RESEARCH ARTICLE OPEN ACCESS

Web Service Information Discovery Using Semi-supervise Approach

ROSHAN O. RAJPAYLE, Prof. S.R.Tayde
Sanmati engineering college, Washim, +918275311865 r.o.rajpayle@sanmati.in
Sanmati engineering college, Washim. +918275311865 s.tayade@sanmati.in

ABSTRACT

Web services are playing an important role in e-business and e-commerce applications. As web service applications are interoperable and can work on any platform, large scale distributed systems can be developed easily using web services.

Now a day the Internet has become the largest marketplace in the world, and online advertising is very popular with numerous industries, including the traditional mining service industry where mining service advertisements are effective carriers of mining service information.

However, service users may encounter three major issues- heterogeneity, ubiquity, and ambiguity, when searching for mining service information over the Internet. And as the number of Web Services is increased, finding the best service according to users requirements becomes a challenge. The Semantic Web Service discovery is the process of finding the most suitable service that satisfies the user request. A number of approaches to Web Service discovery have been proposed, so here it proposes the Semi-supervised for web service discovery.

The idea of semi-supervised learning is to learn not only from the labeled training data, but to exploit also the structural information in additionally available unlabeled data. This dissertation review existing semi-supervised approaches, and propose an evolutionary algorithm suited to learn interpretable service information discovery rules from partially labeled data.

GENERAL TERMS

Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.
Semi-supervised technique falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location). The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semi-supervised learning is also of theoretical interest in machine learning and as a model for human learning.

KEYWORDS

Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.

Semi-supervised technique falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location). The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semi-supervised learning is also of theoretical interest in machine learning and as a model for human learning.

As in the supervised learning framework, we are given a set of $l$independently identically distributed examples
$x_1, \ldots, x_l \in X$ With corresponding labels $y_1, \ldots, y_l \in Y$.

Additionally, we are given $u$unlabeled examples$x_{l+1},\dots,x_{l+u}\in X$.

Semi-supervised learning attempts to make use of this combined information to surpass the classification performance that could be obtained either by discarding the unlabeled data and doing supervised learning or by discarding the labels and doing unsupervised learning.

Semi-supervised technique may refer to either transductive learning or inductive learning.

The goal of transductive learning is to infer the correct labels for the given unlabeled data $x_{l+1},\dots,x_{l+u}$only.

The goal of inductive learning is to infer the correct mapping from $X$to Y.

Intuitively, we can think of the learning problem as an exam and labeled data as the few example problems that the teacher solved in class. The teacher also provides a set of unsolved problems. In the transductive setting, these unsolved problems are a take-home exam and you want to do well on them in particular. In the inductive setting, these are practice problems of the sort you will encounter on the in-class exam.

I. Introduction

The service users may encounter three major issues while searching for mining service information over the internet that are Heterogeneity, Ubiquity, and Ambiguity. First is heterogeneity in the real world, many schemes have been proposed to classify the services from various perspectives, including the ownership of service instruments, the effects of services, the nature of the service act, delivery, demand and supply and so on. Nevertheless, there is not a publicly agreed scheme available for classifying service advertisements over the Internet. Furthermore, many commercial product and service search engines provide classification schemes of services with the purpose of facilitating a search, they do not really distinguish between the product and the service advertisement instead, and they combine both into taxonomy.

The second one is ubiquity Service advertisements can be registered by service providers through various service registries, including global business search engines, such as Business.com2 and Kompass3, local business directories, such as Google™ Local Business Center and local Yellow pages, domain-specific business search engines, such as healthcare, industry and tourism business search engines, and search engine advertising, such as Google™6 and Yahoo!®7 Advertising Home. These service registries are geographically distributed over the Internet. The last issue is ambiguity it includes most of the online service advertising information is embedded in a vast amount of information on the Web and is described in natural language, therefore it may be ambiguous. Moreover, online service information does not have a consistent format and standard, and varies from Web page to Web page. Mining is one of the oldest industries in human history, having emerged with the beginning of human civilization. Mining services refer to a series of services which support mining, quarrying, and oil and gas extraction activities.

Service discovery is an emerging research area in the domain of industrial informatics, which aims to automatically or semi-automatically retrieve services or service information in particular environments by means of various IT methods. Many studies have been carried out in the environments of wireless networks and distributed industrial systems. However, few studies have been planned for industrial service advertisement discovery in the Web environment, by taking into account the heterogeneous, ubiquitous and ambiguous features of service advertising information.

II. Approaches For Semi-Supervised Learning Generative Model Approach

Generative approaches to statistical learning first seek to estimate, $p(\mathrm{x}/\,\mathrm{y})$the distribution of data points belonging to each class. The probability $p(\mathrm{y}/\,\mathrm{x})$that a given point $x$has label $y$is then proportional to $p(\mathrm{x}/\,\mathrm{y})\,\mathrm{p(y)}$by Bayes' rule. Semi-supervised learning with generative models can be viewed either as an extension of supervised learning (classification plus information about $p(\mathrm{x})$) or as an extension of unsupervised learning (clustering plus some labels).

Generative models assume that the distributions take some particular form $p(\mathrm{x}/\,\mathrm{y},\theta)$ parameterized by the vectorθ. If these assumptions are incorrect, the unlabeled data may actually decrease the accuracy of the solution relative to what would have been obtained from labeled data alone. However, if the assumptions are correct, then the unlabeled data necessarily improves performance.

The unlabeled data are distributed according to a mixture of individual-class distributions. In order to learn the mixture distribution from the unlabeled data, it must be identifiable, that is, different parameters must yield different summed distributions. Gaussian mixture distributions are identifiable and commonly used for generative models.

The parameterized joint distribution can be written as $p(\mathrm{x}/\,\mathrm{y},\theta)=\mathrm{p}(\mathrm{y}/\,\theta)\,\mathrm{p}(\mathrm{x}/\,\mathrm{y},\theta)$ by using the Chain rule. Each parameter vector $\theta$is associated with a decision function $f\theta(\mathrm{x})=\mathrm{argmaxp}(\mathrm{y}/\,\mathrm{x},\theta)$. The parameter is then chosen based on fit to both the

labeled and unlabeled data, weighted by :

$$\arg\max(\log\mathrm{p}(\{\mathrm{x}_{\mathrm{i}},\mathrm{y}_{\mathrm{i}}\}_{\mathrm{i}=1}^{\mathrm{l}})\,|\,\theta)+\lambda\log\mathrm{p}(\{\mathrm{x}_{\mathrm{i}}\}_{\mathrm{i}=\mathrm{l}+1}^{\mathrm{l}+\mathrm{u}}\,|\,\theta))$$

2.1 Low-Density Separation Approach

Another major class of methods attempts to place boundaries in regions where there are few data points (labeled or unlabeled). One of the most commonly used algorithms is the transductive support vector machine, or TSVM (which, despite its name, may be used for inductive learning as well). Whereas support vector machines for supervised learning seek a decision boundary with maximal margin over the labeled data, the goal of TSVM is a labeling of the unlabeled data such that the decision boundary has maximal margin over all of the data. In addition to the standard hinge loss $(1-yf(\mathrm{x}))$ for labeled data, a loss function $(1-|f(\mathrm{x})|)$ is introduced over the unlabeled data by letting $y=sign f(\mathrm{x})$. TSVM then selects $f*(\mathrm{x})=\mathrm{h}*(\mathrm{x})+\mathrm{b}$ from a reproducing kernel Hilbert space $\mathcal{H}$by minimizing the regularized empirical risk:

$$f*=\arg\min(\sum_{i=1}^{l}(1-\mathrm{y}_{\mathrm{i}}\mathrm{f}(\mathrm{x}_{\mathrm{i}}))_{+}+\lambda_{1}\|h\|_{X}^{2}+\lambda_{2}\sum_{i=i+1}^{l+u}(1-|\mathrm{f}(\mathrm{x}_{\mathrm{i}})|)_{+})$$

Other approaches that implement low-density separation include Gaussian process models, information regularization, and entropy minimization (of which TSVM is a special case).

2.2 Graph-Based Approach

Graph-based methods for semi-supervised learning use a graph representation of the data, with a node for each labeled and unlabeled example. The graph may be constructed using domain knowledge or similarity of examples; two common methods are to connect each data point to its $k$nearest neighbors or to examples within some distanceϵ. The weight $W_{ij}$of an edge between $x_i$and $x_j$is then set to.

$$e^{\frac{-\|x_i-x_j\|^2}{x}}$$

Within the framework of manifold regularization, the graph serves as a proxy for the manifold. A term is added to the standard Tikhonov regularization problem to enforce smoothness of the solution relative to the manifold (in the intrinsic space of the problem) as well as relative to the ambient input space. The minimization problem becomes

$$\underset{f\in\mathcal{H}}{\operatorname{argmin}}\left(\frac{1}{l}\sum_{i=1}^{l}V(f(x_i),y_i)+\lambda_A||f||_{\mathcal{H}}^2+\lambda_I\int_{\mathcal{M}}f(x)||\nabla_{\mathcal{M}}f(x)||^2dp(x)\right)$$

Where $\mathcal{H}$is a reproducing kernel Hilbert space and $\mathcal{M}$is the manifold on which the data lie. The regularization parameters $\lambda_A$and $\lambda_I$control smoothness in the ambient and intrinsic spaces respectively. The graph is used to approximate the intrinsic regularization term. Defining the graph Laplacian $L=D-W$ where

$D_{ii}=\sum_{j=1}^{l+u}W_{ij}$ and $\mathbf{f}$the vector$[f(x_1)\dots f(x_{l+u})]$

The Laplacian can also be used to extend the supervised learning algorithms regularized least squares and support vector machines (SVM) to semi-supervised versions Laplacian regularized least squares and Laplacian SVM.

III. Proposed Work

To design a semi-supervised approach by aggregating the unsupervised approach and the supervised approach, with the purpose of achieving the best service information discovery with the help of various algorithms and tools of Semi-supervised learning to implement the web discovery, while keeping the optimal performance without considering the limitation of the training data set.

3.1 System Description/Architecture

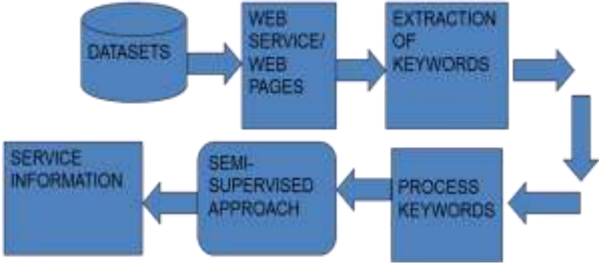

Figure 3.1: System Description

IV. Details Of Blocks

Datasets: Service information discovery includes various datasets that can be used as a input. It may be URL or web pages.

4.1 Web Service/Web Pages:

A Web Service is a software application identified by a URL, whose interfaces and binding are capable of being defined, described and discovered by XML artifacts and supports direct interactions with other software applications using XML based messages via internet-based protocols' will include the web pages or web services. It also includes the various web sites that provide the web service.

4.2 Extraction Of Keywords:

It will extract the keywords from the web service by using various algorithms.

Process Keywords Discovery: The keyword discovery includes various contents with respect to what properties should be described? And How to efficiently query against them? The process keywords depend on Composition, Invocation and Monitoring.

It include various features that are specifying goals of composition, Specifying constraints on a composition, building a composition, analysis of compositions as well as keeping enactments separated and providing transactional guarantees, with the help of monitoring with recovering from failed enactments.

V. SEMI-SUPERVISED APPROACH

The following system shows the semi-supervised approach for web service discovery

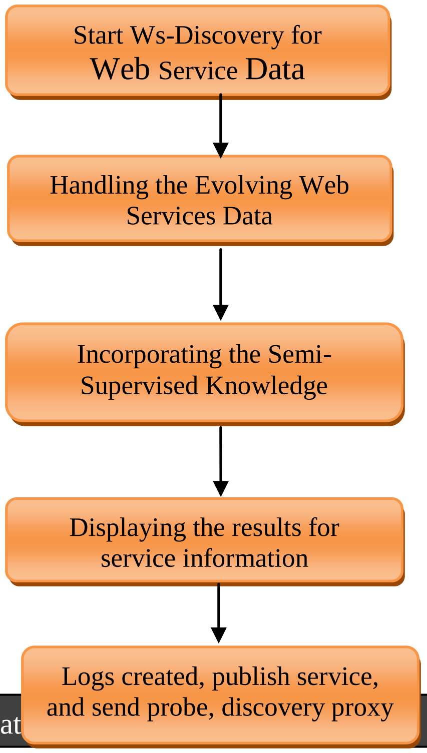
Start Ws-Discovery for Web Service Data
Handling the Evolving Web Services Data
Incorporating the Semi-Supervised Knowledge
Displaying the results for service information
Logs created, publish service, and send probe, discovery proxy

Figure 5.0: Semi-supervised approach

5.1 Start Ws-Discovery for Web Service Data

We represent the web services data with the starting of the web service discovery using a algorithm and techniques.

5.2. Handling the Evolving Web Services Data

In the evolving environment the data changes at each time step. The number of terms, operations and services that define the web service may increase or decrease or the same words might be used in different domains to define a new web service. The clustering algorithm must handle these evolving changes in the data at every time step and perform clustering of the current data in accordance with the historic data .That is, the current clustering must not differ too much from the previous clustering results, and should be able to capture the shift or drift over multiple time steps .We use a parameter α which takes on values between 0 and 1, and specifies the emphasis on current or historic data. Lower value indicates more emphasis towards historic data. As a result, current data is clustered in accordance with historic data. Similarly, a value closer to 1 indicates that more weight is assigned to the current time step data and the clustering is performed according to the current data. There are also different scenarios that need to be handled in the evolving environment as we discuss later.

5.3. Incorporating the Semi-Supervised Knowledge

The user provided knowledge about the web services is incorporated in the form of must link and cannot link constraints on the web service terms. These Constraints ensure that similar terms that define a web services belonging to the same cluster must be placed together and similar terms that define web services belonging to different clusters must be placed apart. The current semi-supervised knowledge will guide the clustering algorithm to cluster similar web services into one cluster.

5.4. Displaying The Results For Service Information

Most of the real world datasets are sparse in nature. In the web service datasets, we have few terms, operations and services that define a web service. To efficiently compute the lager datasets we use the semi-supervised approach.

5.4. Logs created, publish service, and send probe, discovery proxy

In this section the web service discovery created the logs and publishes the service and also sends the probe .It uses the Discovery proxy for web service discovery.

VI. CONCLUSION

The Dissertation work studies the different approaches for web service discovery. It also gives the detailed study of Semi-supervised learning technique. Thus we have solved the issues that lead to service information discovery. Thus UDDI manages the discovery of Web services by relying on a distributed registry of businesses and their service descriptions implemented in a common XML format. The web service discovery mechanisms with Semi-supervised approach discussed above try to achieve objective to enhance efficiency in the matching and binding appropriate web service. The work on discovery mechanisms tries to obtain resulting mechanism not only applicable to web services, but web-based or other software components in general. We have presented a novel approach to perform semi-supervised service information discovery of web services. The semi-supervised knowledge is incorporated in the evolving data by applying the algorithm. The state of the art in knowledge discovery techniques is dominated by supervised approaches. However, in many current real-world problems the assignment of labels for all objects is a severe problem and thus calls for semi-supervised methods.

VII. Acknowledgement

It gives me immense pleasure to express my gratitude to **Prof.S.R.Tayde**, who provided me constructive criticism and a positive feedback during this project work.

I am indebted to Prof. S.A.Vyawhare M.E. Department and other teaching non-teaching staff who were always there whenever I needed any help. Without them and their co-operation, completion of this project work would have been inevitable and their presence behind me is totally indispensable.

I also express my sincere gratitude to our respected Principal ,**Dr.M.S.Kathne** for providing me necessary facilities.

My sincere thanks to the staff of **"Computer Engineering"**, without their help, I could not have even conceived the accomplishment of this report. This work is virtually the result of their inspiration and guidance. I would also like to thank the entire Library Staff and all those who directly or indirectly were the part of this work. I am also thankful to my parents whose best wishes are always with me.

REFERENCES

[1]. H. Dong, F. K. Hussain, and E. Chang, "A service search engine for the industrial digital ecosystems," IEEE Trans. Ind. Electron., vol. 58, no. 6, pp. 2183–2196, Jun. 2011

[2]. Mining Services in the US: Market Research Report IBISWorld2011.

[3]. B. Fabian, T. Ermakova, and C. Muller, "SHARDIS – A privacy-enhanced discovery service for RFID-based product information," IEEE Trans. Ind. Informat., to be published.

[4]. M. Ruta, F. Scioscia, E. D. Sciascio, and G. Loseto, "Semantic-based enhancement of ISO/IEC 14543–3 EIB/KNX standard for building automation," IEEE Trans. Ind. Informat., vol. 7, no. 4, pp. 731–739, Nov.2011

[5]. M. Delamer and J. L. M. Lastra, "Service-oriented architecture for distributed publish/subscribe