International Journal of u- and e- Service, Science and Technology
Vol.9, No. 5 (2016), pp.419-428
http://dx.doi.org/10.14257/ijunesst.2016.9.5.38

Annotating Tourism Resources with Social Tags and Classifying Them with Bayesian Algorithm

Hui Peng
Education Technology Center, Beijing International Studies University,
100024, Beijing, China
penghui@bisu.edu.cn

Abstract

Semantic annotation is the basement of semantic retrieval which has become a hot motivation of research in recent years. Social annotation makes it possible for semantic annotation of large amount of web resources. Because social tags express the semantic information of the resource, resources can be organized and managed by their semantics with the help of social tags. Information classification will help semantic retrieval greatly. We propose a kind of tourism classification which includes concepts related with tourism tags and a Bayesian based algorithm which aggregates concepts from social tags in this paper. Applying the algorithm to the automatic classification of tourism resources, the results show the effectiveness of the proposed algorithm.

Keywords: *social annotation, Bayesian classification, concepts aggregation, semantic retrieval*

1. Introduction

The semantic annotation of resources is the foundation of the semantic computing. The traditional way of semantic annotation is to describe resources with ontology concepts, instances, and ontology formulas. This way of annotation will be very difficult and arduous when the number of resource is huge (such as web resource). The following semantic computing is unable to carry out without semantic annotation. With the application of social tags, it becomes possible for automatic semantic annotation of huge resources.

Social tags was firstly put forward by Golder and Huberman in 2005 [2]. Social tags is a kind of action which allows users to attach any tags freely for any network resources.

Social tags can be directly published on the Web, which can be used for various types of resources. Social tags can be a kind of user's understanding and summary for the resource he has read. So it can be understood as the semantic information of the resource [9]. Many users individual behaviour of annotation are pooled together to form social tags in Internet. In 2010, Tim Berners-lee and Jim Handler pointed out social tags will have a huge impact on Web information search. [8]

The structure of social tags is user oriented. It allows users to tag various resources on the Internet: urls, images, video and other resources for independent tag. A set of social annotation includes a resource set(the resources which are annotated), a tag set(tags of the resources) and a user set(users of resouces). The data structure of a social annotation can be defined as a triple: F:= (U, T, R), which U is the user's finite set, T is the finite set of tags, R is a finite set of resources[10]. This triple reflected the semantic relations between resources and users, the semantic relations between resources and the semantic relations between users which are shown in Figure 1:

ISSN: 2005-4246 IJUNESST
Copyright ⓒ 2016 SERSC

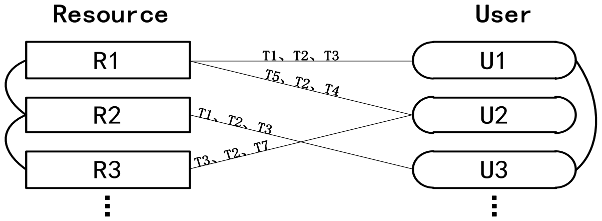

Figure 1. The Structure of a Social Annotation

2. Concept Aggregation and Resources Classification Based on Social Tags

2.1 The Probability Model of Resources and Tags

If there are a series of resources which annotated by tags, we can build the relation model between resources and domain concepts through the relations between tags and concepts. This relation model can be defined by Bayesian statistic model as show below because there is prior probability and conditional probability of a tag belongs to a set of concepts. Bayesian statistic model was broadly applied in prediction and classification based on statistics [1].

If *R* is the resource set of a domain, *C* is the category concepts set of this domain and there is a category tag c_j and $c_j \in C$。 Then, there is a probability between *R* and *C*. R^C :=(*R*, *C*, *φ*), and *φ*:= $R \times C$。 *φ* defines the relationship between *R* and *C*. It means there is a certain probability between a tag of a resource and a tag of classification. for example, a tag of a tourism resource r_i can be either belong to the category tag c_1, also can be belong to another category tag c_2, so there is a probability relationship between resources and category tags[6]. Then *φ*:= $R \times C$ can be expressed as:

$$\forall c_i \in C, \forall r_j \in R, r_j = \left(s_{r_j}\right),$$

$$\varphi\left(C \times R\right) = P\left(c_i \times s_{r_j}\right) \tag{1}$$

In formula (1), s_{r_j} is the social tag of resource $r_{j.}$ From (1), the relationship between *R* and *C* can be expressed by the relationship between social tag *S* and category tag *C*. This probability relationship meets the Bayesian decision. That means there is a Conditional probability to determine if S_r belongs to C_i. This Conditional probability can be described as formula (2):

$$P\left(c_i \times s_{r_j}\right) = P\left(s_{r_j} \mid c_i\right) = \frac{P\left(c_i \mid s_{r_j}\right) \cdot P\left(s_{r_j}\right)}{P\left(c_i\right)} \tag{2}$$

For each resource can be annotated by more than one tag, the probability of all tags consist of a matrix, which can be expressed as formula (3)

$$\Gamma_R = \begin{bmatrix} P\left(s_{1_1} \mid c_1\right) & P\left(s_{1_2} \mid c_2\right) & P\left(s_{1_3} \mid c_3\right) & \ldots & P\left(s_{1_m} \mid c_m\right) \\ P\left(s_{2_1} \mid c_1\right) & P\left(s_{2_2} \mid c_2\right) & P\left(s_{2_3} \mid c_3\right) & \ldots & P\left(s_{2_m} \mid c_m\right) \\ P\left(s_{3_1} \mid c_1\right) & P\left(s_{3_2} \mid c_2\right) & P\left(s_{3_3} \mid c_3\right) & \ldots & P\left(s_{3_m} \mid c_m\right) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ P\left(s_{n_1} \mid c_1\right) & P\left(s_{n_2} \mid c_2\right) & P\left(s_{n_3} \mid c_3\right) & \ldots & P\left(s_{n_m} \mid c_m\right) \end{bmatrix}_{m \times n} \tag{3}$$

Formula (3) defines the probability determination which every social related to tag category tags. In this matrix, one line vector expresses the probability of C_i to all tags of one resource which in resources set R. And one colon expresses the probability of one concept to tags s_{r_j}. Through the analysis of the matrix Γ_R, we can aggregate concepts from catalog tags and automatic classify resources as the following parts of this paper.

2.2 The Procedure of Concepts Aggregation

We define the semantic overlay of a resource is the catalog tags which described the resource. We can get the semantic overlay about a resource from Γ_R.

Every element $P(s_j|c_i)$ in Γ_R express the probability of social tag S_j related to catalog tag C_i. The value of $P(s_j|c_i)$ is determined by two factors. One is from the classification of R. because $s_j \in R$, if R is classified to c_i, then s_j is determined by c_i. The other factor is from the connotation of s_j itself. This connotation can be concluded by the situation that s_j belongs to different resources.

From this perspective, using $P(s_j|c_i)$ to recalculate the relation between the resources and the catalog concept can increase the contribution of the social tag to the catalog concept and decrease the error of artificial classification about a resource[6].

If resource $r_t \in R$, from (3), we sum the elements $P(s_i|c_j)$ in Γ_R can get formula (4):

$$\Delta R = \sum_{\mathrm{i}} \sum_{\mathrm{j}} P\left(s_j \mid c_i\right) \cdot c_i \tag{4}$$

In formula (4) ΔR is the linear sum which for resource r_t. catalog tag c_i is as the independent variables, the Probability determination of catalog tag to social tag is the coefficient.

Calculate all resources on formula (4), we can get A series of linear function. They show the relationship between current resource r_t and catalog tags set C. From this we can get the linear description of every catalog tag c_i about resource set R. This description is shown in formula (5).

$$\Delta C = \sum_{\mathrm{j}} \sum_{\mathrm{i}} P\left(s_j \mid c_i\right) \cdot r_j \tag{5}$$

Formula (5) can be expressed as a matrix. Every column vector in the matrix expresses the decision function which a catalog tag c_i related to the resource r_t The similarity of two catalog tags can be calculated from the two column vectors, which is shown in formula (6).

$$corr\ (\vec{c}_i, \vec{c}_j) = \frac{\vec{c}_i \bullet \vec{c}_j}{\left|\vec{c}_i\right| \cdot \left|\vec{c}_j\right|} \tag{6}$$

In formula (6), $\vec{c}_i \bullet \vec{c}_j$ is the dot product operation and $\left|\vec{c}_i\right|$ is the module operation on $\vec{c}_i$.

If the threshold of concept aggregating is δ, then when $corr\ (\vec{c}_i, \vec{c}_j) \geq \delta$, catalog tag c_i and c_j can be aggregated to be a semantic overlay.

2.3. The Procedure of Resource Classification

Because Γ_R expresses the probability determination of every social tag related to every catalog tag it can be used to classify the resource which described by a serial social tags. If the resource r_t is annotated by a serial social tags, it is marked as r_t (s_{t1}, s_{t2}, ..., s_{tm}). And the resource r_t is to be classified. Then how to select a set of catalog tags c_i for r_t to accurate overlay r_t is the problem of automatic classification. We can adopt the posteriori probability algorithm to accomplish this aim.

The Posterior probability algorithm can be described as the followings.

We can classify the new resources by matrix Γ_R through the Posterior probability algorithm. From formula (2), we can get formula (7)

$$P\left(c_i \mid s_{t_j}\right) = \frac{P\left(s_{t_j} \mid c_i\right) \cdot P\left(c_i\right)}{P\left(s_{t_j}\right)} \tag{7}$$

From formula (7), we can calculate the Posterior probability $P\left(c_i \mid s_{t_j}\right)$ if $P\left(s_{t_j} \mid c_i\right)$ is known. That means we can be got the concept analysis of a resource if we know the probability determination of every catalog tag related to social tag.

Because a resource often includes more than one social tag, formula (7) should be transformed into formula (8)

$$P\left(c_i \mid s_{t_1}, s_{t_2}, \dots, s_{t_n}\right) = \frac{P\left(s_{t_1}, s_{t_2}, \dots, s_{t_n} \mid c_i\right) \cdot P\left(c_i\right)}{P\left(s_{t_1}, s_{t_2}, \dots, s_{t_n}\right)} \tag{8}$$

If every social tag is independent, which means each social tag affects concept classification independently and two social tags have no relevance affection to concept classification, the formula (8) can be expansion as formula (9)

$$P\left(c_i \mid s_{t_1}, s_{t_2}, \dots, s_{t_n}\right) = \frac{\prod_{j=1}^{n} P\left(s_{t_j} \mid c_i\right) \cdot P\left(c_i\right)}{\prod_{j=1}^{n} P\left(s_{t_j}\right)} \tag{9}$$

According formula (9), if a resource rt includes social tag set $r_t\left(s_{t_1}, s_{t_2}, \dots, s_{t_n}\right)$, we can calculate the concept analysis about rt by the data in ΓR. Every catalog tag ci will contribute to rt theoretically, we can use the top K catalog tags as the tag of resource rt in fact.

3. An Example of Tourism Resource Classification

3.1. Selecting the Set of Training Resources

About the classification of tourism information, there exist many classification catalog. Take China as an example, there are at least two kinds of catalog. One is the industry catalog which is show in Figure.2, the other is the academy catalog which includes at least three different kinds of catalog: the Chinese library classification, CNKI(Chinese National Knowledge Infrastructure) classification and RUC(Renmin University of China) database classification. Because most websites use industry classification to exhibition their information, we take the industry classification as domain concepts so that the tourism knowledge may be shared intensively according this classification.

In this paper, some tourism web resources were collected on the Internet as the experimental data. The tags of a resource are collected according their rank in all tags, Top 5 social tags of a resource are selected to catalog the classification. Based on the industry classification, these resources were divided into seven categories, as shown in Figure 2.

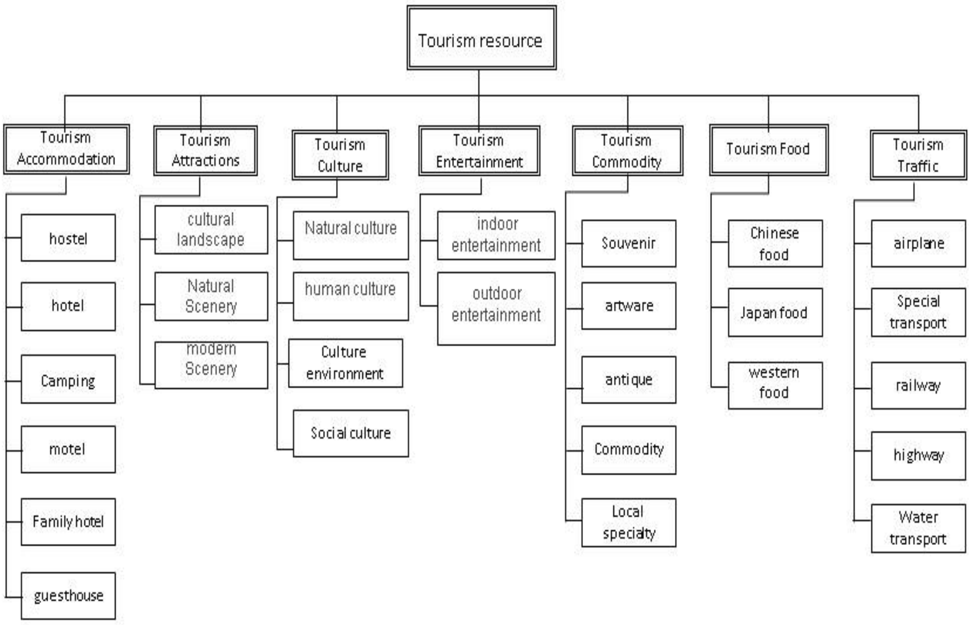

Figure 2. Tourism Resources Classification Hierarchy Diagram

Then assign abbreviated id for these seven categories, for the convenience of later calculation, as shown in Table 1.

Table 1. Tourism Resources Catalog Tags

Id	Tourist Resources Classification
C.1	Tourism Traffic
C.2	Tourism Attractions
C.3	Tourism Commodity
C.4	Tourism Accommodation
C.5	Tourism Food
C.6	Tourism Entertainment
C.7	Tourism Culture

This paper presents a model to analyse training data set of tourism resources. Through online questionnaire survey of these resources and statistics, we got a collection of social tags of these resources, as shown in Table 2.

Table 2. Tourism Resources Social Tags

Id	Social Tags	Id	Social Tags
ST01	Railway	ST11	Hotel
ST02	Self-Driving	ST12	Inn
ST03	Ship	ST13	Chinese Food
ST04	Landscapes	ST14	Foreign Food
ST05	Historical Sites	ST15	Outdoor Sports
ST06	Plaza	ST16	Indoor Entertainment
ST07	Park	ST17	Historical Culture
ST08	Souvenir	ST18	Folk Culture
ST09	Specialty Snacks	ST19	Art Appreciation
ST10	Shopping		

It should be noted that some tags with very low frequency occurrence were excluded by the extraction of social tags set. This is because of their low word frequency, so that generating more noise than contribution for the model.

3.2 Building Calculation Model

Through the analysis of social tags, we acquired probability determination of catalog tags to social tags by the model, as shown in Table 3.

Table 3. Probability Determination of Catalog Tags to Social Tags

	Catalog tags						
Social Tags	C.1	C.2	C.3	C.4	C.5	C.6	C.7
ST01	0.347						
ST02	0.149	0.037		0.046	0.041	0.057	
ST03	0.058	0.029					0.024
ST04	0.186	0.186		0.193	0.172	0.119	0.103
ST05		0.133	0.083	0.042	0.037		0.139
ST06			0.072		0.064	0.044	
ST07		0.052				0.120	0.043
ST08		0.104	0.217	0.087	0.077	0.027	0.202
ST09			0.054		0.048		
ST10			0.072		0.064	0.044	
ST11		0.036	0.045	0.181	0.120	0.083	
ST12	0.058	0.029	0.036	0.108	0.096	0.044	0.024
ST13		0.029	0.144	0.036	0.160	0.044	0.048
ST14			0.072		0.064	0.044	
ST15	0.076	0.076		0.142	0.126	0.204	
ST16			0.054		0.048	0.067	
ST17		0.186	0.093		0.041		0.186
ST18		0.029	0.072	0.108	0.064	0.022	0.072
ST19			0.093		0.041	0.057	0.062

3.3. Aggregation of Concepts

According to the data in Table 3, the system aggregated the classification of tourist resources in the training set. Results are as Table 4.

Table 4. Aggregation of Classification Concepts

	Catalog tags						
Catalog tags	C.1	C.2	C.3	C.4	C.5	C.6	C.7
C.1							
C.2	0.629						
C.3	0.208	0.714					
C.4	0.774	0.830	0.553				
C.5	0.673	0.805	0.741	**0.919**			
C.6	0.719	0.669	0.462	**0.880**	**0.891**		
C.7	0.379	**0.919**	**0.860**	0.646	0.676	0.440	

Setting aggregation threshold δ = 0.85, after calculation of the training set, the system got the concept aggregation that were C.2 and C.7, C.3 and C.7, C.4, C.5 and C.6, that is, "Tourism Attractions" and "Tourism Culture", "Tourism Products" and "Tourism Culture", "Tourism Accommodation", "Tourism Food" and "Tourism Entertainment". Based on the correlation coefficient, we got four semantic overlay of C_2C_7, C_3C_7, C_4C_5 and C_6C_5.

3.4. Classification of Tourist Resources

System selected some other tourist Web resources as test set, as shown in Table 5. The tourist resources of test set were already classified by relevant tourism experts, according to the method shown in Figure 2.

Table 5. Sample and Classification of Test Set

Id	Test Resource	C. Tags	Id	Test Resource	C. Tags
TR01	Yunmeng Mountains	C.2	TR06	Sun Yat-sen Mausoleum	C.7
TR02	Military Museum	C.7	TR07	Yuanming Yuan	C.2,C.7
TR03	Bird's Nest	C.2,C.7	TR08	Shanhaiguan	C.7
TR04	Wangfujing	C.3	TR09	Ditan Book Fair	C.2,C.7
TR05	Nanjing Yangtze River Bridge	C.1,C.2	TR10	Eighty-one Film Base	C.6

With the model presented in this paper, we first employed posteriori probability algorithm to obtain the classification probability matrix of the tourist resources in test set, with the maximum value of each sample's classification probability indicated by bold font, as shown in Table 6.

Table 6. Results of Classification on Tourism Resources

	Catalog tags							Classification Results
Test Resources	C.1	C.2	C.3	C.4	C.5	C.6	C.7	
TR01	0.350	0.088		**0.641**	0.451	0.465		C.4
TR02		0.453					**0.880**	C.7
TR03		0.158				**0.328**		C.6
TR04			**0.903**		0.803	0.555		C.3
TR05	**0.464**	**0.464**					0.258	C.1，C.2
TR06		0.929					**0.967**	C.7
TR07		0.929					**0.967**	C.7
TR08		**0.808**			0.069			C.7
TR09						**0.104**		C.6
TR10	0.232	0.116		0.218	0.193	**0.625**		C.6

3.5. Overall Evaluation

From the experimental results, the model correctly determined the catalog tags of tourist resources by six times, accounting for 60% of the total; partly determined two times, accounting for 20%; determination of errors was two times, accounting for 20%. Effective determining probability of the system about the test set was around 80%.

There are some Common automatic classification algorithms based on statistics, such as Dicision Tree, Neural Net, Rocchio and Bayesian[5].

The complexity of Dicision Tree is decided by the numbers of nodes of tree. When the number of nodes exceeds 200, the complexity of this algorithm will low the praticability.

The algorithm of Neural Net fits complex data relations and its has the power of learning capability. But this method has not been applied in classification area because it's difficult to be realized.

Table 7 shows parts of the results of the first contest of auto Chinese web pages classification[7]. The precision of category "Business and economy" and "Entertainment and relaxation" is much lower than "education" and "Natural science". Because the concepts and tags in category "Business and economy" and "Entertainment and relaxation" are often miscellaneous and not standard as concepts and words in "education" and " Natural science". And most travel information are related to the category "Business and economy" and "Entertainment and relaxation". The Effective determining probability of our system means a good precision.

Table 7. Classification Precision of Different Category

category	precision
Business and economy	42%
Entertainment and relaxation	38%
education	95%
Natural science	85%

Furthermore, compared with other classification algorithms which with the fixed category, our algorithm calculates the semantic overlay and aggregate concepts through semantic overlay. Concept aggregation can expand and modify the original category. New category can adapt the change of web information and improve the precision continually.

4. Related Works

After the concept of social tags ware put forward by Golder and Huberman in 2005[2], the researches about social tags are concerned on the SIGIR,WWW conference and other publishers. Lambiotte researched social tags as triples which include users, resources and tags [10]. Li researched the application of social tags in semantic Web. It promoted a new annotation method by analysing the concepts relations through context in blogs [3]. Liu constructed domain ontology by social tags[4]. Lu combined social tags with Wordnet or ontology to implement semantic retrieval[11]. Many Chinese tourism websites such as ctrip, elong introduced social tags to annotate tourism logs and documents. These tags can be used to construct tourism concept and classification to improve retrieval results.

Acknowledgment

This work is supported by the National Social Foundation of China (No. 14BTQ031) and the scientific planning project of Beijing municipal commission of education (SQKM201410031001).

References

[1] J. Yanan, L. Ruixuan and W. Kunmei, "A Survey on Social Annotation and Its Application in Information Retrieval", Journal of Chinese Information Processing, vol. 24, no. 4, **(2010)**, pp. 52-62
[2] H.-Q. Li, F. Xia, D. Zeng, F.-Y. Wang and Wen-Ji Mao, "Exploring Social Annotations with the Application to Web Page Recommendation", Journal of Computer Science and Technology, vol. 24, no. 6, **(2009)**, pp. 1028-1034.
[3] F. S. Cong and W. J. Min, "The Analysis of a Contest Result on Chinese Web Page Automatic Categorization", Journal of Chinese Information Processing, vol. 17, no. 5, **(2003)**, pp. 34-40
[4] K. Liu and B. Fang, "Ontology Construction by Social Tags", Chinese Journal of Computers, (in Chinese), vol. 33, no. 10, **(2010)**, pp.1824-1833.
[5] S. A Golder and B. A. Huberman, "The Structure of Collaborative Tagging Systems. Arxiv preprint, 2005
[6] Z. Shi. Knowledge discovery", Tsinghua press, (in Chinese), **(2001)**.
[7] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization", Information Retrieval, vol. 1, no.1/2, Kluwer Academic Publishers, **(1999)**, pp. 60-69
[8] H. Yu, "Design and Implement of Chinese Web Page Automatic Collection and Classification", Master Paper of Beijing university of posts and communications, **(2010)**.
[9] J. Hendler and T. Berners-Lee, "From the semantic web to social machines: A research challenge for AI on the World Wide Web, Artificial Intelligence", vol. 174, Issue 2, **(2010)** February, pp. 156-161.
[10] R. Lambiotte and M. Ausloos, "6th International Conference on Computational Science (ICCS 2006), Reading, UK, **(2006)**.
[11] Y. Lu, M. Castellanos, U. R. Dayal and C. X. Zai, "Automatic Construction of a Context-Aware Sentiment Lexicon:An Optimization Approach", WWW, Hyderabad, India, **(2011)**.

Authors

Hui Peng, She received her PhD in Computer Sciences (2009) from Institute of Computer Technology Chinese Academy of Science. Now she is a full professor of computer department, Beijing International Studies University. Her current research interests include different aspects of Artificial Intelligence and Distributed Systems.

Copyright ⓒ 2016 SERSC