International Conference on Internet Studies,
July 18-19, 2015, Tokyo, Japan

AUTOMATIC CLASSIFY PLACES BASED ON CONTENT-AWARE MECHANISMS

Zhong-Xing Xie, Hsin-Wen Wei, Wei-Tsong Lee
Tamkang University, Taiwan
603450197@s03.tku.edu.tw, hwwei@mail.tku.edu.tw, wtlee@mail.tku.edu.tw

ABSTRACT

As information technology continues to progress and the prevalence of Internet, there are more and more data shared on websites. How to correctly classify data into useful information or knowledge is an interesting, important, and challenging issue. Therefore, this paper proposes a system to automatically classify places into two category: diet or not for location-based application. In this paper, a web crawler is developed to obtain data from the website and the content of each related webpage is identified by Chinese knowledge information processing (CKIP). Then a keyword table is designed to regulate data for training in SVM. By applying SVM, our system can determine what category of the input place is.

Keyword: machine learning, support vector machine, place classification

1 INTRODUCTION

As information technology continues to progress and the prevalence of Internet, there are more and more data shared on websites. Many social webs, such as Facebook and Google Plus, provides geographical and location information for users, so the users can share their status and location to their friends. Those webs also allow users to upload information about places if the users are not able to find out needed information about current location.

However, user may not provide enough information about the unknown place, they may just key in place name only. Therefore, how to correctly classify places for providing more information to other users is a challenging issue. For solve this problem, we design a system to automatically classify places into two category: diet or not for providing more information to users.

In this paper, we first develop a web crawler to get webpages from the website and utilize Chinese knowledge information processing to process each page for gathering useful and meaningful terms or phrases. Then, we design a keyword table to regulate data for training in Support Vector Machine (SVM). Finally, we apply SVM to predict what category of the unknown place is.

2 METHODOLOGY

In this section, we present the prediction method used in our system and the flow chart of the method is shown in Figure 1.

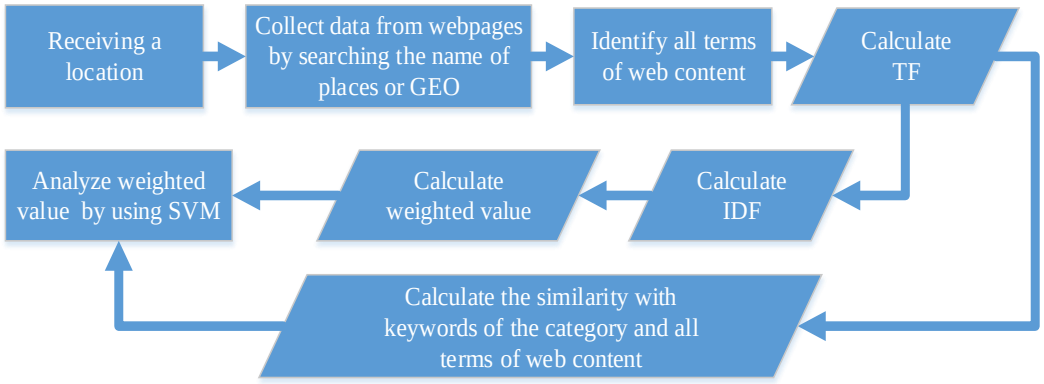

Figure 1. The flowchart of prediction method

● Data Collection

In this research, we design a web crawler to obtain data from the website. When the system receiving a location (or a place name) from an application or users, data about the location is collected from websites by searching the name or GEO of the location (or place) by using Google Search API [2]. Then, system identify all terms of web content by using Chinese knowledge information processing (CKIP) [3] and determine a weighted value for each of these terms. The weighted value of a term is calculated by its Term Frequency (TF) and Inverse Document Frequency (IDF) as shown in the following equations.

$$term's\ weighted\ value(i) = TF(i) \times IDF(i) \tag{1}$$

$$TF(i) = \frac{number\ of\ i^{th}\ terms\ in\ this\ webpage}{total\ number\ of\ terms\ in\ this\ webpage} \times 100\% \tag{2}$$

$$IDF(i) = log\frac{total\ number\ of\ webpages}{number\ of\ webpages\ with\ i^{th}\ term} \tag{3}$$

Though, the terms in a webpage can be identified by CKIP, the meaning of each term is unknown and hard to be automatically recognized. Therefore, the system still needs more information to classify the place. To help the system identify the category of unknown place, we further define a keyword table as follows.

● Keyword Table

We use the content in "iPeen" website to define a keyword table, in which all terms are related to diet. iPeen [6] is website for people to post their comments about dinning places, foods, and drinks. Therefore, we can collect the terms which appear in a webpage that describing a place and related to diet by searching iPeen.

First, we get 64 webpage about "食+site:http://www.ipeen.com.tw/comment" by using Google Search API and identify all terms in all webpages by using CKIP. Second, we calculate the number of occurrences of each term and delete some useless terms according to their syntactic functions which are identified by CKIP. For example: conjunctions, Foreign language…etc [4]. Then, as Table 1 shows, the number of occurrences of a term is transformed into the percentage among all occurrences of all terms, and the numbers of occurrences of all terms are sorted in

descending. Finally, we delete some kind terms are not related to diet enough again, and guarantee the keywords on the top 75 percent are actually related to diet. The other 25 percent goes to hidden or latent additional attributes.

Table 1 Occurrences of keywords in iPeen web.

keyword	occurrences	Percentage of occurrences	Accumulation of occurrences percentage
吃(Vt)	3094	4.275%	4.275%
吃到(Vt)	2783	3.846%	8.121%
火鍋(N)	2728	3.770%	11.890%
飽(Vi)	2501	3.456%	15.346%
吃到飽(Vt)	1745	2.411%	17.758%
鍋物(N)	1381	1.908%	19.666%
麻辣(A)	1219	1.684%	21.350%
⋮			
主廚(N)	12	0.017%	74.975%
紅茶(N)	12	0.017%	74.991%
胡椒(N)	12	0.017%	75.008%
脆脆(Vi)	12	0.017%	75.025%
⋮			
讚讚(Vi)	1	0.001%	100.000%
鑽木取火(Vi)	1	0.001%	100.000%
鱸魚(N)	1	0.001%	100.000%
豔麗(Vi)	1	0.001%	100.000%
Total	72369	100.000%	

● Similarity calculation

After obtained the keyword table about diet, the similarity of unknown webpage with the table can be calculated to help system classify webpage about the place more precisely. To calculate the similarity between terms in unknown webpage and terms in keyword table, Cosine similarity function is used and is shown as below:

$$Similarity = \frac{\sum_{i=1}^{n} keywords_i \times terms_i}{\sqrt{\sum_{i=1}^{n} keywords_i^2} \times \sqrt{\sum_{i=1}^{n} terms_i^2}} \qquad (4)$$

The value of similarity is in between 0~1, where 0 means not similar and 1 is the opposite.

● Classification with SVM

The weighted value of each term in a webpage can be obtained from equation (1) and the weighted value of all terms in each webpage can be summed up as below:

$$weighted\ value\ = \sum_{i=1}^{n} term's\ weighted\ value(i)$$

Then, the similarity and weighted value of a webpage are normalized between [0, 1] as SVM required and we analyze these two features with SVM.

Libsvm[5] is a useful SVM tool. Before using SVM, we use "svm-scale" function to normalize the features as between [0, 1]. Then we use "svm-train" function to train given data, and use "svm-predict" function to check the accuracy of training data, and try to tune the training model via libsvm tool. After obtained the training model, we applied this model to help system to classify the unknown data and also utilize "svm-predict" function to find the accuracy of unknown data. .

3 RESULT

In the experiment, we use the names of 60 stores nearby Tamkang University and 30 nearby Taipei University. Then we input the names of stores into our crawler program and obtain the URLs of webpages related to the names of stores which are shown in Table 2. When system obtained URLs of webpages, it will further crawl the content of webpages and use CKIP to identify terms in the webpages. After that, the TF and IDF value of all terms in the webpage are calculated and shown in Table 3. Further, the weighted value of all terms in the webpage and the similarity between the webpage and keywords that related to diet are shown in Table 4. In Table 4, we can see that the similarity of a webpage of a marketing consultant cooperation, called "網路通科技有限公司" is 0.119955573, which is higher than that of the webpage of a dinning place, called "七嘴八舌滷味". The result is inaccurate because the website of this marketing consultant cooperation offers many business cases which related to foods and drinks. Therefore, the number of terms related to diet category is higher, we can find the second webpage about "網路通科技有限公司" has 355 terms of all which is more than that of "七嘴八舌滷味". Although "網路通科技有限公司" is not a dinning place, the number of additional attributes might be much more than that of "七嘴八舌滷味". For this reason, "網路通科技有限公司" has higher similarity. Finally, we analyze these two features with SVM.

To avoid the difference of dinning culture or user behavior between Tamkang University and Taipei University that may cause the accuracy of classification decreased, we do not use the data that are collected around Tamkang University (Taipei University) as training data and the data that are collected around Taipei University(Tamkang University) as unknown data. Instead, we randomly select 120

data as training data and the remaining 140 data are treated as unknown data. The results of classification and accuracy are shown in Figure 2, Figure 3 and Figure 4.

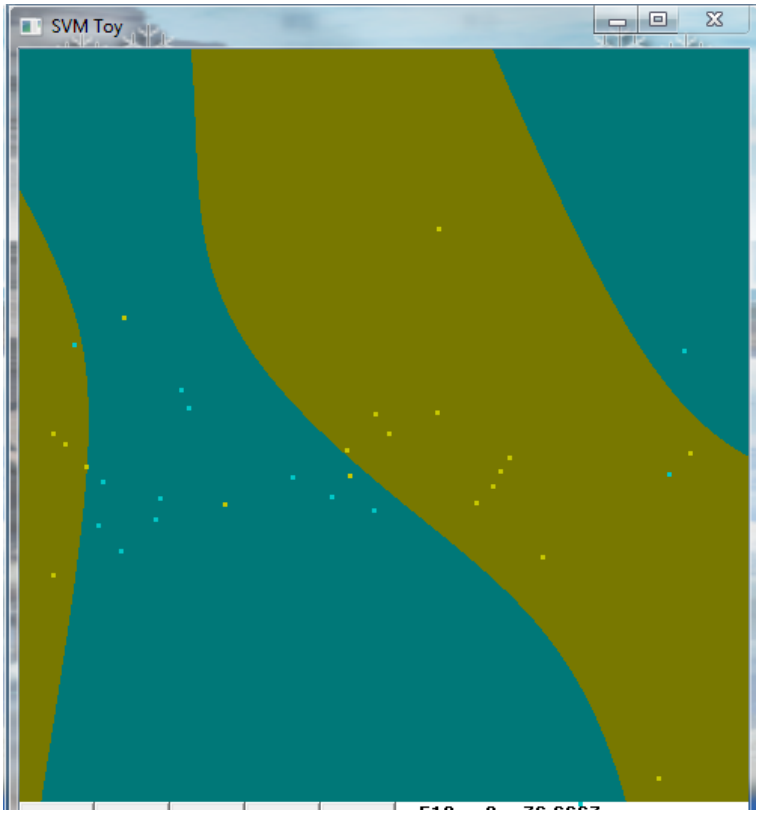

Figure 2. training data in libsvm

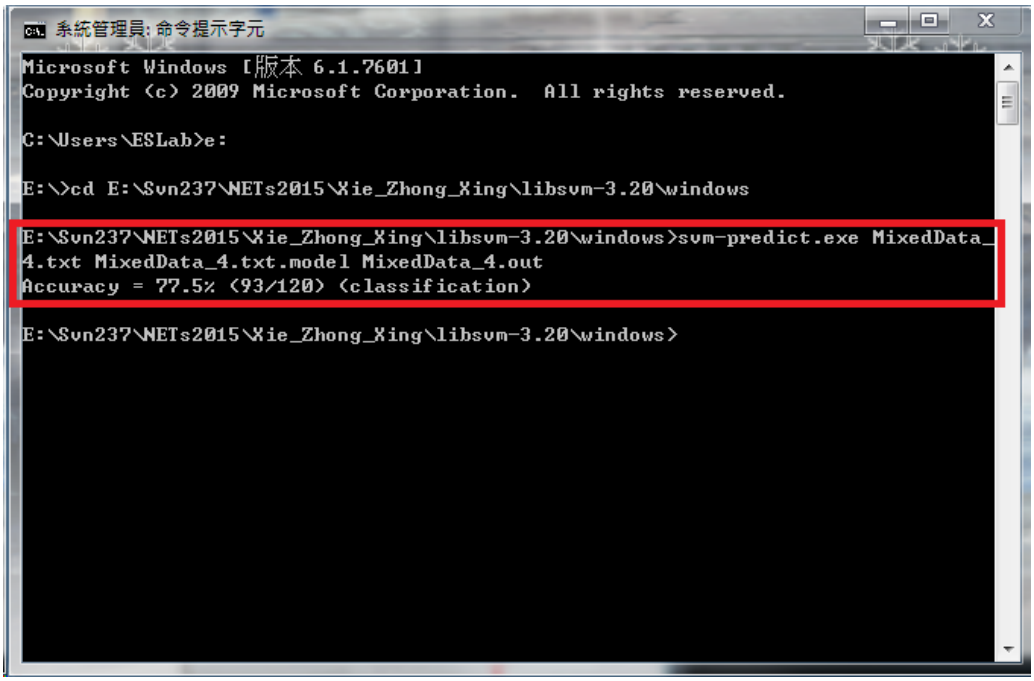

Figure 3. accuracy of training data

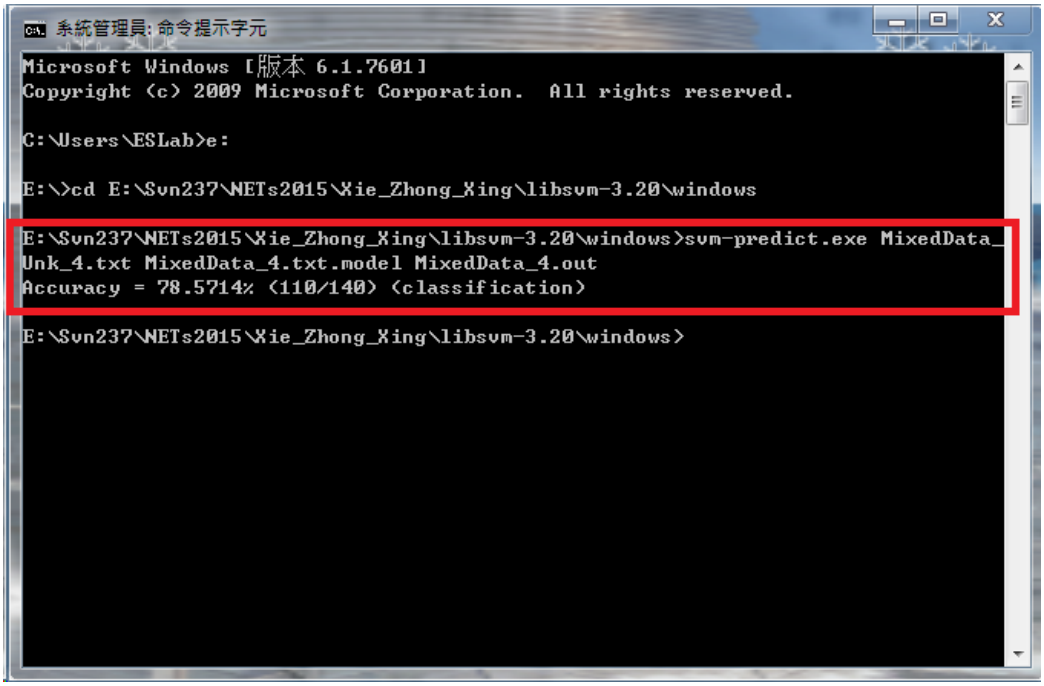

Figure 4. accuracy of unknown data

Table 2. part of URLs that are collected by the crawler program in the experiment.

Search name	URL
七嘴八舌滷味	http://www.7-8.com.tw/
	http://www.7-8.com.tw/menudetail.php%3FidNo%3D61
	http://www.7-8.com.tw/photoindex.php%3Fcid%3D38
	http://www.7-8.com.tw/photoindex.php%3Fcid%3D46
淡水亞太飯店	http://ap.hotel.com.tw/
	http://www.tripadvisor.com.tw/Hotel_Review-g1432365-d1726473-Reviews-Asian_Pacific_Hotel-Xinbei.html
	http://www.booking.com/hotel/tw/asia-pacific.zh-tw.html
⋮	⋮

Table 3. part of terms and their weight of webpages

Search name URL	Term	TF	Weight
七嘴八舌滷味 http://www.7-8.com.tw/	滷味(N)	4.0609%	0.16198920
	泡菜(N)	3.0457%	0.14288369
	當成(Vt)	1.0152%	0.05466492
	味道(N)	1.5228%	0.05229422
	七嘴八舌(Vi)	1.0152%	0.04762790

	的(T)	6.0914%	0.04222216
	特殊(Vi)	1.0152%	0.04049730
	傳統(N)	1.0152%	0.04049730
	辣(Vi)	1.0152%	0.04049730
七嘴八舌滷味 http://www.7-8.com.tw/menudetail.php%3FidNo%3D61	泡菜(N)	3.2609%	0.15297874
	滷味(N)	3.2609%	0.13007557
	韓式(N)	1.6304%	0.07648937
	味道(N)	2.1739%	0.07465190
	包心白菜(N)	1.0870%	0.05099291
	七嘴八舌(Vi)	1.0870%	0.05099291
	清脆(Vi)	1.0870%	0.05099291
淡水亞太飯店 http://ap.hotel.com.tw/	幅(M)	3.5088%	0.16460870
	飯店(N)	3.5088%	0.10141655
	饗宴(N)	1.7544%	0.09446483
	飽享(Vt)	1.7544%	0.09446483
	收進(Vt)	1.7544%	0.09446483
	復古式(N)	1.7544%	0.09446483

Table 4. part of similarity and weighted value of webpages

Search name URL	Similarity	number of terms about diet / all of terms	Weight	category
七嘴八舌滷味 http://www.7-8.com.tw/	0.090174751	40 / 134	2.9723535	diet
七嘴八舌滷味 http://www.7-8.com.tw/menudetail.php%3FidNo%3D61	0.129172264	34 / 114	3.0731430	diet
淡水亞太飯店 http://ap.hotel.com.tw/	0.21618104	11 / 45	2.7818575	diet
吃呼義料 http://www.591.com.tw/community-index.html%3Fcid%3D6265	0.235914381	5 / 56	3.062626	diet
網路通科技有限公司 http://company.zhaopin.com/P8/CC1370/2447/CC	0.015009416	72 / 276	2.338928	Not diet

137024478.htm				
網路通科技有限公司 http://www.net4p.com/aboutck.php	0.119955573	79 / 355	3.085666	Not diet
化學館 http://www.ch.ncku.edu.tw/	0	0 / 71	1.869868	Not diet
化學館 http://chemweb.tongji.edu.cn/	0.075102546	2 / 13	2.106275	Not diet

4 CONCLUSION AND FUTURE WORK

We develop a crawler program to obtain the data from webpage and give two features: similarity and weighted value to the webpage of unknown place. Then use SVM to classify these two features that makes successes to classify the unknown place into two classes: diet or not, and accuracy is 78.57%.

In the future, we will enhance the accuracy of classification method and the categories will be expanded to the range about living, like Clothing, Accommodation, Transportation, Education, and Recreation not only food or drink.

5 ACKNOWLEDGMENT

This work is supported by the Ministry of Science and Technology, Taiwan, R.O.C., under the grant No. MOST 103-2221-E-032 -035 -.

6 REFERENCES

[1] S. H. Wang and C.H. Lee, Research on Applying Text Mining, Image Annotation of Scenic Spots and Fusion Techniques for Geospatial Knowledge Discovery, 2010.
[2] http://ajax.googleapis.com/ajax/services/search/web
[3] CKIP. http://ckip.iis.sinica.edu.tw/CKIP/engversion/index.htm
[4] http://ckipsvr.iis.sinica.edu.tw/papers/category_list.doc
[5] LIBSVM--A Library for Support Vector Machines. Chih-Chung Chang and Chih-Jen Lin. http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[6] iPeen. http://www.ipeen.com.tw/