A Novel Approach of Identifying User Intents in Microblog*

ChenXing Li, YaJun Du, Jia Liu, Hao Zheng and SiDa Wang

School of Computer and Software Engineering, Xihua University, Chengdu 610039, China

E-mail: 272463637@qq.com

Abstract. Social micro-blogging platforms facilitate the emergence of citizen's needs and desires which reflect a variety of intents ranging from daily life (e.g., food and drink) to leisure life (e.g., travel and physical exercise). Identifying user intents in microblog and distinguishing different types of intents are significant. In this paper, we propose a novel approach to classify user intents into three categories, namely Travel, Food&Drink and Physical Exercise. Our method exploits Wikipedia concepts as the intent representation space, thus, each intent category is represented as a set of Wikipedia concepts. The user intents can be identified through mapping the microblogs into the Wikipedia representation space. Moreover, we develop a Collaborative User Model, which exploits the user's social connections to obtain a comprehensive account of user intents. The quantitative evaluations are conducted in comparison with state-of-the-art baselines, and the experimental results show that our method outperforms baselines in each intent category.

Keywords: User intents, Microblog, Wikipedia, Collaborative User Modeling

1 Introduction

With the explosive development of social network, micro-blogging services have become a popular platform for people to express their needs and desires. For example, the microblog "I want to take part in physical exercise to lose weight" explicitly indicates that the user has the intent to do some physical exercises. If the intent is identified accurately, information providers can push related advertisement to the user and recommend relevant microblogs as well as users with similar intents to the user.

Intent stands for a purposeful action. We perform intent behaviors every day from querying a search engine to buying a smartphone. A large number of studies focus on identifying the query intent in Web[1], [2], which can be classified to navigational, informational and transactional intent, however the research of identifying user intents in microblog domain is much different. Microblogs often contain sentences which often explicitly express user intents. Moreover, microblogs often contain more infor-

* Project supported by the National Nature Science Foundation of China (No.61271413,61472329,61532009). Innovation Fund of Postgraduate, Xihua University.

mation than queries, e.g., friendship and context [9]. We exploit a classification form for identifying specific intent classes of microblogs.

We meet with two key challenges in the intent classification of microblogs. One is that the microblogs cover diverse intents, then how to define the space of intent representation that can precisely identify the intent of the content is critical. Moreover, we need to clarify the semantic boundary of the intent domain so that the intent classifier can accurately detect whether the microblog falls under the specific domain. Another is that the social interaction is an important issue in intent classification and it requires a careful selection of relevant social features.

For the first challenge, we utilize the Wikipedia concepts as the space of intent representation. All of the relevant concepts for specific intent could construct the Wikipedia Link Graph where each concept will be assigned an intent score by using random walk algorithm. The semantic boundary of the intent domain will be clearly identified by the intent scores of the Wikipedia concepts. Moreover, we exploit explicit semantic analysis (ESA)[5] for the words that are not contained by the Wikipedia concepts. For the second challenge, we construct the Collaborative User Model leveraging the interactions (e.g., mentions and re-microblogs) to give a comprehensive account of user intents. At last, we demonstrate the effectiveness of our method in three applications, namely Travel, Food&Drink and Physical Exercise. The numerous evaluations in comparison with state-of-the-art baselines show that our method outperforms other methods in each intent domain. The rest of the paper is organized as follows. We review related work in Section 2, Section 3 presents our method, Section 4 describes our experiments and the conclusion and future work are in Section 5.

2 Related work

2.1 Query intent identification

Researchers have designed approaches to mine intent in queries using data from user search logs, including clicks, click sequence graphs and query terms. The query intent can be broadly categorized to navigational, informational and transactional intent[1], [2], [3], [4]. However query is different from microblogs. Query is too short to obtain effective characteristics for intent identification. In addition, the intents of queries are typically implicit. For example, a keyword query like “laptop” does not explicitly express user intents. Conversely, microblogs often contain sentences that explicitly express the user intent. For example, the microblog “I want to buy a laptop.” explicitly expresses the user intent. Our focus in this paper is to identify the intent of microblogs that explicitly express user intents, which is different from query intent classification.

2.2 Online intention identification

Prior research on online intention classification mainly focused on commercial and crisis domains. Dai et al. [6], who first proposed the Online Commercial Intention (OCI), presented the framework of building machine learning models to learn OCI

based on Web page content and search queries. Chen et al. [7] aimed at identifying intents expressed in posts of forums. Hollerit et al. [8] firstly defined the commercial intent(CI) and employed traditional classification models like Naïve Bayes with n-gram and part-of-speech tags as features to classify tweets. The most related is the work [9], Wang et al. proposed a semi-supervised learning approach to classify intent tweets into six categories. However they did not take the interactions between users into consideration like the work we do in this paper.

3 Our method

In this paper, we firstly construct the Wikipedia Link Graph ultilizing Wikipedia concepts to represent the intent space, then the Random Walk are performed on the graph to obtain an intent score(i.e. intent probability) for each concept. For the words that are not covered by the Wikipedia concepts, we exploit Explicit Semantic Analysis (ESA)[5] to obtain the most related Wikipedia concepts for the words. At last, the Collaborative User Model exploiting the interactions between users is constructed to further identify user intents in microblog.

3.1 Wikipedia Link Graph Construction

For each intent, we firstly choose several most representative words for specific intents and then search them in the Wikipedia. Considering the structure of the Wikipedia, for each word, we will get an article describing the word in detail and containing a large number of the words which are associated with the specific word. Through browsing these concepts it contains, the category it belongs as well as its sibling concepts, we can easily get enough concepts to cover the specific intent domain. For example, if we want to acquire the concepts about the “Physical Exercise”, the first step would be searching the query “Physical Exercise” in the Wikipedia. Through browsing the corresponding articles, the concepts such as “aerobic exercise”, “body-weight exercise”, “walking” would be collected. Moreover, we can obtain more sibling concepts such as “Swimming” and “Yoga”.

With the collection of the concepts we could build the Wikipedia Link Graph $G = (V, E)$ where V represents the collected concepts while E presents the link relations within concepts. Only when two concepts include each another then there is one edge between them. Conversely, there is no edge between them. Based on the graph G, we can construct a weight matrix $\mathbf{W}$ where the element w_{ij} equals the link count connecting vertices between v_i and v_j in the matrix. The weight matrix $\mathbf{W}$ is symmetric due to $w_{ij} = w_{ji}$.

3.2 Random Walk on the Wikipedia Link Graph

Taking Fig. 1 and Fig. 2 as examples to represent random walk algorithm. The four labeled concepts “Yoga”, “Category: Physical exercise”, “Aerobic exercise”, and “Bodyweight exercise” are seeds for the Physical Exercise intent. For the reason that Wikipedia concepts, which link to each other through article or category links, often share similar topics, we assume that their immediate neighbors also have the same kind of intent to some extent [5]. Starting from a seed concept, it moves to it’s neigborhood node j with probability P_{ij} after the first step. The walk continues until it converges to a stable state, and all the concepts in the graph have probabilities that they belong to Physical Exercise intent. In this example, “Running”, “Walking”, “Plank”, and “Muscle” have a higher Physical Exercise intent probability after first step.

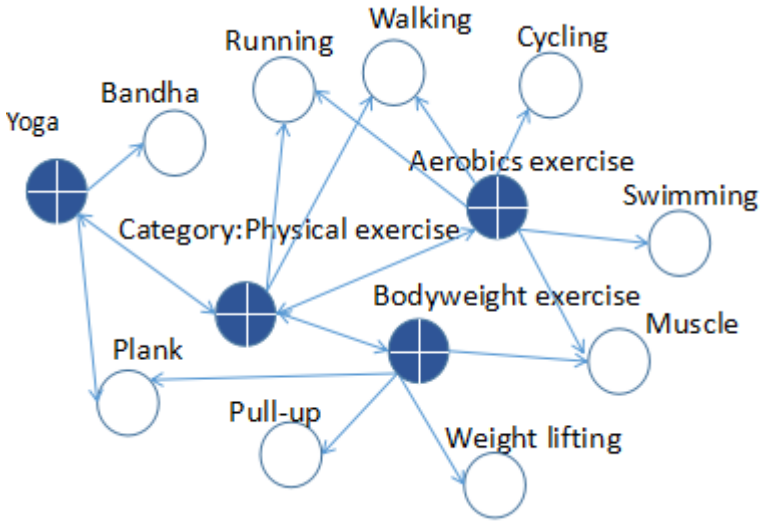

Fig. 1. A link graph where blue nodes represent seed concepts labeled +.

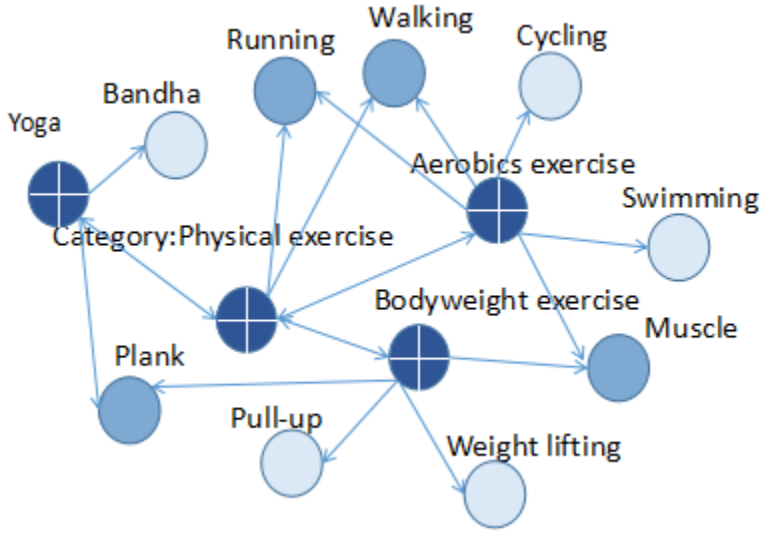

Fig. 2. Label information is propagated from the seed concepts to unlabeled concepts (the darkness of blue denotes the intent probability).

We define transition probabilities $P_{t+1|t}(v_k \mid v_i)$ from the vertex v_i at time t to $v_k (v_i, v_k \in V)$ at time $t+1$ by normalizing the score out of node v_i, and it can be represented as:

$$P_{t+1|t}(v_k \mid v_i) = w_{ik} / \sum_j w_{ij} \ . \tag{1}$$

where j ranges over all vertices connecting to v_i. The score w_{ik} is symmetric, but the transition probabilities $P_{t+1|t}(w_k \mid w_i)$ generally are not because of the normalization varies across nodes. We rewrite the one-step transition probabilities in a matrix form as $\mathbf{P} = [P_{t+1|t}(w_k \mid w_i)]_{ik}$. The matrix $\mathbf{P}$ is row stochastic so that rows sum to 1.

Based on the concepts of weight matrix $\mathbf{W}$, we can select a small set of seed concepts as positive examples and it can be denoted as $S (S \subseteq V)$. The concepts in S are labeled as +1 while the rest are assigned zero. After that, we initiate an intent label vector $v_0 = (p(v_i))_{i=1}^m$ (where m is the total number of vertexes) with values:

$$p(v_i) = \begin{cases} v_i / \sum_j s_j, & if \ v_i \in S, \\ 0, & otherwise. \end{cases} \tag{2}$$

where $S_j \in S$, $p(v_i)$ is the probability that a random walk starts from v_i. The vector v_0 is updated as:

$$v_0 \leftarrow \alpha \mathbf{P}^T v_0 + (1-\alpha) v_0, \text{ where } \alpha \in [0,1). \quad (3)$$

After t iterations, the transition probability from vertex v_i to vertex v_k, denoted by $P_{t|0}(w_k \mid w_i)$, is equal to $P_{t|0}(w_k \mid w_i) = [\mathbf{P}^t]_{ik}$. The random walk sums the probabilities of all paths of length t between two vertices and gives a measure of the amount of the paths from one vertex to another. If there are many paths, the transition probability will be higher. Since the matrix $\mathbf{P}$ is a stochastic matrix, the largest eigenvalue of $\mathbf{P}$ is 1 and all other eigenvalues are in [0,1). Consequently, the vector v_0 will converge to v_*.

The value in the entry of the vector v_* is the probability that the vertex v_i is associated with a specified intent. Each Wikipedia concept is assigned a probability reflecting the degree of intent. We can treat the multiple intents classification as a set of binary intent classification. Consequently, this algorithm can be directly applied for each intent separately.

3.3 Explicit Semantic Analysis (ESA)

Since there are a large number of new words included in microblogs which are not covered by Wikipedia concepts, we address this problem by exploiting explicit semantic analysis (ESA)[5], which will provide a semantic interpreter that maps the fragment of text into some related concepts from Wikipedia ordered by their relevance to the input text fragment. Each Wikipedia concept is represented as a vector of words that occur in the corresponding article. Entries of these vectors are weighted by TFIDF scheme. Then we build an inverted index, which maps each word into a list of concepts in which it appears. Given a text fragment, we first represent it as a vector using TFIDF scheme. The semantic interpreter iterates over the text words, retrieves corresponding entries from the inverted index, and merges them into a weighted vector of concepts ordered by their relevance to the given text.

For words that are not contained in Wikipedia, we firstly use search engine such as Google to augment the representation of the word with top 10 search result snippets which include the titles and description parts. Then, we merge each snippets into sentences. Consequently, we can get a list of relevant Wikipedia concepts for each sentence by exploiting ESA. After the max-min normalization of rank scores of retrieved concepts, we get a desired number of highest-scoring concepts. Thus, we can predict the intent of the word according to the sum of the intent probabilities for the highest-scoring concepts.

For a given microblog, if the words in the microblog are covered by the Wikipedia concepts, we can get the probability that the microblog belongs to the defined intent using the intent probabilities of the mapped Wikipedia concepts. For the words that are not covered by Wikipedia concepts, we exploit ESA to map the words to the most related Wikipedia concepts and make the intent judgment based on the intent probabilities of the mapped Wikipedia concepts.

Let $\mathbf{K}$ be the set of intent categories (three intent categories in this paper). Each microblog will be associated with a vector of $||\mathbf{K}||$ elements and each element d_i^k in the vector represents the confidence score(i.e., intent probability) of the microblog belonging to category k estimated by our proposed method. Then the category with the highest intent score for each microblog is chosen as the inferred category, i.e., $\hat{k} = \arg\max_{k \in K} d_i^k$.

3.4 Collaborative User Model(CM)

One of the most important features of microblogs is its social network structure, which enables interactions between users. User *u* could mention his/her friend *f* with the symbolic @ sign and repost the microblogs posted by friend *f*. All the mentions and re-microblogs could be assigned an intent category k according to the intent probabilities of mapped Wikipedia concepts. The Collaborative User Model weights each friend *f* of user *u* on intent k by exploiting the Intent-interaction between them.

Intent-interaction: We first retrieve all microblogs which include the mentions of *f* by *u*, and re-microblogs of *f's* microblogs by *u* and then assign each microblog to an intent k. The intent-interaction weight between user *u* and the friend *f* is normalized by $w(u, f, k) = \log_{10}(1 + c(u, f, k))$ where $c(u, f, k)$ is the count of interactions assigned to intent k.

The Collaborative User Model can be denoted as Eq. 4, where F_u is the set of *u's* friends.

$$\theta_{u,k}^{CM} = \sum_{f \in F_u} w_{u,f,k} \,. \tag{4}$$

3.5 Intent Predictor

For a given microblog, it will be associated with a vector of $||\mathbf{K}||$ elements and the intent category k is determined by $\hat{k} = \arg\max_{k \in K} d_i^k$. With the Collaborative User Model, we can re-evaluate the d_i^k by simply add the $\theta_{u,k}^{CM}$ to it. If d_i^k is greater than a predefined threshold n, then the given microblog has the intent k and vice versa.

4 Experiments

4.1 Intent Applications

In this work, we apply our algorithm on three applications, namely Travel, Food&Drink and Physical Exercise, our approach is general enough to be applied to other applications as well.

- Travel is a complex social activity that involves with various services including agency services, transportation services, accommodation services, and other hospitality industry services. Therefore, we mark a microblog with a Travel intent if it is directly or indirectly related to the services mentioned above.
- The microblog that contains words related to food or drink is considered to have the Food&Drink intent.
- Physical Exercise includes various exercises such as Aerobics exercise, Bodyweight exercise, Walking, Yoga an so on. We mark a microblog with a Physical Exercise intent if it is directly or indirectly related to the exercises mentioned above.

4.2 Data collection

The data in this paper include Wikipedia data and microblog data. For the Wikipedia data, we identify over million distinct Wikipedia concepts for the three intents. There are 96,000 categories with an average of ten subcategories and 18 articles in each category. The seed concepts and the proportion of the three intents are shown in Table 1.

In order to collect enough intent microblogs, we give the definitions of *Intent-indicator* and *Intent-keyword*[9]. *Intent-indicator* refers a verb or infinitive phrase that that expresses intent on a general level. For example, “wanna”, “wanna to” are intent-indicators. *Intent-keyword* 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-indicator, e.g., in microblog “I wanna buy a car”, “buy” and “car” which are contained in the phrase “buy a car” are intent-keywords. Given the idea that a microblog is more likely to be an intent microblog if it contains Intent-indicator and Intent-keyword. We adopt the bootstrapping based method to retrieve intent microblogs.

Specifically, given a seed set of intent-indicators, (1) we extract the intent-keywords that frequently co-occur with intent-indicators, and (2) we use the extracted intent-keywords to extract more intent-indicators if their co-occurrence frequency is above a certain threshold. We repeat these steps until we cannot extract more intent-indicators and intent-keywords. Finally (3) microblogs which contain these extracted intent-indicators and intent-keywords are kept in our test collection for manual annotation. Finally, 43,584 potential intent microblogs are obtained, two annotators are employed to annotate the microblogs according to three categories. We get 19,714 intent microblogs and 23,870 non-intent microblogs with the same label by

two annotators. The Cohens Kappa coefficient between the two annotators is 0.6713. We summarize the statistics of this dataset in Table 2.

Table 1. Number of Wikipedia data

Intent Type	(%)	Seed concepts
Travel	21%	Travel; Hotel; Tourism; Airline tickets; Scenery; Travel Guide
Food&Drink	42%	Food; Drink; Restaurant; Noodle; Coffee; Juice
Fitness exercise	37%	Physical exercise; Aerobic; Bodyweight; Running; Swimming; Yoga

Table 2. Number of labeled dataset

Intent Type	#(%)	Seed concepts
Travel	6379 (49% positive)	I wanna travel to Paris. My dream is to travel around the world.
Food& Drink	7153 (52% positive)	I need some coffee. So hungry, it would be nice to have delicious noodles.
Fitness exercise	6182 (42% positive)	I find great pleasure in aerobic activity. I should go to the gym.

4.3 Algorithms Compared and Results

- **Hollerit's Method[8]:** It uses a Bayes Complement Naïve Bayes classifier, a classification model which attempts to address the shortcomings of the Naïve Bayes classifier with word and part-of-speech n-grams as attributes. The part-of-speech tagger we use is the Stanford POS Tagger.
- **Maximum entropy(MaxEnt) Method:** It exploits a hybrid feature representation including Bag-of-Words, POS tagging, named entity, and dependency trees.
- **Wang's Method[9]:** It formulates the problem of inferring intent categories from a small number of labeled tweets as an optimization problem. It constructs an intent graph with intent tweet nodes and intent-keyword nodes. With the labeled data, it leverages the association between nodes and propagates the evidence of intent categories via the intent graph.
- **Hassan's Method[10]:**It constructs a word relatedness graph with seed words which have known its polarities and words without knowing its polarities, then the random walk is performed. The confidence of a word being positive/negative is determined by the percentage of time at which the walk ends at a seed word and the average time the walk takes to hit a seed word. We use the microblogs for the words, and the seed words are the microblogs with known intent categories.
- **Ours:** Our method utilizes Wikipedia concepts and Collaborative User Model.

The metrics we used are F1, the macro-average and the micro-average. From the Table 3 we can see that our method outperformes other methods in every metric for three intent categories. The MaxEnt method is better than Hollerit's method for the linguistically oriented features such as POS tagging, named entity and dependency trees enhance the recognition of the intent categories. Wang's method and Hassan's method are both graph-based algorithms. Wang's method improves the macro-average and micro-average compared with the Hassan's method. The difference is that the results of the Hassan's method changes with the iterations the random walk takes, while Wang's method produces the unique answer to the optimization problem, which is more stable and efficient. Our method is better than Wang's method does not take the interactions into consideration like our method.

Table 3. The F1 scores on all categories.

Category	Travel	Food& Drink	Fitness exercise	Marco-F1	Micro-F1
Hollerit's	44.74%	42.14%	43.15%	41.91%	46.13%
MaxEnt	45.01%	43.72%	44.23%	42.42%	46.22%
Hassan's	46.24%	45.62%	45.31%	45.21%	47.56%
Wang's	47.12%	45.82%	45.81%	45.53%	47.82%
Ours	**50.12%**	**50.22%**	**50.31%**	**57.53%**	**51.42%**

To demonstrate the importance of the CM in our method, we compare our method with the method without using CM. Table 4 shows the comparison of precision, recall and F1 of two algorithms. It is shown that our method significantly outperformes the method without using CM in all three intent applications based on the F1 measure, which means that CM is important for intent identification.

Table 4. Comparison of the performance of our method and the method without using CM on the three intent identication task.

Intent Type	Our method			Method without using CM		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
Travel	67.46%	39.87%	**50.12%**	44.95%	37.99%	41.18%
Food&Drink	68.23%	39.73%	**50.22%**	45.35%	35.60%	39.89%
Fitness exercise	65.12%	40.99%	**50.31%**	43.47%	37.84%	40.46%

5 Conclusions

In this paper, we aim to solve the problem of identifying user intents in the microblog. We construct the Wikipedia Link Graph, each Wikipedia concept in the graph is assigned an intent probability. We exploit ESA to acquire most related Wik-

ipedia concepts for the words that are not covered by Wikipedia concepts and consequently get their intent probabilities. Then the intent of the microblog is determined by combing the intent probabilities of the mapped Wikipedia concepts and Collaborative User Model. The experimental results demonstrate that our algorithm achieves much better classification accuracy than baselines. Our work differs from previous works since we aim to use a human knowledge base to identify the user intent in microblog, and we do not collect large quantities of examples to train an intent classifier. This approach allows us to minimize the human effort required to investigate the features of a specified domain.

References

1. Broder A. A taxonomy of web search. ACM Sigir forum. ACM, 2002, pp. 3-10.
2. Rose D E, Levinson D. Understanding user goals in web search[C]. Proceedings of the 13th international conference on World Wide Web. ACM, 2004, pp. 13-19.
3. Shen D, Pan R, Sun J T, et al. Query enrichment for web-query classification[J]. ACM Transactions on Information Systems(TOIS), 2006, 24(3): 320-352.
4. Ashkan A, Clarke C L A. Impact of query intent and search context on clickthrough behavior in sponsored search[J]. Knowledge and information systems, 2013, 34(2): 425-452.
5. Hu J, Wang G, Lochovsky F, et al. Understanding user's query intent with wikipedia[C]. Proceedings of the 18th international conference on World Wide Web. ACM, 2009, pp. 471-480.
6. Dai H K, Zhao L, Nie Z, et al. Detecting online commercial intention (OCI)[C]. Proceedings of the 15th international conference on World Wide Web. ACM, 2006, pp. 829-837.
7. Chen Z, Liu B, Hsu M, et al. Identifying Intention Posts in Discussion Forums[C]. HLT-NAACL. 2013, pp. 1041-1050.
8. Hollerit B, Kröll M, Strohmaier M. Towards linking buyers and sellers: detecting commercial intent on twitter[C]. Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013, pp. 629-632.
9. Wang J, Cong G, Zhao X W, et al. Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets[C]. Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015, pp. 213-222.
10. Hassan, A., Radev, D. Identifying text polarity using random walks. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics . Association for Computational Linguistics. 2010, pp. 395-403