap between contexts:

$$\arg\max_{e_i \in E}(C_T \cap C_{e_i})$$

Once we have disambiguated the entity, we retrieve the sub-tree from the folksonomy that contains the entity. This process involves tracing back the "categories" on the Wikipedia pages. At the bottom of most Wikipedia pages are categories that users have assigned for the entity on the page. Each category has a name, and links to its category page. In turn, the category page 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). For instance, the Wikipedia page for Theo Walcott[3] contains a box at the bottom of the page with categories such as "People from Stanmore," "English footballers," "England under-21 international footballers," etc. If one clicks the "English footballers" category, the category page contains a list of players, along with parent categories such as "English sportspeople," "Association football players by nationality," etc.

Therefore, in our Twopics approach we start with the set of categories for the given entity, and trace through the links of each category, collecting the parent categories along the way. At the end of this process, we have a "sub-tree" of the folksonomy, rooted at the most specific term (the current entity's categories). We qualify the name sub-tree because building the trees in the manner actually generates sub-trees where deeper levels in our trees actually represent more general categories. As we discuss below, we must account for the fact that our ontology is inverted. Also, note, we empirically chose to go 5 levels deep, as at this point the categories were sufficiently general and vague (e.g., "Games," or "Living People."), and with a branching factor averaging around 20, this provided a large number of categories.

An example of building the sub-tree by walking through the categories of a given entity is shown in Figure 3. In this example, we start with the category "English footballers," from Theo Walcott's page, which produces four categories: {"English Sportspeople," "Association football Players by nationality," "Football in England," and "British footballers"}. Then we show tracing this one step further, by following the categories from "Association football Players by nationality." The bottom of the figure shows the resulting sub-tree.

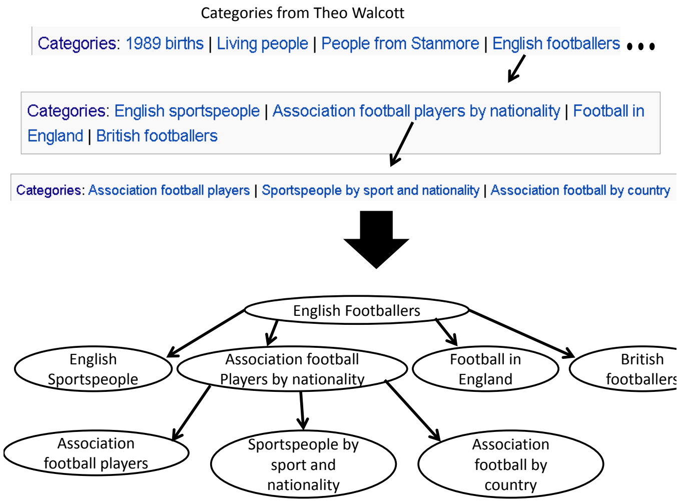

Categories from Theo Walcott
Categories: 1989 births | Living people | People from Stanmore | English footballers • • •
Categories: English sportspeople | Association football players by nationality | Football in England | British footballers
Categories: Association football players | Sportspeople by sport and nationality | Association football by country
English Footballers
English Sportspeople
Association football Players by nationality
Football in England
British footballers
Association football players
Sportspeople by sport and nationality
Association football by country

Figure 3: Part of the Sub-Tree from an entity

As in previous work [17][4] we ignore the categories related to the administration of Wikipedia itself, such as categories with names "Category," "Wikipedia," "Template," etc.

This three step approach generates the set of category sub-trees that contain the disambiguated entities from the Tweets. We note that this is still early work, and we outline some of our future improvements to this approach in Figure 4.

2.2 Discovering User Profiles from Categories

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, then, is to take in this forest of sub-trees and discover the categories (nodes in the trees) that occur frequently and are specific enough to generate useful topic profiles. The key is that the categories are generated from the disambiguated entities, and as such, we can assume that the categories (nodes in the subtrees) already cover the entities in the Tweets. Therefore, if we can find the subset of nodes that occur frequently across all trees generated for all of the Tweets, then we found the subset of categories that covers the entities, which in turn represent a topic profile for the user. That is, these are the categories that represent the topics the Twitter user likes to Tweet about.

[3] http://en.wikipedia.org/wiki/Theo_Walcott

[4] The authors focus on building a taxonomy from Wikipedia

To generate the topic profile, we rank each of the categories, c, in the set of sub-trees according to the following ranking function:

$$Rank(c) = Freq(c) * w_c$$

Where $Freq(c)$ is the frequency of the category's occurrence and w_c is a weight, inverse to the category's level in the sub-tree. That is, we want to give more weight to the roots (which represent specific categories) than the leaves (which are too general) such that a person Tweeting about baseball players and person Tweeting about actors do not get assigned the same topic profile with a category of "People." Therefore, assuming some branching factor b, and given a depth in the tree d for our current category, we define w_c as:[5]

$$w_c = 1/b^d$$

Because our knowledge base is a folksonomy, it may be inconsistent, having categories that occur at various depths, in different parts of the ontology. Therefore, we note that the ranking score of a category is actually the sum of its ranking scores for each depth where it occurs. For instance, if a category occurs 4 times in the 2nd level and 8 times in the 3rd level, its Rank would actually be $4*(1/b^2)+8*(1/b^3)$.

Finally, we define the topic profile as the top-K categories, ranked according to our ranking function.

3. EXPERIMENTAL STUDY

We selected four Twitter users for this study. Two were chosen purposely: the Twitter account of Daily Mirror newspaper writer John Cross, and that of technology blog Gizmodo. Both of these were chosen because they focus on specific topics, which makes them easy to verify. John Cross focuses almost exclusively on football in his Tweets, writing about the World Cup, players, and teams, with an emphasis on English teams (and Arsenal in particular). Gizmodo, as a technology blog, Tweets about gadgets and technology.

We also randomly chose two Twitter users to study the generalization of our method. The Twitter homepage has a selection of users under the "See who's here," heading. These users are both regular users and celebrities, and from these we randomly selected two of the regular users. For privacy, we chose to keep the names of these users anonymous. We then manually read these users Tweets to characterize their profile. Interestingly, the anonymous users are much more varied in what they Tweet about. Anonymous1 writes about schools in the Pacific-10 University sports conference of the USA (schools such as University of California at Los Angeles, USC, etc.) with an emphasis on USC. The user also writes about television shows, and social issues of interest. Anonymous2 focuses on sports in the Chicago area, particularly baseball, with a few posts about hockey. This users also posts frequently about music and bands, with an emphasis on Tweeting about what album, song, and band the user is currently listening to.

The statistics of our experimental data are given in Table 1. We note, as this is early work, we are still in the process of collecting and labeling a much larger experimental data set to study. However, this data set does offer an interesting subset to analyze.

Table 1: Our experimental data

Twitter User	# Collected Tweets
John Cross	280
Gizmodo	599
Anonymous1	340
Anonymous2	180

We ran our Twopics approach over the Tweets for all of these users, and examined the discovered topic profiles for each user. For this analysis, we read through the set of Tweets for each user to gain an understanding about the common categories which they write about. After, for each topic profile, we labeled the categories in that profile as either "relevant" or "not-relevant." This allows us to calculate an average precision for each topic profile as we vary the top-K. We analyzed topic profiles generated from the top-5, top 10, and top 25 ranked categories, and report the average precision@K in Table 2, along with the standard deviation.

Table 2: Average precision at K for categories

Size of K	Avg. Precision (± Std-Dev)
5	0.95 ± 0.10
10	0.90 ± 0.08
25	0.85 ± 0.08

For clarity, Figure 4 shows the topic profile generated from the top-10 ranked categories for each user. The figure shows the relevant categories for the users on a white background, while we gray out the irrelevant categories to make them clear as well.

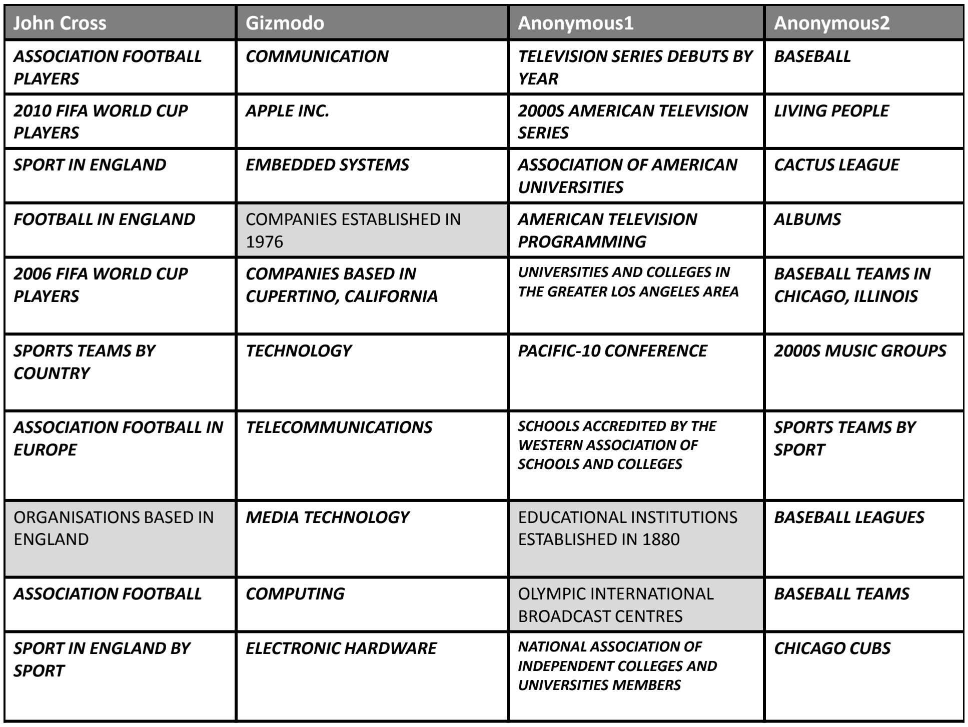
The above demonstrates that even in our early stage, we can generate reasonable topic profiles based on the categories found to represent the Tweets. Next we examine the disambiguation results in detail, as these are a key component of our Twopics algorithm. As labeling and analyzing all discovered named entities from all of the Tweets would be too costly, we instead analyze a subset of Tweets and the named entities within them. For each of our four users, we randomly sampled 50 Tweets and analyzed the named entities in detail. For each entity, we determined whether the disambiguated entity correctly represents the entity implied in the post. In this subset, there were 365 named entities discovered by the system, of which Twopics correctly disambiguated 191, resulting in an accuracy of 52.33%. We compare this to a baseline where Twopics randomly selects one of the candidate entities as the disambiguated entity. The baseline's accuracy is just 5.21%. Therefore, leveraging the context clearly lends some advantage when choosing the disambiguated entity.

In fact, Twopics is able to disambiguate a number of difficult cases. In the running example of this paper, Twopics correctly identified the three soccer related entities Arsenal, Walcott, and Becks (England is easy to disambiguate). Also, in the following Tweet:

```
HAWKS WIN!!!  Stanley Cup is coming to Chicago
```

Twopics correctly identified the "HAWKS" mentioned as the American hockey team, the Chicago Blackhawks. However, there are also a number of errors that leave room for improvement. For instance, in some cases the named en-

[5] Note that we empirically set b to be 20.96, as this was the average branching factor observed for all nodes in all subtrees in our experiments.



John Cross	Gizmodo	Anonymous1	Anonymous2
ASSOCIATION FOOTBALL PLAYERS	***COMMUNICATION***	***TELEVISION SERIES DEBUTS BY YEAR***	***BASEBALL***
2010 FIFA WORLD CUP PLAYERS	***APPLE INC.***	***2000S AMERICAN TELEVISION SERIES***	***LIVING PEOPLE***
SPORT IN ENGLAND	***EMBEDDED SYSTEMS***	***ASSOCIATION OF AMERICAN UNIVERSITIES***	***CACTUS LEAGUE***
FOOTBALL IN ENGLAND	COMPANIES ESTABLISHED IN 1976	***AMERICAN TELEVISION PROGRAMMING***	***ALBUMS***
2006 FIFA WORLD CUP PLAYERS	***COMPANIES BASED IN CUPERTINO, CALIFORNIA***	***UNIVERSITIES AND COLLEGES IN THE GREATER LOS ANGELES AREA***	***BASEBALL TEAMS IN CHICAGO, ILLINOIS***
SPORTS TEAMS BY COUNTRY	***TECHNOLOGY***	***PACIFIC-10 CONFERENCE***	***2000S MUSIC GROUPS***
ASSOCIATION FOOTBALL IN EUROPE	***TELECOMMUNICATIONS***	***SCHOOLS ACCREDITED BY THE WESTERN ASSOCIATION OF SCHOOLS AND COLLEGES***	***SPORTS TEAMS BY SPORT***
ORGANISATIONS BASED IN ENGLAND	***MEDIA TECHNOLOGY***	EDUCATIONAL INSTITUTIONS ESTABLISHED IN 1880	***BASEBALL LEAGUES***
ASSOCIATION FOOTBALL	***COMPUTING***	OLYMPIC INTERNATIONAL BROADCAST CENTRES	***BASEBALL TEAMS***
SPORT IN ENGLAND BY SPORT	***ELECTRONIC HARDWARE***	***NATIONAL ASSOCIATION OF INDEPENDENT COLLEGES AND UNIVERSITIES MEMBERS***	***CHICAGO CUBS***

Figure 4: Twopics: Top-10 Categories

tity recognizer does a poor job at selecting entities. In the Tweeet:

`It is amazing outside today in Chicago! Tomorrow's supposed to be even warmer`

Twopics identified "Tomorrow" as a named entity, and disambiguated it as the song "Tomorrow" by the artist Sean Kingston. An improved recognizer should alleviate this issue. Also, in some cases, there simply is not enough context to exploit (as in Tweets of mostly stopwords). In these cases, Twopics is essentially making a random choice from the candidates and selects the incorrect Wikipedia entry. By forcing some minimum constraints on the context we can alleviate this issue as well.

Our disambiguation results are in line with previous work that leverages Wikipedia to disambiguate queries, where the authors report that a cosine-similarity model averaged an accuracy of 63.8%, while the authors' proposed (supervised) SVM-based approach improved this to an average accuracy of 74.5% [4]. We plan to investigate using this approach in our model in future work. However, this approach is supervised and requires training data. Our approach, in contrast, is unsupervised and allows for ad-hoc matching. Most importantly, even though the precision of our approach is not perfect, it still provides enough disambiguated entities to generate reasonable topic profiles for the users.

4. DISCUSSION

Tweets are short, and there are few users who generate enough of them to create a large enough sized corpus for deep analysis. This sparseness can adversely affect frequency-based topic-models such as Latent Dirichlet Allocation (LDA) [1]. Further, while methods such as LDA (or even TF-IDF based bag-of-words) can give a term-level topic modeling, this may not be appropriate for clustering users and searching users by high-level topic. For instance, a user may want to discover all Twitter users who write about English Football Clubs. However, this concept is ill defined at the term level without a knowledge base to define it. The interested user may never find John Cross's Tweets if he only writes about Arsenal and its players specifically, without mentioning the term "English football club" (which therefore does not appear in an LDA topic). That is, because we leverage an ontology, we are able to generalize our topics to a higher level, based on the instances in the Tweets, rather than relying on the term level. We can group one Twitter user who writes about Fulham and one who writes about Arsenal as both writing about English football clubs, even though neither may have any terms in common in their Tweets. Therefore, we do not propose to directly analyze the Tweet text to discover topics at the term level.

Similarly, we do not leverage "hashtags" to define users by topic. A hashtag is a specially designated word in a Tweet, prefixed with a "#," that represents something of interest to the Twitter community. Other users can then search for specific hashtags, to find any Tweet related to that concept. However, hashtags suffer from the same limitations as using terms directly to define a user's topics of interest. That is, they are often not general enough and do not form an ontology. Table 3 shows the top five most frequent hashtags for our test users (ordered by frequency). Some of them are directly related to a user's interest, but are overly specific. For instance, for John Cross and Anonymous2, the top hashtags are #Arsenal and #Cubs, respectively. Both of these relate to the users' interest, but represent specific football and baseball teams (respectively) rather than the more general categories discovered by our method. Alternatively, there are hashtags that only represent a user's interests in a tangential manner. For instance, Anonymous2's second most frequent hashtag is #nowplaying, which references which music he or she is currently listening to. This does imply that Anonymous2 has an interest in music, but automatically inferring this fact is exceedingly difficult. Another example is the Gizmodo hashtag #photography, which references the blog's pictures of devices, but not the types of devices themselves. Therefore, this hashtag does not actually yield information about the types of devices which represent the topics of interest (e.g., Apple devices). As a final example, consider the hashtags #omgfacts and #spoileralerts from Anonymous1 which provide little insight into this users' topics of interest. Therefore, while hashtags present a compelling mechanism for users to organize and search their Tweets, the information they provide about users' interests are either overly specific or difficult to utilize, because they are at the term level and might be ill defined.

Table 3: Top Five Frequent Hashtags per User

Username	**Hashtags (ordered)**
John Cross	Arsenal, England, wc2010 Spurs, mufc
Gizmodo	iPad, Apple, memoryforever ipadapps, photography
Anonymous1	USC, dadt, glee omgfacts, spoileralert
Anonymous2	Cubs, Nowplaying, Blackhawks Chicago, MLB

Given the problems with analyzing specific terms in the Tweets, we instead analyze the entities in the Tweets, and try to create topics centered around them at the categorical level. This makes an assumption for our method that Tweets contain enough entities to be useful. In our small study, this indeed was the case for our two randomly selected Twitter users (Gizmodo and John Cross were chosen specifically because they represent categories and have entities). In fact, recent work showed that 19% of the Twitter accounts analyzed contained mentions of 50 selected brands [9]. In that study, the authors were looking for a very small subset of brands and still found almost 20% of the Twitter users mentioned them, which is an encouraging result that motivates the use of entities, as we are looking for any entity (not just 50 brands, which is a tiny sample). Another recent large scale study of Twitter found that 85% of the trending topics (most popular topics to Tweet about) were related to headline news, and out of the 41M users in the study, almost 20% of the users posted about these trending topics [12]. We assume that headline news often mentions entities, and again, this provides motivation that users indeed Tweet about entities.

While our early results are encouraging for generating the topic profiles, there are still a number of areas we would like to improve. The first, and most important area is generating a summary of the topic profile. Currently, we define the topic profile as the top-K categories selected for a given user. This is a useful profile for clustering users, searching Twitter users by topic, etc., but it is not satisfactory in terms of human consumption. Ideally, we would want to collapse such profiles into just two or three phrases that really strike at the heart of a Twitter users interests. For instance, John Cross is clearly just interested in Football. The second important area to improve is our study itself. We plan to perform a very large scale empirical study of our method. We will gather many more Tweets from more users, and perform a larger scale analysis which will allow us to more robustly test our method and draw conclusions. Finally, each step of our algorithm can be improved. We want to use a more sophisticated approach for disambiguating the entities, likely focusing on a language model trained on Wikipedia as a corpus (e.g., [4]). Also, the selection of categories from the sub-trees could also be improved, perhaps using other methods for assigning relevance based on the graph structure of the ontology.

5. RELATED WORK

Recently, the ability to discern topics from social media has begun to receive attention as the necessity to search the text and user profiles gains importance. In fact, two recent approaches address different problems using Twitter data.

Chen, et al., [5] 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 to the user. There are a few drawbacks to this approach. First, because a bag-of-words is used, the terms must be very specific. For instance, a Tweet about "Theo Walcott" might not be recommended to a user who writes about the Arsenal Football Club because she never explicitly mentioned Theo. Further, this is a model at the term level, and not truly at the topic level. However, one might replace their topic models with those generated by Twopics for use in their recommendation system.

Another approach to analyzing Twitter that uses topics is TwitterRank, which aims to identify influential microbloggers [22]. This approach leverages LDA by creating a single document from all of a user's Tweets and then discovering the topics by running LDA over this "document." Again, such an approach has the problems of LDA since the Twitter data is sparse, and the generated topics are based on terms rather than concepts.

There is also work that is similar in that the goal is to determine the topics and interests of bloggers by analyzing their blogs [15, 21, 6]. However, blogs are a much richer medium for textual analysis because blog posts are generally much longer than Tweets and usually conform better to the grammatical rules of written English.

Twopics relies heavily on the ability to disambiguate the entities within the Tweets. In fact, disambiguating entities is a (relatively) old problem in natural language processing [7] and there has been previous work on using dictionaries to aid this task (e.g., [23]). However, these approaches are generally statistical (e.g., using co-occurrences) and require large training corpora. Because we lack this corpora, we instead leverage Wikipedia as a knowledge base. Other research has proceeded in this direction, leveraging Wikipedia as the knowledge base for entity disambiguation (and labeling) [11, 13, 14, 8]. We hope to incorporate such methods, though the noisy, ungrammatical nature of Tweets may prove difficult for them. We note a key difference 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. Also, recent work proposed a framework for micro-blogging that includes the capacity to link entities within the post to their disambiguated concept on the Semantic Web [16]. However, this approach relies on the poster to manually annotate the entities in the post. Perhaps an approach such as ours could alleviate this manual linking of the entities, allowing Twopics to do it automatically.

There are other recent, related efforts that leverage knowledge bases for disambiguation. Bunescu, et al. [4] leverage Wikipedia pages to disambiguate queries [4]. In this work, the authors develop an SVM kernel to perform this disambiguation, and run their approach in an supervised setting. This is an interesting extension that we can perhaps use in our approach to disambiguation. Bratus, et al. [3] propose graphical models with outside knowledge (such as domain-specific ontologies) to perform entity extraction and disambiguation from very noisy text. Again, however, this work functions in a supervised setting and requires training data, while our approach does not.

6. CONCLUSION

In this paper we present our early results of Twopics, which aims to discover the topics of interest for Twitter users based on their posts. Our goal is to support clustering and searching of Twitter users based on their topics of interest. Therefore, to support such searching and clustering at the topic level, and to deal with issues of sparseness and noise, we described a knowledge-based approach.

Twopics discovers (and disambiguates) the entities within the posts, and then determines the high-level categories defined by these entities. By analyzing the resulting categories, Twopics can then generate a topic profile for the user. Our early results are encouraging and we hope that our future research in this area will spur even more interest in analyzing Twitter, as the users can be clustered and search by their topics of interest.

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] K. Borau, C. Ullrich, J. Feng, and R. Shen. Microblogging for language learning: Using twitter to train communicative and cultural competence. In *Proc. of the International Conference on Advances in Web Based Learning*, 2009.

[3] S. Bratus, A. Rumshisky, R. Magar, and P. Thompson. Using domain knowledge for ontology-guided entity extraction from noisy, unstructured text data. In *Proc. of the Workshop on Analytics for Noisy Unstructured Text Data (AND)*, 2009.

[4] R. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[5] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proc. of the international conference on Human factors in computing systems*, 2010.

[6] Y. Cheng, G. Qiu, J. Bu, K. Liu, Y. Han, C. Wang, and C. Chen. Model bloggers' interests based on forgetting mechanism. In *Proc. the International Conference on World Wide Web*, 2008.

[7] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[8] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proc. of EMNLP-CoNLL*, 2007.

[9] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, 2009.

[10] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *Proc. of the first workshop on Online social networks*, 2008.

[11] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

[12] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. of the International Conference on World wide web*, 2010.

[13] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proc. of the sixteenth ACM conference on Conference on information and knowledge management*, 2007.

[14] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proc. of the 17th ACM conference on Information and knowledge management*, 2008.

[15] M. Oka, H. Abe, and K. Kato. Extracting topics from weblogs through frequency segments. In *Proc. of the Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[16] A. Passant, T. Hastrup, U. Bojars, and J. Breslin. Microblogging: A semantic and distributed approach. In *Proc. of Workshop on Scripting for the Semantic Web*, 2008.

[17] S. P. Ponzetto and M. Strube. Deriving a large scale taxonomy from wikipedia. In *AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence*, 2007.

[18] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of the International Conference on World wide web*, 2010.

[19] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proc. of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2009.

[20] D. A. Shamma, L. Kennedy, and E. F. Churchill. Tweet the debates: understanding community annotation of uncollected sources. In *Proc. of the first SIGMM workshop on Social media*, 2009.

[21] C.-Y. Teng and H.-H. Chen. Detection of bloggers' interests: Using textual, temporal, and interactive features. In *Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 2006.

[22] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proc. of the third ACM international conference on Web search and data mining*, 2010.

[23] D. Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 14th conference on Computational linguistics*, 1992.