International Journal of Hybrid Information Technology
Vol. 9, No.10 (2016), pp. 51-58
http://dx.doi.org/10.14257/ijhit.2016.9.10.06

# A Semantic Web Services Classification Approach based on IG-C4.5

Chaokai He[1] and Meng Wu[2]

[1]*School of Computer Science, Nanjing University of Posts and Telecommunications,*
*No. 9, Wenyuan Road, Nanjing 210023, China*
*{heck}@njupt.edu.cn*
[2] *College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, No. 9, Wenyuan Road, Nanjing 210023, China*
*{wum}@njupt.edu.cn*

***Abstract***

*Semantic Web services have emerged as the solution to the need for automating several aspects related to service-oriented architectures, such as service discovery and composition, and they are realized by combining Semantic Web technologies and Web service standards. The increasing number of available Web services has raised the need for their automated and accurate classification. Services Classification would be useful to enact powerful capabilities such as automatic selection, invocation, composition, location or discovery of web services. In this paper, We proposes a approach of Semantic Web Services classification taking into account the feature of textual description and the semantic annotations of OWL-S advertisements. An experimental test of the proposed techniques is reported, showing positive results, the efficiency and accuracy of the automatic service portfolio has been improved.*

***Keywords****: Semantic Web Services, Services Classification, IG-C4.5*

## 1. Introduction

Over the last decade, a great amount of effort and resources have been invested in the development of Semantic Web Service (SWS) frameworks. With the expectable growth of the number of Web services available on the WWW and service repositories, service oriented architectures has become the mainstream software architecture. The emergence of the semantic web has been greatly improved the development of service discovery and composition, which are realized by combining Semantic Web technologies and Web service standards. From our point of view, it has two main drawbacks. Fist, WSDL (Web Service Definition Language)-based [1] descriptions, as a means to enable the manipulation of services by software programs easing the automation of such tasks as service selection, invocation, composition, and discovery. Which with limited description ability, second, unable to handle uncertain ontology's information *etc*.

The automated and accurate classification in domain categories that can be beneficial for several tasks related to SWS, The effectiveness and efficiency of service discovery algorithms can be improved using WS classification by filtering out services that do not belong to the domain of interest. The classification of SWS can be used in order to increase the accuracy of SWS composition by examining only the domain-relevant services in each step of the service workflow generation process. The management of large number of SWS in repositories is more effective when services are organized into categories. Furthermore, automated service classification can be utilized during the

ISSN: 1738-9968 IJHIT

process of registering SWS in repositories by recommending service categorizations to the users. [2]

This paper presents a method for the automatic classification of SWSs based on their OWL-S Profile instances. A Profile instance provides descriptive information about the service, such as textual description, as well as semantic annotations of WS's inputs, outputs, preconditions, effects, non-functional properties, *etc*. Because of the text and semantic information service description files, sample data show significant sparse, high-dimensional nature, we propose a high-dimensional, sparse data dimensionality reduction method of classification, and related experiments shows positive results.

The paper is organized as follows: Section 2 introduces some previous work in the domain of Semantic web service classification and other areas related to the work presented here. Section 3 introduces the feature selection based on Information gain-C4.5. Section 4 present some ideas demonstrating why web service semantics are needed in our classification approach. The Experiments and explanation of our classification approach is described in Section 5. Finally, Section 6 presents some conclusions.

## 2. Related Works

Service categorization is commonly used to facilitate service retrieval, be it mainly by automatic discovery mechanisms. Existent web service classification proposals may be divided into different categories according to various criteria. We briefly discuss next the most relevant initiatives in this scope related to our research.

WSs Classification approaches based on the using of structured text elements form various WSDL components. WSDL descriptions provide us with details about the operations a service provides, as well as the input and output information involved. In [3] WSDL text descriptions are used in order to perform automatic classification of WSs using Support Vector Machines (SVMs) .various classification methods like naive Bayes [4,5], SVMs [4], decision trees [6] or even ensemble of classifiers [6, 7].which using structured text elements from various WSDL components, with disadvantage is that no semantic information is taken into account.

If the descriptions were enriched with further semantic information it would be useful for classification. In [8], the classification of WSs is based on OWL-S advertisements and it is achieved by calculating the similarities of I/O annotation concepts between the unclassified WS and a set of pre−classified WSs for each class. The main disadvantage of this approach is that the representation is not flexible enough in order to be used with any machine learning algorithm and that the text of the description is ignored. In [2] provide evaluation results that prove the utility of even short textual descriptions that may appear in the description of the WS advertisement. A similar task to classification is SWS matchmaking. In this case a query WS description is given in order to find a set of similar WSs [9, 10].

Actually, we found that the service data showing a high-dimensional, sparse characteristic, *i.e*. large amounts of data do not contribute to the actual classification task, resulting in a large number of invalid computation storage resource overhead. Therefore, we propose a high-dimensional oriented, service classification methods based on the combination of textual and semantic information.

## 3. Feature Selection based on Information Gain-C4.5

Due to the comprehensive feature set contains a large number of service features, and the data have sparse, high-dimensional features, the feature set that contains a large amount of redundant and irrelevant information. Therefore, feature selection has emerged as a key sticking point to reduce the amount of data that needs to become processed to optimize the classification.

 

$F = \{f_1, f_2, ..., f_N\}$,

$S = (s_1, s_2, ..., s_N), s_i \in \{0,1\}, i = 1,2,...,N$, $s_i = 1$ if i[th] feature is selected, otherwise $s_i = 0$. Thus the problem of feature selection convert to an optimization problem as follow:

$$\max_S G(S) \tag{1}$$

### 3.1. Information Gain

Information gain is sometimes used synonymously with mutual information, which can be used to measure the amount of information contained in a feature. The greater the gain characteristics of the information on behalf of classified information it contains also more role classification. For Feature A, whose information gain is calculated by

$$Gain(A) = I(s_1, s_2, ..., s_m) - E(A) \tag{2}$$

Where m is the number of categories of samples, s is total number of samples, $s_i$ is the number of samples which belong to the subset $C_i$, $P(C_i)$ is the probability of random sample belongs to $C_i$, $P(C_i) = s_i / s$.

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} P(C_i) \log_2 P(C_i) \tag{3}$$

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + ... + s_{mj}}{s} I(s_{1j}, s_{2j}, ..., s_{mj}) \tag{4}$$

$E(A)$ is the total entropy of samples' feature, A=$\{a_1, a_2, \ldots, a_v\}$, $s$ could be divided into $v$ subsets, those sample which' feature A equal to $a_j$ is including in subsets $s_j$.

$(s_{1j} + s_{2j} + ... + s_{mj}) / s$ is the weights of $j$th subset.

$s_{ij}$ is the number of sample $C_j$ in subset $s_j$

$$I(s_{1j}, s_{2j}, ..., s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2 p_{ij} \tag{5}$$

$P_{ij} = s_{ij} / s_j$ is the probability of random sample of $s_j$ belongs to $C_i$,

The division is to obtain information on the characteristics of the gain from A, which represents the highest information gain feature is characteristic of a given record set with a degree of distinction. By calculating the information gain, can resorts the feature in the classification of the importance of an assessment of the rankings.

### 3.2. C4.5 Algorithm

C4.5is an algorithm used to generate a decision tree developed by J.R. Quinlan[11]. C4.5 builds decision trees from a set of training data, using the concept of information entropy. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sublists. Formula for calculating the entropy S is set as follow

$$Info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} * \log_2 (\frac{freq(C_j, S)}{|S|}) \tag{6}$$

$freq(C_j, S)$: The feature's number of subset $C_j$ in set S

$|S|$: The number of sample of set S

Partitioning according to a property after it comes to a number of subsets, the need for these subsets of entropy weighting and count as

$$Info_f(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} * Info(T_i) \quad (7)$$

$T$ Refers to a collection by attribute f after partition, $T_i$ the $i$th subset

$|T_i|$: The sample number of subset $T_i$

$$Gain(f) = Info(S) - Info_f(T) \quad (8)$$

is defined to calculate the entropy of the set after partition

The rate of information gain is calculate as $Ratio(A) = \frac{Gain(A)}{I(S_1, S_2, ..., S_v)}$ (9)

$v$ is the number of branches, Si is the number of *ith* node's record

### 3.3. Feature Selection based on Information gain-C4.5

General characteristics of the hybrid algorithm can be divided into two stages: feature selection algorithm to select a subset of the first candidate to use filtering features based on the data's characteristics; and then concentrated using a specific learning algorithm in the candidate feature subset cross validation, optimization feature subset selection .

The feature selection based on Information gain-C4.5, which make two improvement

First: IGi,c indicates the information gain between Feature Ai and Category C. At the first stage: each feature is sort by IGi,c after calculate by formula (2).

Second: After sorting feature is complete, enter the package type feature selection cycle. In the cycle, as the search strategy generated using GFS feature subset, C4.5 classification as evaluator, to select optimal feature subset. The ratio of 5-fold cross-validation misclassification rate variance and classification error rate obtained from the mean as feature selection loop termination condition, which is defined as follow

$$f_{error} = \frac{MSE(R)}{MEAN(R)} \quad (10)$$

R is the classification error ratio of 5-fold cross-validation

MSE(R) is used to calculate the mean variance of error rate of classification

MEAN(R) is used to calculate the mean error rate of classification

$f_{error}$ 's value is small, which Show cross-validation error rate close to stable classification results. Setting a threshold value $\varepsilon$, Feature selection loop terminates while $f_{error} < \varepsilon$. Several experiment shows that $\varepsilon$ 's ideal value is 1%.

## 4. Semantic Web Services Classification Approach based on IG-C4.5

First, Optimal feature subset is selected using feature algorithm based on IG-C4.5,The extended feature set and the training set as input, after feature's information gain is calculated, selecting features within the ranks δ to form a new set of features. And then go to the feature selection iteration loop.

 

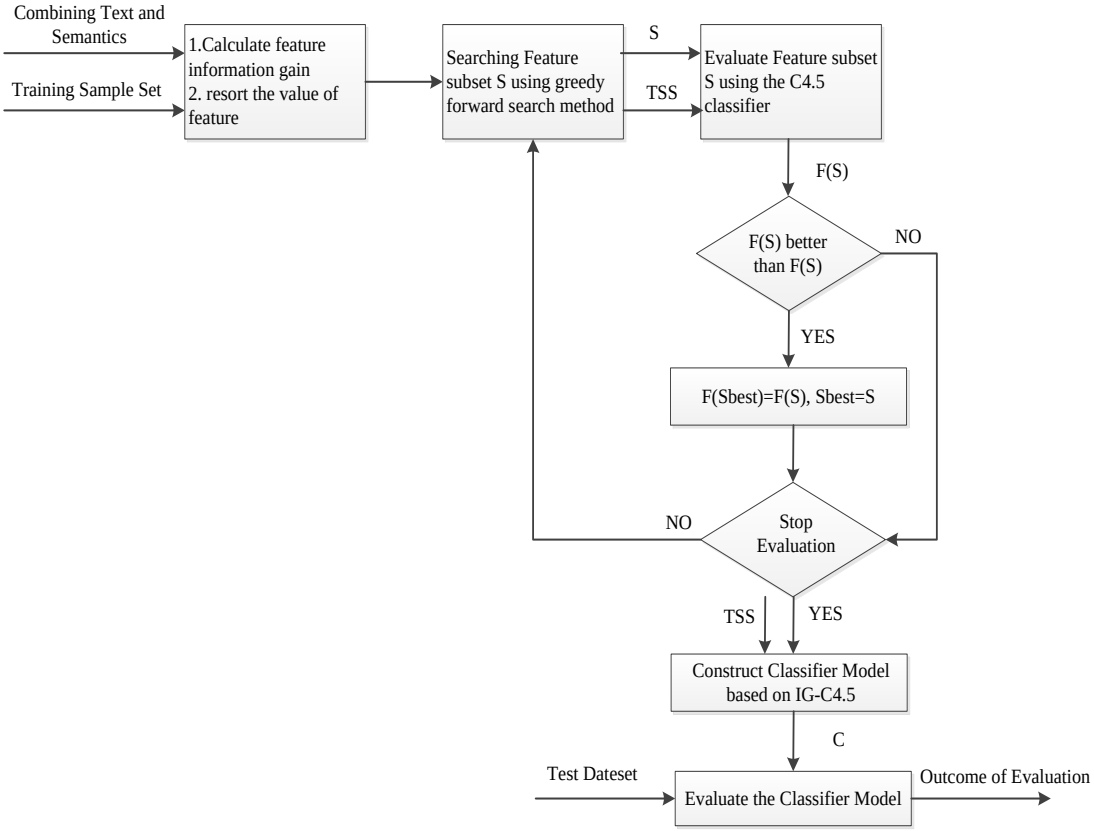


**Figure 1. The Flow of Semantic Web Services Classification Approach based on IG-C4.5**

Second, in each loop, GFS (Greedy forward search) search strategy is used to generate new feature subset S, with C4.5 classification algorithm as evaluator, by 5-fold cross-validation classification accuracy rate F(S) is calculated as a performance evaluation value. If $F(S)<F(S_{best})$ then $F(S_{best})=F(S)$, else $S_{best}=S$. the loop terminates when the value of the feature subset has reached a pre-standard or maximum number of iterations has been reached.

Finally, C4.5 classifier is built according to the feature subset $S_{best}$, which will be used to the semantic web services classification.

## 5. Experiments and Analysis

### 5.1. Experimental Data

The OWLS-TC ver. 2.2 collection is used, which is the only publicly available collection with a relatively large number of advertisements, and it has been used in many research efforts. It consists of 1007 OWL-S advertisements, without any additional modification. The advertisements define profile instances with simple atomic processes, without pointing to physical WSDL descriptions. The advertisements are also classified in seven categories, namely Travel, Education, Weapon, Food, Economy, Communication, and Medical.

### 5.2. Evaluation Method

To assess the effectiveness of the algorithm we proposed, the following series of experiments carried out. The main experiment consists of two parts:

1. Verify that affect the combination feature set of the service classification results
2. Verify the effective of the service classification algorithm we proposed

In evaluating classifiers effect, the experiment used three basic indicators

1. accuracy=the number of samples been correctly identified/total number of samples
2. recall rate= the number of specific category samples been correctly identified/the number of such samples
3. precision= the number of specific category samples been correctly identified / the number of specific category samples been identified by classifier

### 5.3. Results

The same WEB service data sets is classified, using C4.5 classifier based on textual features, semantic features, combination features. The classification accuracy rate respectively reached 89.77%, 92.55%, 93.25%.which shows that, more information is include in the combination feature, resulting in better overall classification. The performance comparison of recall rate as shows in Table1, the performance comparison of precision as shows in Table 2.

**Table 1. Performance Comparison on Recall Rate**

| *category* | *Textual feature* | *Semantic feature* | *Combination feature* |
|---|---|---|---|
| Travel | 0.82 | 0.95 | 0.88 |
| Economy | 0.91 | 0.97 | 0.96 |
| Education | 0.96 | 0.96 | 0.97 |
| Food | 0.73 | 0.79 | 0.73 |
| Communication | 0.93 | 0.97 | 0.97 |
| Medical | 0.81 | 0.63 | 0.81 |
| Weapon | 0.90 | 0.90 | 1 |

**Table 2. Performance Comparison on Precision**

| *category* | *Textual feature* | *Semantic feature* | *Combination feature* |
|---|---|---|---|
| Travel | 0.98 | 0.99 | 1 |
| Economy | 0.98 | 0.98 | 0.98 |
| Education | 0.77 | 0.82 | 0.83 |
| Food | 0.80 | 1 | 1 |
| Communication | 0.92 | 0.98 | 1 |
| Medical | 0.97 | 0.96 | 0.94 |
| Weapon | 0.97 | 0.95 | 0.93 |

### 5.4. The Performance Evaluation of the Algorithm We Proposed

After combination of textual and semantic feature of the OWLS-TC ver. 2.2 collection, the total number of the feature reach to 851. In the experiment, we finished the feature selection using the IG-C4.5 algorithm, the remaining features in the feature vector is 25,

To verify the IG-C4.5 classifier based on combination feature set of classification results, experimental build a classifier on combination feature set selection, use half of the cross-validation to assess the overall results. After using the methods, taken overall classifier accuracy rate was 94.7%, the overall classification of feature selection is better than before, and the use of only 25 data characteristic, the amount of data is significantly reduced. Various categories of recall and precision rates are shown below.

 

**Table 3. Various Categories of Recall and Precision**

| *category* | *Recall rate* | *precision* |
|---|---|---|
| Travel | 0.921 | 0.962 |
| Economy | 0.97 | 0.978 |
| Education | 0.945 | 0.893 |
| Food | 0.788 | 1 |
| Communication | 0.966 | 0.982 |
| Medical | 0.932 | 0.944 |
| Weapon | 1 | 0.929 |

## 6. Conclusion

In this paper introduces a natural language processing methods, combined with semantic information for classification of semantic WEB service conducted a preliminary exploration. After combination of textual and semantic information, semantic service description sample data show significant sparse, high-dimensional nature. According to those features, we propose a data reduction technique, a new service classification. Experiments show that the proposed method can significantly reduce the amount of data to be processed, promote the efficiency of computing storage, while the effect on the detection of minimal negative impact.

## Acknowledgement

This Project is supported by the Innovation project of cultivating graduate of Jiangsu Province（Project Number：CXZZ11_0399）.

## References

[1] E. Christiensen, F. Curbera, G. Meredith and S. Weerawarana, “Web Service Description Language (WSDL)”, v1.1. http://www.w3.org/tr/wsdl.
[2] K. Ioannis, M. Georgios, T. Grigorios, B. Nick and V. Ioannis, “On the Combination of Textual and Semantic Descriptions for Automated Semantic Web Service Classification”. IFIP Advances in Information and Communication Technology, **(2009),** pp. 95-104.
[3] B. Marcello, C. Gerardo, D.P. Massimiliano and S. Rita, “An approach to support web service classification and annotation”. Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service, Washington, DC, USA, **(2005)**, pp. 138–143.
[4] A. Hess and N. Kushmerick, “Learning to attach semantic metadata to web services”, The Semantic Web - Proc. Intl. Semantic Web Conference (ISWC 2003). **(2003)**, pp. 258–273.
[5] N. Oldham, C. Thomas, A. Sheth and K. Verma, “METEOR-S Web Service Annotation Framework with Machine Learning Classification”, Lecture Notes in Computer Science, **(2005)**, pp. 137-146. Semantic Web Services and Web Process Composition - First International Workshop, SWSWPC 2004.
[6] S. Saha, C.A. Murthy and S.K. Pal, “Classification of web services using tensor space model and rough ensemble classifier”. Foundations of Intelligent Systems, 17th International Symposium, ISMIS, Toronto, Canada, **(2008)**.
[7] A. Heß, E. Johnston and N. Kushmerick, “ASSAM: A tool for semi-automatically annotating semantic web services”, The 3rd International Semantic Web Conference, Hiroshima, Japan, **(2004)**.
[8] C. Miguel Angel and C. Pablo, “Semi-automatic semantic-based web service classification”, Business Process Management Workshops, BPM 2006 - International Workshops, BPD, BPI, ENEI, GPWW, DPM, semantics4ws, Proceedings, Vienna, Austria, **(2006)**.
[9] C. Kiefer and A. Bernstein, “The creation and evaluation of isparql strategies for matchmaking”, Proceedings of the 5th European Semantic Web Conference. LNCS, Springer Verlag, Heidelberg, Berlin, **(2008)**.
[10] M. Klusch, P. Kapahnke and B. Fries, “Hybrid semantic web service retrieval: A case study with OWLS-MX”, International Conference on Semantic Computing, Los Alamitos, CA, USA, **(2008)**.
[11] J. R. Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers, **(1992)**.

## Authors

**Chaokai He**, he received the B.S. degree in Computational mathematics from Hebei university and the M.E degree in Computer application technology from china university of mining and technology, china in 2002 and 2005 respectively. He is currently working toward the Ph.D. degree. He is currently researching on network security, trust and reputation computing for Peer-to-Peer network.

**Meng Wu**, he received his B.S. M.S. and Ph.D. degrees in communication engineering and computer science from Zhejiang University, Shanghai Jiaotong University, Southeast University, in 1985, 1990 and 1993, respectively. Currently, he is a professor and Ph.D. supervisor of Nanjing University of Posts and Telecommunications. His main research areas are wireless communications, secure network coding, sensor network and the related information security. He has published more than 100 papers in journals and international conferences.