Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Mining User Intents in Twitter: A Semi-Supervised Approach to Inferring Intent Categories for Tweets

Jinpeng Wang[1], Gao Cong[2], Wayne Xin Zhao[3], Xiaoming Li[1]
[1]Department of Computer Science and Technology, Peking University, China,
[2]School of Computer Engineering, Nanyang Technological University, Singapore
[3]School of Information, Renmin University of China, China
JooPoo@pku.edu.cn, gaocong@ntu.edu.sg, batmanfly@gmail.com, lxm@pku.edu.cn

Abstract

In this paper, we propose to study the problem of identifying and classifying tweets into intent categories. For example, a tweet "I wanna buy a new car" indicates the user's intent for buying a car. Identifying such intent tweets will have great commercial value among others. In particular, it is important that we can distinguish different types of intent tweets. We propose to classify intent tweets into six categories, namely Food & Drink, Travel, Career & Education, Goods & Services, Event & Activities and Trifle. We propose a semi-supervised learning approach to categorizing intent tweets into the six categories. We construct a test collection by using a bootstrap method. Our experimental results show that our approach is effective in inferring intent categories for tweets.

Introduction

Posting short messages (i.e., tweets) through microblogging services (e.g., Twitter) has become an indispensable part of the daily life for many users. Users heavily express their needs and desires on Twitter, and tweets have been considered as an important source for mining user intents (Hollerit, Kröll, and Strohmaier 2013; Zhao et al. 2014). For example, the intent tweet "I am planning to travel to New York" explicitly indicates the user's intent. Such intent tweets can be exploited by interested parties, e.g., companies, government, public organizations, etc. For the example intent tweet, it will be more interesting if we categorize it to the *travel* category for marketing services.

In this paper, we study the task of identifying and inferring intent categories for tweets, which will benefit many commercial applications. Based on our analysis on a tweet corpora and the taxonomy of Groupon website, we propose to divide the intent tweets into six categories namely *Food & Drink*, *Travel*, *Career & Education*, *Goods & Services*, *Event & Activities* and *Trifle*. An example intent tweet of *Goods & Services* category is "I want to have a new car" and an example intent tweet of *Event & Activities* category is "I plan to go to swimming". With the inferred intent categories, one can potentially employ Twitter as a market or auction place, where we can find people with particular intents or desires through their tweets.

To the best of our knowledge, there exists no reported study of inferring intent categories for tweets in the context of commerce marketing. Tweets often contain sentences and user intent is often explicitly expressed in tweets. Moreover, tweets often contain more information than queries, e.g., friendship and context. It is reported in (Morris, Teevan, and Panovich 2010) that users prefer social sites over search engines when talking about opinions and recommendations. Note that although only a small portion of tweets (7% in our corpus by randomly sampling tweets) are sent with specific meanings or contain explicit intents, the number of intent tweets is still very huge since the number of tweets is enormous[1]. Nevertheless, the problem of inferring the intent categories of tweets is still challenging. Tweets are very noisy and often contain slang, misspellings, emotions and hashtags. In addition, it is very time-consuming to generate labeled data if we adopt the supervised approach.

To address the challenge, we formulate the problem as a classification problem and propose a graph-based semi-supervised approach to inferring intent categories for tweets. Our approach has two significant merits. (1) *Noise resistant:* to reduce noise in tweets, we propose to use intent-keywords instead of all the words as basic information units which are derived from a unsupervised bootstrap based method. (2) *Weak supervision:* we propose an optimization model which is built on the intent graph with tweets and intent-keywords as nodes. An edge in the intent graph can be established to model the association between two tweets, two intent-keywords, or a tweet and an intent-keyword. With effective information propagation via graph regularization, only a small set of tweets with category labels is needed as the supervised information.

We conduct experiments on a tweet dataset generated by a bootstrap based method. Our experimental results show that the proposed method outperforms several competitive baselines in inferring the categories of intent tweets.

Related Work

Query intent classification The existing work on query intent classification can be roughly divided into two types based on the classification taxonomy employed. The first

[1] Statistics from Twitter show that approximately 500 million tweets are sent every day: http://goo.gl/407UDX.

type is classifying queries according to the query type, such as informational or navigational or transactional (Cao et al. 2009; Kang and Kim 2003); and the other type is classifying queries according to the user intent, such as "jobs" or "travel" (Hu et al. 2009; Shen et al. 2006; Beitzel et al. 2007; Li, Wang, and Acero 2008). The latter task is related to our work, but it differs in several aspects. User intent deduction from queries suffers from the lack of sufficient words due to the lengths of queries. The existing work on query intent classification focuses on expanding features of queries by including external knowledge, such as search snippets from search engine, click-through (Li, Wang, and Acero 2008) and facets from Wikipedia (Hu et al. 2009). In addition, the intents of queries are typically implicit. For example, a keyword query like "xbox" does not explicitly express user intents. In contrast, tweets often contain sentences that explicitly express the user intent. For example, tweets like "I want to buy an xbox." explicitly express the user intent. Our focus in this paper is to identify intent tweets that explicitly express user intents, which is different from query intent classification.

Online commercial intention identification This task is to identify online commercial intention from queries, documents or tweets. Most studies focus on capturing commercial intent by analyzing search queries (Dai et al. 2006; Strohmaier and Kröll 2012) or click-through (Ashkan and Clarke 2009). Chen et al. (2013) aims at identifying intents expressed in posts of forums. Posts differs from tweets in several aspects. First, posts are longer and thus contain more information than tweets. Second, posts in the same forum have similar topics, and this can be exploited for identifying intent posts. Third, the portion of user intention is much larger than that of tweet, e.g., there are many posts that express the intent to buy phones on a digital forum. The most related is the work (Hollerit, Kröll, and Strohmaier 2013), which attempts to detect commercial intent tweets, but does not consider other types of intent tweets as we do in our work. The method by Hollerit, Kröll, and Strohmaier employs traditional classification models like SVM, and uses n-gram and part-of-speech tags as features. This method can be used to detect other types of intent tweets although the work (Hollerit, Kröll, and Strohmaier 2013) focuses on commercial intent tweets. This work can be treated as one of our subtasks, i.e., identifying *Goods & Services* intent.

Problem Statement

Intent Tweet: Inspired by the definition on intent post in discussion forums (Chen et al. 2013), and the definition on commercial intent tweets (Hollerit, Kröll, and Strohmaier 2013), we define a tweet as an *intent tweet* if (1) it contains at least one verb and (2) explicitly describes the user's intent to perform an activity (3) in a recognizable way.

Example 1: Tweet "I want to buy an xbox, if get A in this examination. Bless me!!!" is an intent tweet and it satisfies all the three conditions in the definition. □

In the first part of the definition, the verb is important in exhibiting user intent (Hollerit, Kröll, and Strohmaier 2013), e.g., the verb "want" in Example 1. In the second part, we require that the intent is explicitly described as it is required in previous work (Chen et al. 2013). This is in contrast with implicit intents that need inference or deduction. An example of tweet with an implicit intent is "Anyone knows the battery life of HTC one" and it is difficult to know whether the author was thinking about buying a HTC one. *Recognizable* (Kirsh 1990) here refers to "the ability to make a decision in constant time" and is also used for defining commercial intent (Hollerit, Kröll, and Strohmaier 2013). Note that with this definition, it would make less ambiguity for both annotating and identifying intent tweets.

We further define *intent-indicator* and *intent-phrase*, which are the key elements in our proposed approach.

Intent-Indicator: It comprises a group of terms that are used by users to express their intents. It is a verb or infinitive phrase that immediately follows a subject word, e.g., "I". For example, in tweet "I want to buy an xbox", "want to" is an intent-indicator, indicating the tweet is likely to be an intent tweet. For Part-of-Speech Tagging, we use Tweet NLP [2].

Intent-Keyword: It is a noun, verb, multi-word verb or compound noun (consisting of several nouns) contained in a verb or noun phrase which immediately follows an intent-indictor, e.g., in Example 1, "buy and "xbox which are contained in the phrase "buy an xbox" are intent-keywords. If one intent-keyword contains another intent-keyword, we keep the longer one. Furthermore, a phrase that immediately follows an intent-indicator is referred to as an *intent-phrase*.

Categories of Intent Tweets

Users' intents exhibited in tweets may belong to different categories, and different categories of intents may be of interest to different applications. However, no existing work has attempted to establish the categories for intent tweets.

To establish taxonomy for intent tweets, we have reviewed a large number of tweets and studied the taxonomy of Groupon[3]. The reasons that we refer to the taxonomy of Groupon are: 1) The intent expressed in tweets by Twitter users are usually related to daily life; and 2) Groupon offers deals that cover a wide range of daily life. Finally, we define six types of intent[4]:

- **Food & Drink**: the tweet authors plan to have some food or drink. It corresponds to the *Food & Drink* category of Groupon.
- **Travel intent**: the tweet authors are interested in visiting some specific points of interests or places. This category corresponds to the *Getaways* category of Groupon.
- **Career & Education intent**: the tweet authors want to get a job, get a degree or do something for self-realization. This category is not in Groupon, but we find that a good portion of intent tweets can be categorized into *Career & Education*. This category also appears in Twellow[5] that organizes twitter users into a taxonomy.
- **Goods & Services intent**: the tweet authors are interested in or want to have some non-food/non-drink goods (e.g., car) or

[2] http://www.ark.cs.cmu.edu/TweetNLP/

[3] http://www.groupon.com

[4] Groupon has 8 first-level categories: *Food & Drink*, *Event & Activities*, *Beauty & Spas*, *Health & Fitness*, *Automotive*, *Shopping*, *Apparel* and *Gateways*.

[5] http://www.twellow.com

services (e.g., haircut). This category corresponds to the combination of four categories in Groupon, namely *Beauty & Spas*, *Health & Fitness*, *Automotive*, *Shopping* and *Apparel*. They are combined because they all belong to *Goods & Services* and each of these categories takes only a very small proportion on Twitter.

- **Event & Activities**: the tweet authors want to participate in some activities which do not belong to the aforementioned categories (e.g., concert). This category corresponds to the *Event & Activities* category of Groupon.
- **Trifle intent**: This category of intent tweets talks about daily routine, or some mood trifles (Java et al. 2007).

Data Preparation

Since the proportion of intent tweets is small, we will obtain few intent tweets by sampling the tweets. We do not want to construct an unbalanced test collection with an overwhelming number of non-intent tweets for intent classification. Based on the definitions of intent-indicator and intent-keyword, we propose a novel method to construct the test collection. The idea is that a tweet is more likely to be an intent tweet if it contains an intent-indicator. We adopt the bootstrapping based method (Riloff, Wiebe, and Wilson 2003; Zhao et al. 2014) to retrieve intent tweets.

Specifically, given a seed set of intent-indicators, (e.g., “want to”), (1) we extract the intent-phrases (e.g., “buy an xbox”) that frequently co-occur with intent-indicators, and (2) we use the extracted intent-phrases to extract more intent-indicators. For instance, we extract intent-phrase “buy an xbox” by using intent-indicator “want to” if their co-occurrence frequency is above a certain threshold, and we further use this intent-phrase to extract more intent-indicators like “wanna to”. We repeat these steps until we cannot extract more intent-indicators and intent-phrases. Finally (3) tweets which contain these extracted intent-indicators are kept in our test collection for manual annotation. In this paper, we focus on the intents of the tweet authors. We also discard tweets that express the negative intents by filtering out those that contain negative words (e.g., “don’t wanna”).

We use the Twitter data (Kwak et al. 2010), which spanned the second half of 2009. We first identify potential intent tweets by using the above bootstrapping based method. We randomly sample 3,000 potential intent tweets, and two annotators with experiences of using Twitter are employed to annotate the category label[6]. For each tweet, each annotator first determines whether it is an intent tweet according to our definition above. If yes, the annotator further labels its intent category according to the description on the six intent categories above. Finally, we get 1,599 intent tweets and 531 non-intent tweets with the same label by the two annotators. In other words, we discard 870 tweets with inconsistent annotations. The Cohen’s Kappa coefficient between the two annotators is 65.35%, which is still a high value. We summarize the statistics of this dataset in Table 1. Note that 75.08% of these tweets identified by the bootstrap method are intent tweets, which indicates that the bootstrap based method is indeed an effective unsupervised approach to retrieving intent tweets.

Table 1: Statistics and examples of intent categories in our test collection.

Category	# (%)	Example
Food & Drink	245 (11.50%)	`hungry...i need a salad......four more days to the BEYONCE CONCERT...`
Travel	187 (8.78%)	`I need a vacation really bad. I need a trip to Disneyland!`
Career & Education	159 (7.46%)	`this makes me want to be a lawyer RT @someuser new favorite line from an ...`
Goods & Services	251 (11.78%)	`mhmmm, i wannna a new phoneeee. ... i have to go to the hospital. ...`
Event & Activities	321 (15.07%)	`on my way to go swimming with the twoon @someuser; i love her so muchhhhh!`
Trifle	436 (20.47%)	`I’m so happy that I get to take a shower with myself. :D`
Non-intent	531 (24.92%)	`So sad that Ronaldo will be leaving these shores...http://URL`

Our methods are built on *intent-keywords* as the information units. Thus, each tweet in the test collection is kept with the annotations of intent-keywords identified by the unsupervised bootstrap method.

Inferring Intent Categories

Given a set of intent tweets and a set of intent-keywords, and a small number of labeled tweets as the input, this task is to infer the intent categories for each tweet in the set of intent tweets. We first construct an intent-graph. The intent-graph models associations between intent tweets and their intent-keywords. Based on the intent-graph, we formulate the problem of inferring intent categories from a small number of labeled tweets as an optimization problem. Note that our approach derives intent-keywords using the unsupervised bootstrap method without relying on any dictionary for intent-keywords. This is desirable since tweets are often written in informal language and a dictionary of intent-keywords is not available.

Intent-graph

We construct an intent-graph as illustrated in Figure 1 from the input data to characterize the relations among tweets and intent-keywords. In Figure 1, there are two types of nodes, namely intent tweet nodes and intent-keyword nodes. The tweet nodes associated with labels (e.g., “food”) are labeled intent tweets, and the tweet nodes associated with “?” are unlabeled intent tweets. There are three types of edges in the intent-graph, i.e., edges between tweets and keywords (black edges), edges between tweets (green edges), and edges between keywords (blue edges). Specifically, we establish an edge between a tweet node and a keyword node if the keyword is contained in the tweet; we establish an edge between two keyword nodes if they co-occur in the same tweet; we establish an edge between two tweet nodes if their intent-keywords share common words. We will present how to compute weight for each edge in the following.

Formulation

We introduce the notations to be used and formally define our task. We denote by $\mathcal{X} = \{x_1, \cdots, x_{|\mathcal{T}|}, \cdots, x_{|\mathcal{T}|+|\mathcal{W}|}\}$

[6] The annotated data set is available at http://joopoo.github.io

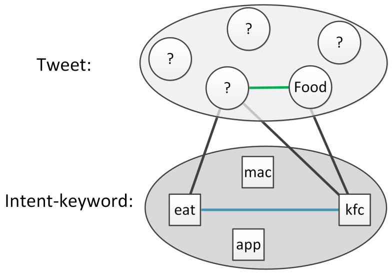

Figure 1: Example intent graph.

the set of nodes in intent-graph, where the first $|\mathcal{T}|$ nodes represent intent tweets and last $|\mathcal{W}|$ nodes are intent-keywords. We assume that these input tweets have been preprocessed in order to obtain a relatively balanced distribution of intent tweets and non-intent tweets. Let the first l nodes $\mathcal{X}^l = \{x_1, ..., x_l\}$ be the labeled intent tweets w.l.o.g., and the remaining nodes $(\mathcal{X} \setminus \mathcal{X}^l)$ are either unlabeled tweet nodes or intent-keyword nodes. In our problem setting, there are only a few labeled instances but many unlabeled instances, i.e. $l \ll |\mathcal{T}|$. Let $\mathcal{C}$ be the set of intent categories (six intent categories in our case). Each labeled tweet node x_i is associated with a vector of $|\mathcal{C}|$ elements, which represent the ground truth category of the tweet. Each element $\acute{y}_i^c$ in the vector represents whether x_i belongs to category c: If x_i has label c, then $\acute{y}_i^c = 1$; otherwise, $\acute{y}_i^c = 0$. Similarly, each node x_i in $\mathcal{X}$ is associated with a vector of $|\mathcal{C}|$ elements, which represent the estimated confidence scores that x_i belongs to each category c in $\mathcal{C}$. Each element f_i^c in the vector represents the confidence score of node x_i belonging to category c estimated by our proposed method.

The problem of inferring categories for unlabeled intent tweets is transformed into estimating f_i^c for each $c \in \mathcal{C}$ and each unlabeled intent tweet t_i. Then the category with the highest intent score for each unlabeled tweet node t_i is chosen as the inferred category, i.e., $\hat{c} = \mathrm{argmax}_{c \in \mathcal{C}} f_i^c$.

The above formulation assumes that the input tweet itself is an intent tweet, which can be classified into one of the six intent categories in Table 1. Not all the tweets are intent tweets, we take a relatively simple but effective method to identify non-intent tweets: if the values for all these elements in the vector f_i^c are smaller than a predefined threshold η, which is set by cross-validation.

Optimization Model

Since we have only a few labeled data, it will be infeasible to train reliable models by only using these labeled data. Our idea is to leverage the association between nodes and propagate the evidence of intent categories via the intent graph. Intuitively, a node should have a similar category label with its neighboring nodes according to the manifold assumption proposed in (Belkin, Matveeva, and Niyogi 2004; Zhu et al. 2003; Wang and Zhang 2006). Besides, we can incorporate the available label information as the supervised information. We adopt the regularization based method to model the association between nodes and incorporate the supervised information. Specially, in our regulation function, we model three types of associations and one penalty function.

Association between tweets and intent-keywords: On the intent graph, an undirected edge (i, j) exists if the i^{th} tweet contains the j^{th} intent-keyword. If tweet t contains intent-keyword w, the confidence scores of t and w should be similar with each other on every category. Note that for each intent-keyword node w, we also model its confidence score of belonging to each category. We model this as follows [7],

$$R_1 = \sum_{t \in \mathcal{T}} \sum_{w \in \mathcal{W}} s(t, w)(f_t^c - f_w^c)^2, \qquad (1)$$

where $s(t, w)$ is the similarity between tweet t and intent-keyword w and $s(t, w)$ is computed by $s(t, w) = \frac{n_{t,w}}{n_t} \times \log \frac{n_w}{|\mathcal{T}|}$, where $n_{t,w}$ is the frequency of word w in tweet t, n_t is the number of intent-keywords in t, n_w is the number of intent tweets that contain intent-keyword w. In the second factor $\log \frac{n_w}{|\mathcal{T}|}$, w is an intent-keyword, which is useful for determining the category of an intent tweet. Unlike the idea of idf in information retrieval, we assume a more frequent intent-keyword should be assigned with a larger weight.

Association between two tweets: If tweet t_1 is similar to tweet t_2, the confidence scores of these two tweets should be similar on every category. We model this as follows,

$$R_2 = \sum_{t_1 \in \mathcal{T}} \sum_{t_2 \in \mathcal{T}} s(t_1, t_2)(f_{t_1}^c - f_{t_2}^c)^2, \qquad (2)$$

where $s(t_1, t_2)$ is the similarity between t_1 and t_2. Each tweet is represented as a vector where each of its elements corresponds to a distinct intent-keyword. We use the cosine similarity to compute the similarity of two tweets. $s(t_1, t_2) = \sum_{w \in \mathcal{W}} \frac{n_{t_1,w} n_{t_2,w}}{\sqrt{(\sum_{w'} n_{t_1,w'}^2)(\sum_{w'} n_{t_2,w'}^2)}}$ where $n_{t_1,w}$ and $n_{t_2,w}$ denote the frequencies of word w in t_1 and t_2, respectively.

Association between two intent-keywords: If intent-keyword w_1 is used in a similar way as is intent-keyword w_2, i.e., they co-occur in many tweets, then their confidence scores on every category should be similar. We model this as follows,

$$R_3 = \sum_{w_1 \in \mathcal{W}} \sum_{w_2 \in \mathcal{W}} s(w_1, w_2)(f_{w_1}^c - f_{w_2}^c)^2, \qquad (3)$$

where $s(w_1, w_2)$ is the similarity between w_1 and w_2. We represent each word as a vector where each of its elements corresponds to an intent tweet, and then we use the cosine similarity to compute their similarity $s(w_1, w_2) = \sum_{t \in \mathcal{T}} \frac{n_{t,w_1} n_{t,w_2}}{\sqrt{(\sum_{t'} n_{t',w_1}^2)(\sum_{t'} n_{t',w_2}^2)}}$, where n_{t,w_1} and n_{t,w_2} denote the frequencies of word w_1 and w_2 in t, respectively.

Penalty Function - Integrating Labeled Data: We have a small number of labeled intent tweets, which are the supervised information. We incorporate the supervised information in our optimization model as follows,

[7] For the convenience of parameter estimation, we use the squared difference to measure the difference between two probability distributions.

$$R_4 = \sum_{t \in \mathcal{X}^l} (f_t^c - \acute{y}_t^c)^2, \tag{4}$$

where $\acute{y}_t^c$ is the label of the intent tweet t on category c.

Final Model The regularization factors and penalty function can be enforced through the terms of the objective function in the following minimization problem where $\mathbf{f}^c = [f_1^c, ..., f_{|\mathcal{X}|}^c]^\top$,

$$\begin{aligned} \hat{\mathbf{f}}^c &= \underset{\mathbf{f}^c}{\operatorname{argmin}} \sum_{t \in \mathcal{X}^l} (f_t^c - \acute{y}_t^c)^2 \\ &+ \gamma\Big(\lambda_1 \sum_{t \in \mathcal{T}} \sum_{w \in \mathcal{W}} s(t, w)(f_t^c - f_w^c)^2 \\ &+ \lambda_2 \sum_{t_1 \in \mathcal{T}} \sum_{t_2 \in \mathcal{T}} s(t_1, t_2)(f_{t_1}^c - f_{t_2}^c)^2 \\ &+ \lambda_3 \sum_{w_1 \in \mathcal{W}} \sum_{w_2 \in \mathcal{W}} s(w_1, w_2)(f_{w_1}^c - f_{w_2}^c)^2\Big) \end{aligned} \tag{5}$$

where $\lambda_1, \lambda_2, \lambda_3 > 0$, and $\lambda_1 + \lambda_2 + \lambda_3 = 1$; the three parameters control the effect of different kinds of edge. The parameter $\gamma \geq 0$ represents the influence of each source of learning (adjacent nodes vs. labeled nodes) on the intent score of f_i^c. Equation 5 can be rewritten as follows,

$$\hat{\mathbf{f}}^c = \underset{\mathbf{f}^c}{\operatorname{argmin}} (\mathbf{f}^c - \acute{\mathbf{y}}^c)^\top \mathbf{I}_l (\mathbf{f}^c - \acute{\mathbf{y}}^c) + \gamma(\mathbf{f}^c)^\top \mathbf{\Delta} \mathbf{f}^c \tag{6}$$

where $\mathbf{I}_l$ is a diagonal matrix which the first l diagonal elements (corresponding to the first l labeled tweets) are all 1 and other diagonal elements are all 0, $\acute{\mathbf{y}}^c = [y_1^c ..., y_{|\mathcal{X}|}^c]^\top$, and $\mathbf{\Delta}$ is the graph Laplacian matrix. Here the first $|\mathcal{X}^l|$ elements of $\acute{\mathbf{y}}^c$ are defined earlier, which represent the ground truth categories of labeled tweets; the remaining elements of $\acute{\mathbf{y}}^c$ are set as 0. We have $\mathbf{\Delta} = \mathbf{A} - \mathbf{S}$, where $\mathbf{S}$ is a $|\mathcal{X}| \times |\mathcal{X}|$ matrix of weighted "edge" weights (i.e., similarities) and

$$S_{ij} = \begin{cases} \lambda_1 \cdot \gamma \cdot s(t_i, w_j), & i \in \mathcal{T}, j \in \mathcal{W} \\ \lambda_1 \cdot \gamma \cdot s(w_i, t_j), & i \in \mathcal{W}, j \in \mathcal{T} \\ 2 \cdot \lambda_2 \cdot \gamma \cdot s(t_i, t_j), & i \in \mathcal{T}, j \in \mathcal{T} \\ 2 \cdot \lambda_3 \cdot \gamma \cdot s(w_i, w_j), & i \in \mathcal{W}, j \in \mathcal{W}. \end{cases}$$

$\mathbf{A}$ is a $|\mathcal{X}| \times |\mathcal{X}|$ diagonal matrix derived from $\mathbf{S}$ as $\mathbf{A}_{ii} = \sum_{j=1}^{|\mathcal{X}|} \mathbf{S}_{ij}$. Then the solution of Eq. 6 can be obtained as follows (see (Belkin, Matveeva, and Niyogi 2004) for details),

$$\hat{\mathbf{f}}^c = (\mathbf{I}_l + \gamma\mathbf{\Delta})^{-1} \mathbf{I}_l \acute{\mathbf{y}}^c \tag{7}$$

Because $(\mathbf{f}^c)^\top \mathbf{\Delta} \mathbf{f}^c > 0$, $\mathbf{\Delta}$ is a symmetric and positive semi-definite matrix. Consequently the above solution is the unique answer to our optimization problem. The solution can be obtained by an efficient power iterative method[8]. With very large amount of tweets, we can consider splitting tweets into small clusters and build intent-graph independently for each cluster.

Experiments

In this section, we evaluate the performance of intent category inference. We used the test collection in Table 1.

Methods

We compare the following methods in our experiments:

- **SVM-Multi**: It is the "one-versus-all" SVM, where a single classifier is trained per class to distinguish that class from all the other classes. We use each binary SVM classifier to predict, and choose the prediction with the highest confidence score. We use bag-of-words of tweets as the features for building classifiers.
- **Hollerit's Method**: It uses the same classifier as does SVM-Multi (Hollerit, Kröll, and Strohmaier 2013). But it uses n-grams and part-of-speech as features.
- **Velikovich's Method**: It is a graph propagation algorithm which is proposed in (Velikovich et al. 2010). The confidence score of a tweet belonging to a category is computed as the sum over the maximum weighted path from every labeled node of the category to the tweet.
- **Hassan's Method**: It is a graph propagation algorithm proposed in (Hassan and Radev 2010). The confidence score of a tweet belonging to a category is computed as the excepted number of steps from the tweet node to the labeled nodes.
- **Ours**: Our semi-supervised approach.

To tune the parameters in the various methods, we first randomly sample 50 instances for each category from all the labeled data, denoted as $\mathcal{D}_A$, and the rest labeled data is denoted as $\mathcal{D}_B$. Then we use $\mathcal{D}_A$ for parameter tuning with five-fold cross-validation. For the rest experiments, each method uses the corresponding optimal parameters derived by cross-validation. Then, we randomly sample 10 labeled instances for each category from $\mathcal{D}_A$ as training data, and use $\mathcal{D}_B$ as test data. Such experiments were performed ten runs, and the average of ten runs is reported.

Metrics

To measure the average performance of a classifier over multiple categories, the macro-average and the micro-average are widely used. The macro-average weights equally all the categories. The micro-average weights equally individual tweets, thus favoring the performance of larger classes.

Overall Performance

In this set of experiments, we set the number of labeled instances at ten for each of the six categories, and the rest of labeled instances are used for testing. Our method and two graph propagation baselines rely on graphs built with textual similarities between tweets, between words, or between a word and a tweet.

Table 2 present the F1 results of individual categories, as well as Macro-F1 and Micro-F1 over the six categories. We can see that our method performs much better than the four baselines in terms of both Macro-F1 and Micro-F1. In particular, it yields large improvements over the best baseline (i.e., *Velikovich's* method), by 20.30% and 25.25%, respectively, for Macro-F1 and Micro-F1. For individual categories, our method achieves the best performance in all categories except *Trifle*. Inferring tweets for this category is difficult, because some tweets in *Trifle* are similar to tweets in

[8] http://goo.gl/rj7rKf

Table 2: The F1 scores on all categories (with 10 labeled instances per category as training data).

Category	Food	Travel	Self	Goods	Event	Trifle	Non-intent	Marco-F1	Micro-F1
SVM-Multi	38.89%	52.70%	37.74%	28.66%	16.75%	20.16%	32.53%	32.49%	33.11%
Hollerit's	45.35%	50.62%	34.78%	30.70%	23.25%	**21.48%**	14.21%	31.48%	31.37%
Velikovich's	44.89%	49.56%	45.21%	29.72%	26.36%	19.79%	21.38%	33.84%	33.70%
Hassan's	28.40%	18.05%	43.07%	24.27%	20.10%	16.05%	31.74%	25.96%	26.46%
Ours	**54.63%**	**58.64%**	**45.73%**	**43.25%**	**27.13%**	20.04%	**35.56%**	**40.71%**	**42.21%**

Travel and *Event & Activities*. All methods perform poorly on this category. For the *non-intent* category, our method also performs best by using the effective threshold filtering method described in the section of Formulation.

We next compare the performance of four baselines. We can see that *SVM-Multi* is a robust method, which achieves better results than *Hollerit's* method. For the two graph propagation algorithms, *Velikovich's* method performs the best while *Hassan's* method is the worst among these four baselines. Their main difference is that they adopt different evaluation methods to estimate the relationship between unlabeled tweets and labeled nodes—*Velikovich's* method uses weighted path while *Hassan's* method uses excepted steps. The excepted steps may be too coarse to estimate the relationship between unlabeled tweets and labeled nodes.

Varying the Number of Labeled Instances

This set of experiments is to study how the number of labeled instances affects the classification accuracy. We vary the number of labeled instances in a category from 5 to 50. Figure 2 shows the performance of all methods in terms of Macro-F1, and our method is consistently better than other baselines with different numbers of labeled instances.

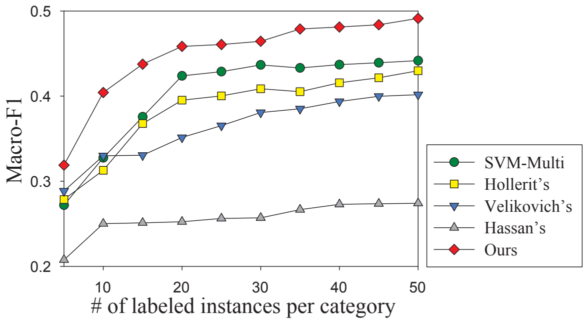

Figure 2: The overall performance of intent category inference in terms of Macro-F1.

Table 3: Performance comparison with different types of words in terms of Macro-F1.

Method	Velikovich's	Hassan's	Ours
All words	20.16%	18.06%	30.32%
Intent-keywords	**33.84%**	**25.96%**	**40.71%**

Table 4: Top five intent-keywords for every intent category.

Food.	Travel	Career.	Goods.	Event.	Trifle
snack	quay	hire	pen	movie	space
nestea	travel	student	kitten	soccer	sex
eat	train	lawyer	t-shirts	songs	texting
pudding	august	incomes	shelf	bruno	buddies
corn	houston	fighter	print	television	tweet

Intent-keywords or Bag-of-words?

In our earlier experiments, we only consider intent-keywords for computing textual similarities but not all the words. This set of experiments is to evaluate whether intent-keywords can better capture intent-oriented associativity between textual units (e.g., a tweet or a word) than all the words. Thus, we use our method and two graph propagation baselines as testing methods as they compute the textual similarity. We consider two ways to compute the tweet-tweet similarity and tweet-word similarity: either using all the words or using only intent-keywords.

As shown in Table 3, the performance of using intent-keywords is consistently better than that of using all words for these three methods. This is because tweets are usually noisy and not all the tokens in a tweet are related to intents. In contrast, intent-keywords are more discriminative features to infer intent categories. Having observed the effectiveness of intent-keywords, we present top intent-keywords for every intent category in Table 4, where are meaningful.

Parameter Analysis

The first parameter that requires tuning for our model is the trade-off coefficient γ in Eq. 5. We examine F1 scores of our model by varying γ from 0.1 to 2.0 with a gap of 0.1. We found that γ = 0.7 gives the best performance in terms of Macro-F1 and Micro-F1. For $\lambda_1, \lambda_2, \lambda_3$ in Eq. 5, we did a line search to find an optimal set of parameters and our results reveal that simply setting $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$ gives good performance which is only slightly worse than what can be obtained by the optimal parameters. Hence, for simplicity, we set all λs to be 1/3 in our experiments.

Conclusions

In this paper, we propose to study the problem of inferring intent categories for tweets. We formulate the problem as a classification problem, and propose a graph-based semi-supervised approach to solve this problem. Our experimental results on a tweet dataset demonstrate the effectiveness of the proposed approach in labeling the categories of intent tweets.

Acknowledgments

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Prime Minister's Office, Singapore, under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office. This research was also supported by the National Key Basic Research Program (973 Program) of China under grant No. 2014CB340400, NSFC Grant 61272340 and 61472013. Xin Zhao was partially supported by the pilot project under Baidu open cloud service platform on the special topic "mining users' commercial intents from online social media data".

References

Ashkan, A., and Clarke, C. L. 2009. Term-based commercial intent analysis. In *SIGIR*, 800–801.

Beitzel, S. M.; Jensen, E. C.; Lewis, D. D.; Chowdhury, A.; and Frieder, O. 2007. Automatic classification of web queries using very large unlabeled query logs. *ACM Trans. Inf. Syst.* 25(2).

Belkin, M.; Matveeva, I.; and Niyogi, P. 2004. Regularization and semi-supervised learning on large graphs. In *In COLT*, 624–638.

Cao, H.; Hu, D. H.; Shen, D.; Jiang, D.; Sun, J.-T.; Chen, E.; and Yang, Q. 2009. Context-aware query classification. In *SIGIR*, 3–10.

Chen, Z.; Liu, B.; Hsu, M.; Castellanos, M.; and Ghosh, R. 2013. Identifying intention posts in discussion forums. In *NAACL-HLT*, 1041–1050.

Dai, H. K.; Zhao, L.; Nie, Z.; Wen, J.-R.; Wang, L.; and Li, Y. 2006. Detecting online commercial intention (oci). In *WWW*, 829–837.

Hassan, A., and Radev, D. 2010. Identifying text polarity using random walks. In *ACL*, 395–403.

Hollerit, B.; Kröll, M.; and Strohmaier, M. 2013. Towards linking buyers and sellers: Detecting commercial intent on twitter. In *WWW*, 629–632.

Hu, J.; Wang, G.; Lochovsky, F.; Sun, J.-t.; and Chen, Z. 2009. Understanding user's query intent with wikipedia. In *WWW*, 471–480.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why we twitter: Understanding microblogging usage and communities. In *WebKDD/SNA-KDD*, 56–65.

Kang, I.-H., and Kim, G. 2003. Query type classification for web document retrieval. In *SIGIR*, 64–71.

Kirsh, D. 1990. When is information explicitly represented. *The Vancouver studies in cognitive science* 340–365.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *WWW*, 591–600.

Li, X.; Wang, Y.-Y.; and Acero, A. 2008. Learning query intent from regularized click graphs. In *SIGIR*, 339–346.

Morris, M. R.; Teevan, J.; and Panovich, K. 2010. What do people ask their social networks, and why?: A survey study of status message q&a behavior. In *CHI*, 1739–1748.

Riloff, E.; Wiebe, J.; and Wilson, T. 2003. Learning subjective nouns using extraction pattern bootstrapping. CONLL, 25–32.

Shen, D.; Sun, J.-T.; Yang, Q.; and Chen, Z. 2006. Building bridges for web query classification. In *SIGIR*, 131–138.

Strohmaier, M., and Kröll, M. 2012. Acquiring knowledge about human goals from search query logs. *IP&M* 48(1):63–82.

Velikovich, L.; Blair-Goldensohn, S.; Hannan, K.; and McDonald, R. 2010. The viability of web-derived polarity lexicons. In *ACL*, 777–785.

Wang, F., and Zhang, C. 2006. Label propagation through linear neighborhoods. In *ICML*, 985–992.

Zhao, X. W.; Guo, Y.; He, Y.; Jiang, H.; Wu, Y.; and Li, X. 2014. We know what you want to buy: A demographic-based system for product recommendation on microblogs. In *KDD*, 1935–1944.

Zhu, X.; Ghahramani, Z.; Lafferty, J.; et al. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, 912–919.