CASE: A Platform for Crowdsourcing Based API Search

Tingting Liang(✉), Liang Chen, Zhining Xie, Wei Yang, and Jian Wu

Zhejiang University, Hangzhou, China
{liangtt,cliang,lynntse,victor0118,wujian2000}@zju.edu.cn

Abstract. With the rapid growth of Web APIs on the Internet, searching appropriate APIs is becoming a challenging problem. General API search systems (e.g., ProgrammableWeb) implement API search through simple keywords matching leading to unsatisfactory search results. In this paper, we presents a crowdsourcing based API search engine CASE. Specifically, the API search engine leverages social information, *Twitter List*, a tool used by individual users to organize accounts that interest them on semantics. Based on the lists information, Latent Semantic Indexing (LSI) model is employed to compute the semantic similarity between the APIs and queries. Furthermore, the popularity of APIs inferred from the lists number is integrated with the semantic similarity to generate the final search result.

1 Introduction and Motivation

With the development of Web and mobile applications, a form of service called application programming interface (API) becomes prevalent. An API is a gateway to other people's software and can be employed to invoke a third-party software component over the Internet. Compared to the traditional web services, APIs are in a more popular style and can make solution more accessible and more useful, which lead to a rapid growth of Web APIs. According to the analysis offered by ProgrammableWeb (PW)[1], a platform dedicated to manage APIs and mashups, the number of its following APIs increased fast in recent years and has reached 10850 by October 2014. Thus, how to discover the appropriate APIs becomes a hot issue. Since general API search systems consider the keywords match rather than the real semantic or topic information of APIs, the performance of API search is limited. Taking PW as an example, if a consumer takes "travel" as the query in search system, the top result is "Webcams.travel" which is a directory of touristic webcams and classified in *Video* category. Absolutely, "Webcams.travel" is not the objective API. The reason is that PW search system only seeks APIs whose names or descriptions contain the query.

To alleviate the limitation of general API search systems, introducing crowdsourcing information into the search method is a popular and novel strategy. And it has been proved effective in many other fields, such as information retrieval,

[1] ProgrammableWeb: http://www.programmableweb.com.

A. Barros et al. (Eds.): ICSOC 2015, LNCS 9435, pp. 482–485, 2015.
DOI: 10.1007/978-3-662-48616-0_34

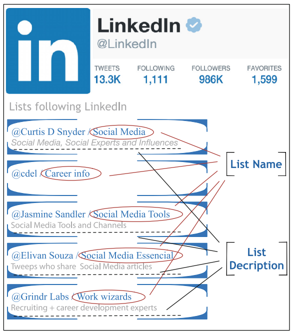

LinkedIn
@LinkedIn
TWEETS 13.3K
FOLLOWING 1,111
FOLLOWERS 986K
FAVORITES 1,599
Lists following LinkedIn
@Curtis D Snyder / Social Media
Social Media, Social Experts and Influences
@cdel / Career info
@Jasmine Sandler / Social Media Tools
Social Media Tools and Channels
@Elivan Souza / Social Media Essencial
Tweeps who share Social Media articles
@Grindr Labs / Work wizards
Recruiting + career development experts
List Name
List Decription

Fig. 1. An example of list

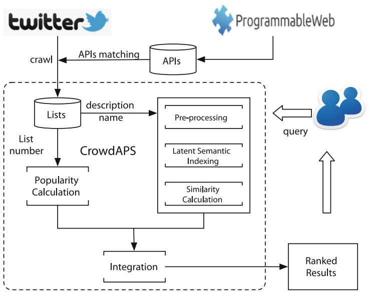


Fig. 2. Framework of crowdsourcing based API search

social network, etc. [2,4]. In this paper, we introduce extra crowdsourcing information, named as Twitter lists [3], to improve the performance of API search. Twitter lists, which are used to help users organize the Twitter accounts they are interested in, include list names and descriptions that contain a wealth of semantic information. Figure 1 illustrates the contents of lists by showing several lists including LinkedIn. Most of the list names are related to 'Social Media' and 'Career' along with the corresponding list descriptions, which distinctly reflect the topic of LinkedIn. After being matched with APIs in PW, we crawled 3877 APIs' Twitter accounts and the lists involving them. One API could be involved in many lists and the most reaches 84511. 30.8 % of total APIs are organized by more than 100 lists, and 70.1 % are included by at least 10 lists. Thus the list information of most APIs is full of richness.

In this paper, we build a demonstration of API search engine called CASE. The critical idea of CASE is to infer an API's topic by analyzing semantics of lists including the API Twitter account. Based on the lists information, Latent Semantic Indexing (LSI) [1] is employed to compute the semantic similarity between APIs and queries. Additionally, the popularity of APIs inferred from the lists number is integrated with the semantic similarity to generate the final search result. Therefore, CASE could provide semantics matched and popularity satisfied results for a given query.

2 Algorithm Framework

Figure 2 illustrates the whole framework of algorithm applied in the API search engine. The list data for API search is crawled from Twitter after being matched with APIs crawled from PW. Generally, the process of the algorithm could be divided into two parts. The first part is utilizing the name and description information of lists to compute the semantic similarity, including three steps: pre-processing, LSI and similarity calculation. After the data pre-processing like tokenization, stop words removal, stemming, etc., the features extracted from lists and the given query could be mapped into a conceptual space through LSI.

The returned relevant APIs are ranked according to cosine similarity, thus the semantics based rank score of an API and a given query is defined as:

$$score_s(q, API_i) = sim(q, d_i) = \frac{\sum_j w_{q,j} w_{i,j}}{\sqrt{\sum_j {w^2}_{q,j}} \sqrt{\sum_j {w^2}_{i,j}}}, \quad (1)$$

where $\mathbf{w}$ denotes term vector of an API or query in latent semantic space. Another part is about popularity calculation for each API based on its list number, which reflects how popular an API is among Twitter users to a certain extent. The popularity based rank score of an API is define as:

$$score_p(API_i) = \frac{\log_{10} N - \log_{10} min}{\log_{10} max - \log_{10} min} \quad (2)$$

where N represents the number of lists including the ith API, min and max respectively denote the minimal and maximal numbers of APIs' lists.

The final rank score is decided by the integration of semantic similarity and API's popularity, since the two measures synthetically satisfy the demands of users. Thus, combining the last two equations, we define the final rank score as:

$$score(q, API_i) = \lambda score_s(q, API_i) + (1 - \lambda) score_p(API_i), \lambda \in [0, 1] \quad (3)$$

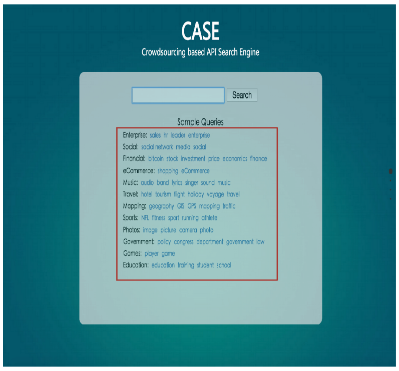
where λ is a weight to balance the importance of semantic similarity and popularity.

3 User Interface

Our crowdsourcing based API search engine is online available, users can use it to search APIs by visiting http://zjumsi.com/projects/case/index.php. The overall pages in the search platform are designed based on HTML5. The first page offers the crowdsourcing based API search function of the platform, and the next pages show some information and analysis about APIs, Twitter lists, and the demo video[2].

Figure 3(a) depicts the search interface and 52 sample queries from 12 categories are offered here. Users can click any sample query or input a query manually for searching the appropriate APIs. Figure 3(b) shows the search result page while user chooses a sample query *holiday*. It can be easily found that each search result entity mainly includes four parts: (1) API name; (2) API description; (3) the category API belongs to; (4) the lists number including API which shows API's popularity. Moreover, the first icon beside the API name links to the API page in ProgrammableWeb, which offers more detailed information about the API. When click the second icon, it will show the page of Twitter lists including the API, helping users to better understand the crowdsourcing knowledge about API.

[2] The demo video also can be found at https://youtu.be/D6xTzFkAXjQ.



CASE

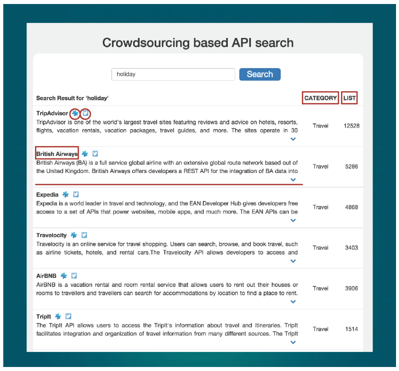
Crowdsourcing based API Search Engine

Search

Sample Queries

Enterprise: sales hr leader enterprise
Social: social network media social
Financial: bitcoin stock investment price economics finance
eCommerce: shopping eCommerce
Music: audio band lyrics singer sound music
Travel: hotel tourism flight holiday voyage travel
Mapping: geography GIS GPS mapping traffic
Sports: NFL fitness sport running athlete
Photos: image picture camera photo
Government: policy congress department government law
Games: player game
Education: education training student school

(a) Search Page



Crowdsourcing based API search

holiday — Search

Search Result for 'holiday'	CATEGORY	LIST
TripAdvisor TripAdvisor is one of the world's largest travel sites featuring reviews and advice on hotels, resorts, flights, vacation rentals, vacation packages, travel guides, and more. The sites operate in 30	Travel	12528
British Airways British Airways (BA) is a full service global airline with an extensive global route network based out of the United Kingdom. British Airways offers developers a REST API for the integration of BA data into	Travel	5286
Expedia Expedia is a world leader in travel and technology, and the EAN Developer Hub gives developers free access to a set of APIs that power websites, mobile apps, and much more. The EAN APIs can be	Travel	4868
Travelocity Travelocity is an online service for travel shopping. Users can search, browse, and book travel, such as airline tickets, hotels, and rental cars.The Travelocity API allows developers to access and	Travel	3403
AirBNB AirBNB is a vacation rental and room rental service that allows users to rent out their houses or rooms to travellers and travellers can search for accommodations by location to find a place to rent.	Travel	3906
TripIt The TripIt API allows users to access the TripIt's information about travel and itineraries. TripIt facilitates integration and organization of travel information from many different sources. The TripIt	Travel	1514

(b) Search Result Page

Fig. 3. User interface of CASE

4 Conclusion and Outlook

This paper demonstrates a crowdsourcing based API search platform named CASE. CASE applies LSI model to calculate semantic similarity based on Twitter lists information, and integrates it with API popularity to infer the final search result. Therefore, the search platform offers semantics matched and popularity satisfied results for a user query.

To enrich the function of the API search platform, it is considered to assign tags for each API which can be used to find APIs with similar function, and recommend APIs for a specific mashup in our future work.

Acknowledgment. This research was partially supported by the National Technology Support Program under grant of 2011BAH16B04, the National Natural Science Foundation of China under grant of 61173176, Science and Technology Program of Zhejiang Province under grant of 2013C01073, National High-Tech Research and Development Plan of China under Grant No. 2013AA01A604.

References

1. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. JAsIs **41**(6), 391–407 (1990)
2. Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., Gummadi, K.: Cognos: crowdsourcing search for topic experts in microblogs. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 575–590. ACM (2012)
3. Kallen, N.: Twitter blog: Soon to launch: Lists (2009)
4. Lamere, P.: Social tagging and music information retrieval. J. New Music Res. **37**(2), 101–114 (2008)