

MINISTRY OF SCIENCE AND HIGHER EDUCATION  
OF THE RUSSIAN FEDERATION  
Federal state autonomous educational institution of higher education  
“South Ural State University (National Research University)”  
Polytechnic institute  
Power engineering faculty  
Industrial Heat Power Engineering department  
Classifier number and the name of the Master's program  
13.04.01 “Heat Power Engineering and Heat Engineering”

THESIS WORK  
IS VERIFIED BY

Director  
of Repair & constructions Ltd.

\_\_\_\_\_ K.A. Khasanov  
« \_\_\_\_\_ » \_\_\_\_\_ 2020

ALLOW TO DEFEND

Head of department  
“Industrial Heat Power Engineering”  
Candidate of technical sciences,  
associate professor

\_\_\_\_\_ K.V. Osintsev  
« \_\_\_\_\_ » \_\_\_\_\_ 2020

**Research on service discovery based on semantic Web**

THESIS WORK OF THE MASTER'S PROGRAM  
«HEAT POWER ENGINEERING»  
SUSU-13.04.01.2020. 290.06.EN TW

Head of the Master's program,  
Candidate of technical sciences,  
associate professor

\_\_\_\_\_ K.V. Osintsev  
« \_\_\_\_\_ » \_\_\_\_\_ 2020

Head of work,  
Grand PhD, professor

\_\_\_\_\_ A.A. Alabugin  
« \_\_\_\_\_ » \_\_\_\_\_ 2020

Author of work,  
Student of Master's program  
of P-284 group

\_\_\_\_\_ Y. Sang  
« \_\_\_\_\_ » \_\_\_\_\_ 2020

Chelyabinsk 2020

## ANNOTATION

- Y. Sang. Research on service discovery based on semantic Web
- Chelyabinsk: SUSU, PI, PEF; 2020, 57 p., 14 figure, references
- 54, 10 slides of presentation

The main research content of this paper is to face the problem of semantic Web service discovery, and it is expected to improve the clustering algorithm service by combining the document topic model to perform service discovery. And the article finally introduces the application of semantic-based Web services in thermal power.

TW purpose – Aiming at the description method of semantic-based services, we have studied the leading achievements in the field of semantic web-based service discovery at home and abroad, optimized and improved on this basis, put forward our own research ideas, and improved the service discovery algorithm to make it good scalability and efficiency. And based on the SWOT matrix, it analyzes the application of semantic-based Web services in thermal power.

TW contains: the development of semantic Web discovery; word vector, word2vec model; LDA theme model; clustering algorithm; service matching; service discovery processing; SWOT; thermal power; Gantt chart.

					<i>13.04.01.2020.290.06.EN TW</i>						
	<i>Page</i>	<i>Document #</i>	<i>Signature</i>	<i>Date</i>							
<i>Student</i>	<i>Y.Sang</i>				<i>Research on service discovery based on semantic Web</i>			<i>Letter</i>	<i>Page</i>	<i>Pages</i>	
<i>Head of work</i>	<i>Alabugina A.A.</i>			<i>T</i>				<i>W</i>	<i>3</i>	<i>57</i>	
<i>N.controller</i>	<i>Alabugina R.A.</i>			<i>SUSU Department «Industrial Heat Power Engineering»</i>							
<i>Head of dep.</i>	<i>Osintsev K.V.</i>										

## TABLE OF CONTENTS

ABSTRACT.....	6
1 INTRODUCTION .....	7
1.1 The background of the subject and the purpose and significance of the Research .....	7
1.1.1 Background.....	7
1.1.2 Research purpose and significance.....	8
1.2 Research status at home and abroad .....	8
2 OVERVIEW OF THEORY RELATED TO SEMANTIC WEB SERVICE.....	11
2.1 Web service.....	11
2.1.1 WSDL .....	12
2.1.2 UDDI .....	13
2.1.3 SOAP .....	13
2.2 Semantic Web services .....	14
2.2.1 Semantic Web .....	14
2.2.2 Ontology .....	15
2.2.3 Overview of semantic Web services .....	15
2.2.4 Semantic Web service description language .....	16
2.3 Similarity analysis of semantic Web services .....	17
2.3.1 Similarity based on semantic distance.....	18
2.3.2 Similarity based on information content .....	19
2.3.3 Attribute-based similarity .....	20
2.4 Summary of this chapter .....	20
3 SERVICE PREPROCESSING BASED ON WORD2VEC AND LDA....	21
3.1 Semantic Web service processing .....	21
3.1.1 Web service semantic description abstract definition .....	21
3.1.2 Service description preprocessing .....	21
3.2 Word vector research based on word2vec.....	22
3.2.1 Word vector .....	22
3.2.2 Word Frequency-Inverse file frequency.....	23
3.2.3 Word2Vec model.....	24
3.2.4 Semantic expansion .....	26
3.3 LDA theme model .....	27
3.3.1 Multinomial distribution and dirichlet distribution .....	27
3.3.2 Model introduction .....	28
3.3.3 Parameter estimation .....	29
3.4 Summary of this chapter .....	31
4 WEB SERVICE DISCOVERY METHOD BASED ON CLUSTERING.	32
4.1 Clustering algorithm analysis .....	32
4.1.1 Evaluation of clustering effectiveness.....	32

4.1.2	Classification of clustering algorithms .....	33
4.1.3	K-means clustering .....	34
4.1.4	Hierarchical clustering.....	35
4.2	KMHC clustering.....	36
4.2.1	Contour coefficient .....	36
4.2.2	KMHC clustering.....	37
4.3	Semantic Web service discovery .....	38
4.3.1	Service matching.....	38
4.3.2	Service discovery process.....	40
4.4	Experimental analysis .....	41
4.4.1	Experimental settings.....	41
4.4.2	Results analysis.....	42
4.5	Summary of this chapter .....	45
5	USE OF TECHNOLOGY OF THE SEMANTIC WEB SERVICES .....	46
	IN WORK OF THERMAL POWER PLANT .....	46
5.1	The analysis strong and weaknesses of technology based on the semantic Web services, opportunities and threats of its application use of technology in work of thermal power plant .....	46
5.2	Gantt's schedule of actions for implementation of technology based on the semantic Web services in work of thermal power plant .....	49
5.3	Summary of this chapter .....	49
6	SUMMARY AND OUTLOOK.....	51
6.1	Summary .....	51
6.2	Outlook .....	52
	CONCLUSION .....	53
	REFERENCES .....	54

## ABSTRACT

Due to the large number of Web services available on the Internet, the process of Web service discovery is becoming complicated and time-consuming. Web service discovery is based on the description of the target service user, and according to a certain service matching algorithm from the service registration center to find the service that matches the user's demand description. Since the number of Web services is growing rapidly, how to quickly and accurately return the service requested by the service requester is the hotspot of current research. Here we have conducted in-depth research on the problem of semantic Web-based service discovery.

Faced with the massive amount of information provided by Web service providers, before using service discovery, the word2vec method fused with TF-IDF is used to represent the text vector, and the text description part of the Web service is semantically expanded by training the Wikipedia corpus. The model, Gibbs distribution is assigned, and the document-topic probability distribution matrix is obtained. In order to improve the efficiency of service discovery and reduce the service matching time, cluster processing is performed. Aiming at the problem that the K-means algorithm is sensitive to the initial clustering center and the calculation amount of the hierarchical clustering algorithm is particularly large and the processing speed is slow, an improved KMHC clustering algorithm is proposed based on the K-means clustering and hierarchical clustering algorithms. The Web services are clustered according to the service text description information to obtain K cluster clusters. In each cluster cluster obtained, there is a Web service as a cluster center, which represents this cluster cluster. When a service requester requests a service, calculating the Euclidean distance between the requested service and the representative service in the cluster cluster, and determining which service cluster the service belongs to can reduce the matching time. Then, by calculating the similarity of the input and output parameters based on the semantic similarity calculation method of the ontology concept, the functions are matched to obtain the service that meets the service request in the service cluster, and return the matched service to the user. Finally, through experiments, the optimal number of topics under the LDA topic model is first determined, and then when the optimal number of topics K is 20, it is verified that this method has better service discovery efficiency than traditional methods, and the accuracy rate The recall rate has also been improved. Finally, we analyzed the application of semantic-based Web services in thermal power through the SWOT matrix, and showed our work through the Gantt chart.

# 1 INTRODUCTION

## 1.1 The background of the subject and the purpose and significance of the research

The development of information technology makes the network environment more and more complicated, and at the same time, many Web service technologies have emerged. With the in-depth research and development of semantic Web service technology, the emergence of a large number of Web service technologies has brought a lot of convenience to users, but also brought new challenges. For example, how to improve the degree of automation of information integration and realize dynamic and integrated business processing.

### 1.1.1 Background

Due to the popularity and scale of Web applications, the types of services have become more and more abundant. As a standard for remote access, Web services are also changing user needs. Web service discovery refers to the way to quickly and efficiently select the correct matching method from multiple services according to user needs (such as functional requirements) to find the service they need. Service matching is the focus and hotspot in the field of service research, which directly determines the efficiency and performance of users in obtaining services. Changed many restrictions on application communication between users and enterprises. Web service discovery relies on keyword matching between the service and the request. However, due to the limited service description ability based on keywords and grammar, it is impossible to distinguish the case of one-time polysemy. For the problem of data scale and diversified types, service discovery Efficiency is low. In order to improve the flexibility and scalability of Web service discovery, scholars have proposed semantic Web services to promote the automation of Web service publishing, discovery and execution. This paper focuses on the semantic-based service description method, researches the domestic and foreign frontier achievements in the field of semantic Web-based service discovery, optimizes and improves on this basis, proposes its own research ideas, and improves the service discovery algorithm to make it have Good scalability and efficiency.

Semantic Web services add semantics that computers can understand on the basis of traditional services, expand the semantic level of interpretation, and make the entire Internet a common medium for information exchange. Traditional Web services are interpreted using the WSDL [1] (Web Services Description Language) language. WSDL is a syntactic-level Web service description that does not have semantic interpretation. Semantic Web services are explained with OWL-S [10] (Ontology Web Language for Service). The core of semantic Web services is Ontology, which uses Ontology's description and interpretation language. Ontology is a formalized description of shared concepts, which can be used for conceptual reasoning. At present, most of the content described in Web services can only be read and understood by humans. Because seman-

tic information is added to the Semantic Web, through semantic interpretation, the computer can read and analyze the information on the Web before the system can effectively be Become an entity that can interact with the machine. In the process of service matching, make full use of ontology reasoning ability. Due to the introduction of semantic level information and semantic level matching, the results can be more accurate, allowing the computer to read and analyze the information on the Web to coordinate the optimization results.

### 1.1.2 Research purpose and significance

Based on the existing Web service discovery research, this paper gives a cluster-based semantic Web service discovery method. The discovery process should first locate the corresponding service in the registry. We use the proposed clustering algorithm to lock the service in Similar service clusters. Before clustering, the text description content of each Web service is first preprocessed, and then the obtained data set is semantically expanded to establish an LDA topic model, and then clustered. Semantic augmentation helps improve the accuracy of service matching when requesting the service matching process. Calculate the distance between the requested service and the service cluster center according to user service needs, determine which cluster cluster the service is in, and store the services in the cluster into the priority queue. Finally, the concept similarity is calculated according to the input and output parameters of the service's functional attributes, and the optimal service is selected and returned to the user. With the wide application of service-oriented architecture technology and the explosive growth of the number of services on the Internet, the research content of this topic has certain practical application prospects.

### 1.2 Research status at home and abroad

As one of the important implementation methods of service-oriented architecture, Web services are developing rapidly. The advantage of Web services is that it can search a large number of services in the registry according to the different needs of users, and then provide users with functions that meet the needs in the form of services. As the demand for SOA (Service Oriented Architecture) grows, Web services have become an outstanding technology that can provide good solutions for the interoperability of different types of systems. Web services mainly support interoperability, which is its main purpose. It eliminates the shortcomings of existing technologies and becomes popular with the implementation of SOA. Web services have three main platform elements, which are SOAP [4] (Simple Object Application Protocol), XML (Extensible Markup Language), WSDL (Web Services Description Language) and UDDI [6] (Universal Description, Discovery and integrated). The main advantage of using Web services is the interoperability of attributes. All communication between providers and consumers is based on XML. Using XML technology as part of Web services will become reliable. SOAP is a communication protocol between service providers and ser-

vice consumers and UDDI. It is a protocol mainly based on XML for Web services, and is a standard given by W3C. UDDI is a registry for storing Web service information. SC can use the stored information to access the Web service, and can use standard predefined formats to access the information provided in the registry.

How to improve the discovery effect of Web services is a key problem to be solved in the field of service-oriented computing. Traditionally, the discovery of Web services has functional attributes such as input, output, prerequisites and effects, ignoring the multi-faceted docking framework for Web service discovery using service quality parameters [11]. In response to this problem, Zhong Mei et al. [9] extended the OWL-S language and gave an understanding of the neighborhood construction method. For service quality issues, add a description of quality of service (QoS), and propose to use additional relationships, which represent a feasible solution. The search is performed using semantic similarity, a five-level three-stage multi-level matching process is proposed, and it provides a basis for designing a powerful and effective tool for distributed search algorithms. Xu Guopeng et al [12] established a QoS model for progressive adaptation and screening, proposed an improved semantic Web service discovery method within QoS constraints, introduced candidate service management and achieved high quality, and provided the basis for implementing a service configuration plan. Literature [13] proposed the use of fuzzy particle swarm optimization to improve the deficiencies in the existing methods of semantic Web service discovery based on quality of service (QoS). According to the service discovery problem, the position and velocity of the particle are defined. For the immature convergence problem of the particle swarm algorithm, fuzzy theory and incremental inertia factor are introduced to improve the precision of the semantic web service discovery method based on the particle swarm algorithm. Literature [14] believes that current service discovery methods assume that the QoS attributes of services are represented by an accurate real number, without considering the ambiguity of QoS attributes. Due to the ambiguity of certain QoS attributes, they are described by an accurate real number Its value is unreasonable. Therefore, a service discovery algorithm that supports fuzzy QoS is proposed, and the method of using interval numbers is used when describing QoS attributes with ambiguity.

Xia et al. [15] proposed an effective social-like semantic awareness service discovery mechanism called SLSA by imitating human-like social behavior. And through cooperative intelligence, you can discover the services you need in a fast and scalable manner. And considering the semantic similarity and semantic relativity of the two concepts in the domain ontology, fuzzy logic methods are introduced to calculate their relevance for service ranking. Reference [16] takes the idea of semantic similarity calculation of information content into consideration and proposes a method that uses service-based IO (input, output) semantic matching and semantic content-based similarity calculation. Ou Weijie et al [17] improved the shortcomings of the original bipartite graph optimal matching algorithm. Aiming at the problem that the traditional keyword-based and semantic matching method has low recall rate and low efficiency, which can not meet the practical application, the concept similarity calculation based on hierarchical ontology is realized, and a prototype system of Web service discovery is realized ac-



according to the algorithm. Farrag [18] proposed a matching algorithm based on semantic distance (SDMA). The basic idea is to use the measurement of the semantic distance between the user request and the test service as an indicator of the degree of correlation between them, and a concept tree is proposed to facilitate the calculation of the concepts distance. First, extract the I / O concept set of the Web service, and then use the bipartite graph matching method to calculate the most phrase meaning distance of the two concepts in WordNet, measure the similarity between the concepts, and achieve Web service matching. However, due to the concise interface information, it is easy to cause information loss, which may reduce service.

Literature [19] studied the influence of community relations among multiple users on Web service discovery results. According to the principle that the correlation between Web services and user interest background is strong to weak, the community relations between users are decomposed into Preference relationship, cluster relationship and trust relationship give the formal methods of these three relationships. Based on the formalized community relationship, a differential service discovery strategy is proposed, and a web service discovery system framework based on user community relationship is constructed to gradually retrieve or recommend candidate web services. Literature [20] introduced situation factors to learners who were dissatisfied or unstable with the learning services provided by the e-Learning service discovery system, and designed a learning service discovery algorithm-eLSDACA. The algorithm perceives the learner's learning situation, forms the learner's situation ontology, and participates in the process of service discovery. Reference [21] proposes a Web service discovery method based on user context clustering to quickly discover Web services with high user suitability. Incorporating clustering and inverted index technology into Web service discovery algorithms, using BIRCH clustering ideas to cluster user contexts, effectively narrowing the search range of Web services, and inverted index technology can quickly locate services and further optimize the Web The time the service was discovered. Reference [22] matches the user interest model with the academic resource model. The article builds an academic resource model through the citation structure network, LDA topic distribution, and feature word distribution. Combined with the academic resource model, it calculates the user 's interest in the browsed academic resources and calculates its Similarity, and finally the top N academic resources with the highest calculated user interest value are returned to the user. As the current service clustering algorithm can not achieve multi-functional clustering of services, literature [23] proposes a multi-functional clustering method of Web services based on LDA topic model and fuzzy C-means algorithm.

## 2 OVERVIEW OF THEORY RELATED TO SEMANTIC WEB SERVICES

### 2.1 Web service

Web service is a communication mechanism between applications. They have the advantages of independence, low coupling, independence and programmability, and can be released and called. By using Open XML (Extensible Markup Language) to describe, publish, search, tune and define these applications, provide interoperable applications for cross-platform application services and eliminate system heterogeneity [26]. In a deeper sense, Web services are an extension of Web applications. It can be used to describe, publish, search, and call the Web through independent, self-describing modular applications. Typical network services are DNS, FTP, Telnet, WINS, SMTP, etc. Web services use common Internet standards such as HTTP and XML, and are widely used in e-commerce, e-government, company business process electronic and other fields, allowing people to access data on the Web through different terminal devices in different places, such as online booking, Check reservations, online reservations, etc. Therefore, Web services have the advantages of object-oriented technology based on XML. The system structure of the Web service has three main roles. The distribution is the service provider, the service request, and the service registry. The relationship between them is shown in figure 2.1.

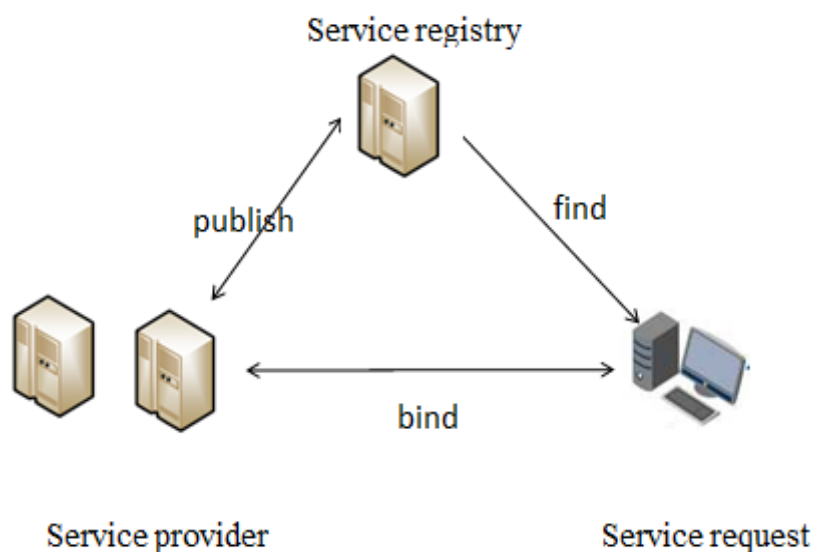


Figure 2.1 – Web service architecture

From a technical point of view, the service provider describes the service through WSDL (Web Service Description Language), and then publishes it on the commercial registration center UDDI, and accesses it through SOAP (Simple Object Access Protocol), which can deploy the Web service as a Web Objects and reusable software mod-

ules, so that developers and e-commerce applications can search and locate the service. When WSDL defines specific deployment access service points, it combines related access points to obtain an abstract overall Web service. Then, it abstracts the request / response message information used during the access and associates the abstract information with a specific transmission protocol and message format [24]. UDDI provides service registration annotations, mainly launched by Ariba, IBM, Intel and other companies. The core component is UDDI commercial registration. The company registers the service together and combines the online services that other companies want to find so that other companies can find the service. For Web services, SOAP is a lightweight protocol for exchanging information. It can remotely call services on other computers, and at the same time send information on the local computer to the remote computer. This is the same as the purpose of the remote call protocol (RPC). of. The parameters of the SOAP transfer method are mainly used to make Web calls through XML documents. The specific contents of WSDL, UDDI, and SOAP are divided into three sections to introduce.

### 2.1.1 WSDL

WSDL [1-3] is an abbreviation of Web Services Description Language (Web Services Description Language). WSDL is used to accurately describe the documents of Web services. It is written in XML and is an interface definition language that can describe the interface information of Web services.

The WSDL document can be divided into two parts: abstract definition and specific description. WSDL describes the Web service and explains how to communicate with the Web service, provides users with detailed interface instructions, specifies the location of the service and the operations provided by the service. The elements contained in the WSDL document are shown in table 2.1.

Table 2.1 – Document composition

Composition	Feature
Port Type	Specific protocol and data format specifications for specific port types
Binding	Bundled protocols and information formats
Port	A single endpoint defined as a combination of binding and network address
Service	A collection of related service access points, including related interfaces, operations, messages, etc
Type	Data type defined container
Message	The abstract type definition of the communication message, which is composed of one or more parts
Operation	Abstract description of the operations supported in the service

WSDL defines four information exchange methods, as is shown in table 2.2.

Table 2. 2 – Information exchange methods

Information exchange methods	Meaning
One-way	Endpoint accepts information
Request-response	The server accepts the request message from the service requester and sends a response message
Solicit-response	The service requesting end sends a message and then receives a reply message
Notification	Endpoint sends message

### 2.1.2 UDDI

UDDI [6-7] is the abbreviation of Universal, Description, Discovery and Integration, which means unification, description, discovery and integration. This is a Web-based, distributed cross-platform description paradigm, which builds a platform-independent and open framework to provide information registration standard specifications for Web services. Its main purpose is to describe services through the Internet, discover business, and integrate business services, create a global standard to provide self-service for each company, so that each company can understand each other, so as to make common progress on the network. UDDI is a public registration and one of the most important public technologies. The working principle of the UDDI registry is to build assembly technology, which is a multi-purpose application.

The system integration framework based on UDDI is composed of three parts: system providing layer, system integration layer and user layer. It includes a third-party integration platform and a metadata that meets the user's dynamic state source acquisition and business requirements, implements special needs supervision, and provides standards-based specifications for description and discovery services.

### 2.1.3 SOAP

SOAP [4-5] (Simple Object Access Protocol) stands for Simple Object Access Protocol. It is a communication protocol. It is lightweight, simple, and based on XML standardization. It is designed to exchange structured and Cured information. The SOAP protocol consists of four parts [25]. SOAP can be used in conjunction with many existing Internet protocols and formats such as Multipurpose Internet Mail Extension Protocol (MIME), Simple Mail Transfer Protocol (SMTP), Hypertext Transfer Protocol (HTTP), and so on. The purpose of SOAP is to transfer data between various systems distributed across the network. When applications and services communicate, the most

common method of exchanging data between the two systems is to use messages. The message sent to the service will call the method provided by the service, and then ask the service to perform a specific operation. The service uses the information contained in the message to perform functions when necessary, and returns the result in another message. These modes define the format of the message sent over the network, including that the message may contain some type of data, and the message must be configured so that another server can correctly interpret it. The SOAP message format includes four parts, namely: SOAP envelope, SOAP encoding Style, SOAP header, SOAP body. SOAP is located above the interconnection protocol and can be used to send and receive data with other networks. Most firewalls can receive service requests, so they are used to allow services to communicate through the firewall.

## 2.2 Semantic Web services

Semantic Web service is a new Web service model, which aims to solve the problem of using semantic analysis to match or search traditional Web services. In order to better understand the computer, the semantic Web service technology combines semantic and Web service technology. Combining ontology technology with Web services, using semantic ontology to model Web services, provide more accurate semantics, and perform service matching based on the similarity of domain ontology concepts. The semantic level describes the service interface, service message, service structure and service interaction. We propose a semantic Web service description language that uses description logic and logical reasoning to perform automatic detection, tracking, and recovery of Web services.

### 2.2.1 Semantic Web

The Semantic Web is a network that describes things in a way that computers can understand. The Semantic Web will define a structure for meaningful content on a web page and create an environment where software agents that jump from one page to another can easily perform complex tasks for users. The Semantic Web describes things in a way that computer applications understand. It is a general framework designed to bring machine-readable information online. When the Semantic Web first appeared, its content was human-readable, and the computer could not understand and manipulate it. The Semantic Web is a network with data at its core, in which information can be understood and processed by machines. "Semantic" means that the computer can understand the meaning behind the rich expressions of numbers, can understand the pictures and links on the web page, and can clearly define the relationship between the web page pointed to by the link and the current web page. The Semantic Web chooses a richer way to express the meaning of the data so that the machine can understand the data. "Network" hopes to interconnect various forms of data to form a huge information network, such as the Internet with linked web pages, but the basic unit is not subdivided. To search or access the Semantic Web, we need "Semantic Web Agent" or "Semantic

Web Service". These "agents" or "services" will help us find what we are looking for on the Semantic Web. The Semantic Web describes the relationship between things (for example, X is part of Y, and A is a member of B) and the attributes of things (such as weight, place of origin, price, raw materials, etc.). The Semantic Network describes network resources using RDF (Resource Description Framework). RDF can describe network information and resources. This is a markup language. Computer programs can search, analyze, and process data in RDF.

### **2.2.2 Ontology**

Ontology is a formal sharing concept and a key factor in solving heterogeneous semantics and systems. It contains rich semantic information and can provide important support for research and related applications in the fields of semantic web, knowledge representation, question answering system and information extraction, information retrieval, etc. Therefore, how to construct the ontology quickly and effectively has a very important research value. Researchers have proposed a variety of effective ontology construction methods from different perspectives. Overall, these ontology construction methods can be divided into manual construction methods, automatic and semi-automatic technology construction methods. Manual ontology construction methods need to rely on experts in the ontology field to participate. However, because this method relies on domain experts' understanding of concepts and relationships, experts in different fields have not reached a consensus. This method has many disadvantages, such as high construction cost, low efficiency, strong subjectivity, and inconvenience for transplantation. The step-by-step construction method of ontology based on automatic and semi-automatic technology does not require manual participation (or requires little participation), and can easily use the latest research results in other research fields (machine learning, artificial intelligence, natural language processing, etc.), More comprehensively and quickly build different data sources for Ontology. In this way, a large number of concepts and relationships between concepts are obtained from a large amount of text, machine-recognizable databases, dictionaries and other text data sources. The advantage of this method is the large amount of operable data, easy access, and the ability to mine more potential knowledge. More and more researchers begin to pay attention to how to effectively use text resources to automatically build ontology.

### **2.2.3 Overview of semantic Web services**

Semantic Web service is independent of the development platform and is a modular application with self-contained and self-describing features. Service providers publish a large number of services on the Internet and use open standards to provide services to users. However, because many Web services lack some necessary semantic description information and the dynamics and heterogeneity of the environment, the basic functions of Web services lack accurate descriptions. Therefore, the problem of Web service in-

teraction can be solved by increasing the semantic information interaction between the service requester and the service provider and adding the semantics to the Web service.

Semantic Web service is a combination of semantic Web technology and Web service technology. Ontology technology is introduced on the basis of Web service. In order to accurately describe the semantics of Web service, the rich semantic description function of Semantic Web and the powerful logic reasoning function are used to enrich Semantic information of Web services. Through the description with semantic information and the reasoning function of the ontology, it is convenient for the computer to understand and access, so that the Web service can be understood by the computer and transparent to the user, and the service can be automatically discovered, combined, monitored and invoked. Figure 2.2 shows the idea of combining semantic Web and Web services into a new semantic Web service technology.

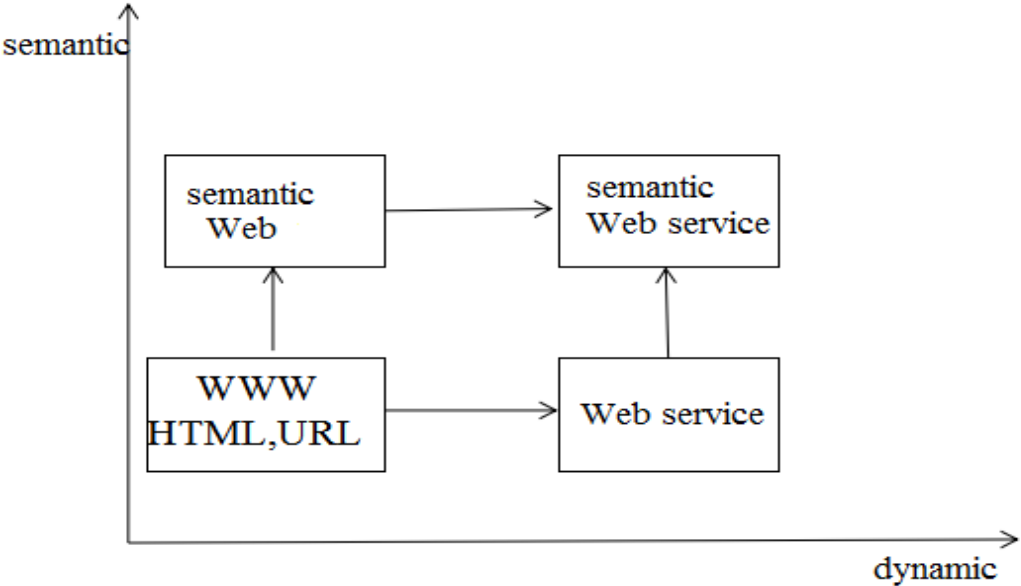


Figure 2.2 – Semantic Web service

The logical reasoning based on semantic ontology can realize the automatic discovery, invocation, matching, management, tracking and realization of Web services, which can realize the direct interaction of describing service interfaces, service structures and service messages on the semantic level. The combination of Web services and Semantic Web technology realizes service-oriented computing and service-oriented architecture.

**2.2.4 Semantic Web service description language**

At present, in the research of semantic Web services, the description languages with greater influence are mainly WSDL, OWL-S, WSMO / WSML, SWSO / SWSL and so on. Among them, OWL-S [32] (Ontology Web Language for Service), which represents

the web service ontology language, is a new generation of markup language built on OWL (Web Ontology Language, ontology web language) launched by Darpa and used to describe web services, Using description logic to achieve its reasoning. It describes the semantic information of Web services according to the W3C standard ontology language. It was once recommended by W3C as a standard semantic Web Service description language as an ontology in 2005. OWL-S's Web service description model can describe semantic Web services. Its basic description of services includes basic information, functional information and attributes of services. OWL-S describes Web service attributes and functions based on ontology, which is a service description specification. In OWL-S, a service is usually composed of Service Profile (ontology of service description information), Service Process (service process), and Service Grounding (service base point). Description. The function description follows in figure 2.3.

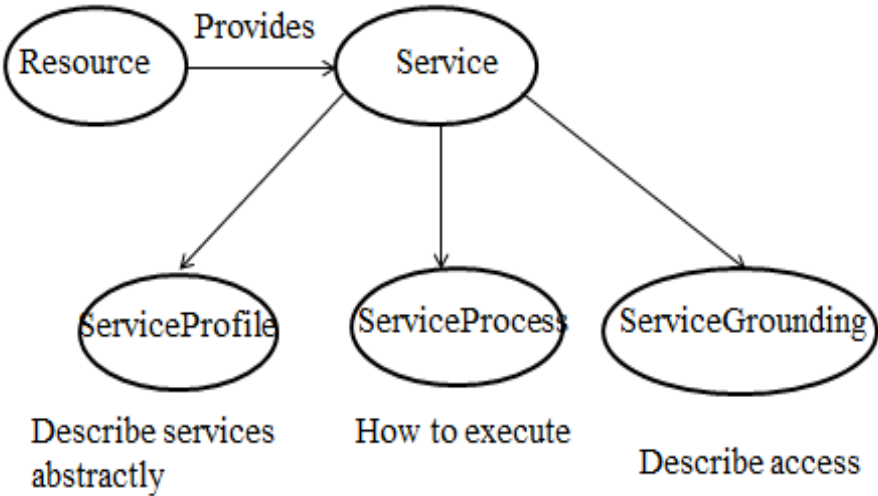


Figure 2.3 – OWL-S service description specification diagram

**2.3 Similarity analysis of semantic Web services**

Semantic similarity refers to calculating the similarity between concepts or language words. These words are not necessarily lexical similar, also known as semantic association. It attaches metrics to the document set or terms that are marked in the document according to semantics. The goal of the Semantic Web is to build a network system based on information semantics rather than grammar. The lack of common words in the document does not mean that they are different or irrelevant. With the development of ontology research and the development of more extensive application of ontology technology, more and more researchers suggest to use structural domain ontology to calculate concept similarity, especially the semantic information in WordNet ontology is widely used in semantic similarity Calculation [28]. Various algorithms are used to calculate the logical similarity to prove whether the documents or sentences with different vocabulary are semantically the same or similar. Semantic matching faces the problems



of polysemy, synonyms, and concept matching, and the semantic similarity measure indicates the taxonomic closeness between concepts.

### 2.3.1 Similarity based on semantic distance

Semantic distance is a mathematical metric that expresses the shortest correlation length in the inheritance relationship or binary relationship chain existing in two different concepts in the same ontology. The shorter the connection path between the two concepts in Ontology, the more similar they are. We calculate the semantic similarity between concepts by calculating the semantic distance between concepts in WordNet. According to the semantic similarity, the matching degree of the two Web services can be measured. The application principle of the semantic similarity algorithm based on semantic distance is: the shorter the semantic distance between concepts, the stronger their semantic similarity. The basic idea of the semantic distance similarity algorithm is to treat the concept in the ontology as a node in the ontology tree. The semantic similarity between two nodes is expressed by the distance between the nodes and the depth of the node position in the ontology tree to express the depth of concept. On the basis of the inheritance relationship of ontology concepts, the connection relationship only refers to the is-a relationship between two concepts. The relationship between the concepts is a unidirectional relationship, which is inversely proportional to the geometric distance between concepts. The closer the two concepts are, the higher the semantic similarity, and the further the distance, the lower the semantic similarity between the concepts.

In order to quantify the matching degree between semantic Web services and make it more accurate, we use [0,1] as the service matching similarity metric to measure the matching degree between Web services. The formula for calculating the semantic distance can be defined as (2.1):

$$Dist(s_i, s_j) = \frac{depth(s_i) + depth(s_j)}{depth(LCA(s_i, s_j))} \quad (2.1)$$

where  $s_i, s_j$  represents two concepts in the ontology tree and  $LCA(s_i, s_j)$  represents the nearest common ancestor of  $s_i, s_j$ .  $Dist(s_i, s_j)$  is used to represent the semantic distance between two concepts and  $depth(s_i)$  represents the depth of the  $s_i$ . Here, the root node depth of the ontology tree is defined as 1, and the depth is increased by 1 for each additional layer. The similarity formula is defined as (2.2) and (2.3):

$$sim(s_i, s_j) = (1 - Dist(s_i, s_j) * \theta)^{\frac{1}{2}} \quad (2.2)$$

$$\theta = \frac{m^2 - 1}{m^2} \left( \frac{m+1}{m} \right)^{-1} \quad (2.3)$$

where among them,  $m$  is the total level of the ontology tree,  $l$  The number of layers where the ontology concept is located.

Formula (2.2) inputs the semantic distance of two ontology, and the output is the semantic similarity. By calling this function, we can preliminarily filter out the corresponding semantic Web services.

### 2.3.2 Similarity based on information content

The key to the method based on information content is to calculate the amount of information (IC), which is based on the evolution of information theory, and uses IC to measure how much information a concept contains.

Resnik et al. [29] believed that the similarity of two concept nodes in the ontology can be converted into the amount of information they contain together. Therefore, the smallest common contained IC of the two concept nodes can represent the similarity between them. To improve the information content of the node with the largest amount of information among the common parent nodes of the ontology concept, a method based on information content is proposed. The formula is as follows (2.3):

$$sim(c_1, c_2) = IC(les(c_1, c_2)) \quad (2.3)$$

Among them,  $les(c_1, c_2)$  represents the closest common parent node of concepts  $c_1$  and  $c_2$ .

Liu [30] and others simulated human thinking and proposed an improved algorithm model (2.4):

$$sim(c_1, c_2) = \frac{\alpha depth(les(c_1, c_2))}{\alpha depth(les(c_1, c_2)) + \beta path(c_1, c_2)} \quad (2.4)$$

where  $depth(les(c_1, c_2))$  represents the depth of the nearest common parent node, and  $path(c_1, c_2)$  represents the shortest path between concepts  $c_1$  and  $c_2$ . Here  $\alpha$  and  $\beta$  are smoothing parameters, the shortest path used to measure the depth of the closest common parent node and the similarity between concepts.

Xu et al. [31] believe that the semantic relationship between ontology concepts should also be used as part of the similarity of information content. Concept pairs mainly have synonymous relations, Is-a relations, Part-whole relations, or Other relations. In comparison, the similarity of information content of concept pairs with synonymous relations should be significantly higher than the latter relations. Define the contribution as (2.5):

$$sr(c_1, c_2) \begin{cases} 1, & \text{Synonymous -relationship} \\ 0.6, & \text{Is - a-relationship} \\ 0.3, & \text{Part - relationship} \\ 0.1, & \text{Other -relationship} \end{cases} \quad (2.5)$$

And give the similarity calculation formula based on information content (2.6):

$$Sim(c_1, c_2) = \frac{2 \times IC(RCPN(c_1, c_2)) \times sr(c_1, c_2)}{IC(c_1) + IC(c_2)} \quad (2.6)$$

### 2.3.3 Attribute-based similarity

The basic principle of the attribute-based semantic similarity calculation model is to measure the similarity of the attribute sets corresponding to the two concepts. The attribute-based semantic similarity is used to calculate the common attributes between transactions, and the concept semantic similarity is calculated according to the characteristics of the vocabulary. This method It is believed that the degree of association between things is proportional to the number of attributes they share. The attribute-based semantic similarity method attempts to solve the problem reflected by the distance-based method. Using the overlapping degree of ontology attributes, the more common attributes of the two concepts, the greater the similarity. Such as the use of synonymous relations in WordNet.

The attribute-based semantic similarity calculation model proposed by Xun Endong [27] et al. Extracts synonyms from WordNet and uses the vector space method to calculate the semantic similarity. First, extract candidate synonyms from the three aspects of WordNet's sense interpretation (Sense Explanation), class information (Class) and synonym word set (Synset), extract feature words, and then calculate the meaning between each concept. Then calculate the distance between attributes in the three feature spaces, and finally calculate the similarity of words based on the similarity of concepts to attribute values.

## 2.4 Summary of this chapter

This chapter mainly introduces the theoretical knowledge related to semantic Web services, first introduces Web services, and introduces the three major technologies of Web services: WSDL, UDDI, SOAP. Secondly, it introduces the Semantic Web Service, first introduces the relevant knowledge of the Semantic Web and Ontology, enriches the semantic information of the Web Service, and introduces the OWL-S, a semantic Web service description language. Finally, the similarity calculation of semantic Web services is explained based on semantic distance, information content, and attribute, which provides a theoretical basis for the following specific applications.

### 3 SERVICE PREPROCESSING BASED ON WORD2VEC AND LDA

Web service preprocessing is an important prerequisite for Web service clustering and service discovery. The results of Web service preprocessing will directly affect the efficiency of service clustering and service discovery. This chapter implements a preprocessing of documents before clustering, and uses the topic model for feature modeling.

#### 3.1 Semantic Web service processing

##### 3.1.1 Web service semantic description abstract definition

We define Web service semantic description abstraction as follows:

Definition 1: The service is represented by a quadruple:  $WSIR = \{iServiceName, itextDescription, iInput, iOutput\}$ , where  $iServiceName$  represents the name of the web service;  $itextDescription$  represents the text description information of the web service;  $iInput = \{Input1, Input2, \dots\}$  Represents the set of input concepts of the Web service;  $iOutput = \{Output1, Output2, \dots\}$  represents the set of output concepts of the Web service.

Definition 2 Service description information textDescription set IDS: IDS is a collection of description information of all services in the Web service registry, that is,  $IDS = \{ItextDescription1, ItextDescription2, \dots\}$ , where  $ItextDescription1 \in WSIR1$ ,  $ItextDescription2 \in WSIR2$ .

##### 3.1.2 Service description preprocessing

###### 1. Feature extraction

When we perform service discovery, we need to parse and obtain the service name, content description information, input and output and other information to get the corresponding WSIR document collection. We use the widely used Jena to parse document information. Table 3.1 shows the input, output, name, and text description parameters of a service obtained by parsing.

Table 3.1-Service parameters

ServiceName	BookPriceService
Input	"#_BOOK"
Output	"#_PRICE"
TextDescription	This service returns list of current purchase prices of a given book title. The prices include both new and used versions of the book.

It can be seen from the definition in 3.1:  $WISR1 = [BookPriceService, This service returns$

list of current purchase prices of a given book title. The prices include both new and used versions of the book. "#\_BOOK", "#\_PRICE"].

2. Word segmentation processing

In order to get the word vector describing the characteristics of the Web service, after obtaining the WSIR document collection, we need to do word segmentation processing on the content in it. For example, when the service name is "Cannon Camera Price Service", after word segmentation, the vocabulary vector is {Cannon, Camera, Price, Service}.

3. Go to stop words and tags

Stop words are words used to express the grammatical relationship between objects or actions. The actual information provided is relatively small, such as also, and, will, etc., which have little meaning for retrieval. Tags like "#\_" It should also be removed. In the process of removing stop words, these words cannot be removed blindly, and the frequency of some feature words should also be noted.

4. Stem reduction

Stem reduction refers to the reduction of words into archetypal form. English words contain many parts-of-speech transformations, such as ing and ed, likes and like, etc., so it is necessary to perform stem reduction. Stem reduction is to merge these different expressions together.

5. Results analysis

6. Natural language processing toolkit (Natural Language Toolki, NLTK) is used for English text preprocessing [33]. Text preprocessing is based on python3 language, and the tool is pycharm. Text preprocessing process:

1. Install the NLTK toolkit-> pip install nltk
2. Word segmentation: call word\_tokenize word segmentation
3. To stop words: call stopwords function
4. Stemming: call PorterStemmer □

The service text description content vector obtained by WISR1 after processing:  
 Wordvec1 = [service, return, list, current, purchase, price, give, book, title, price, include, new, use, version, book].

**3.2 Word vector research based on word2vec**

**3.2.1 Word vector**

The word vector (Distribution representation) was first proposed by Hitton, which maps the word into a new space, so that the related words are closer to each other in the mathematical sense. For example, the word "like" can be expressed as a vector [0,0,0,0,1,0 ..., 0], which is very helpful for many natural language processing problems. However, when solving practical problems with sparse representation, we often encounter dimensionality disasters and need to reduce the dimensionality. In this paper, we calculate the word frequency-reverse file frequency to reduce the dimensionality, and use Word2vec to represent the word vector. Using the dense matrix obtained in this way not

only solves the dimensional disaster problem, but also can express the linear relationship between words to a certain extent, and mines the related attributes between words, thereby improving the accuracy of vector semantics.

### 3.2.2 Word Frequency-Inverse file frequency

The term frequency-inverse document frequency [40] (Term Frequency-Inverse Document Frequency, TF-IDF) method is the most typical measure of text similarity. TF-IDF is widely used as a weighting technique for finding information and text mining. At the same time, it is used as a statistical method to evaluate the importance of documents or words in a corpus. Word frequency (TF) indicates the number of occurrences of a word in a document, but it is impossible to express the importance of a word using this parameter only. The importance of two words with the same word frequency may vary greatly. In order to indicate the importance of a word, on the basis of word frequency, each word is assigned an importance weight. This weight is called inverse document frequency (IDF), and its size is inversely proportional to the common degree of a word. Search engines often use various forms of TF-IDF weighting to measure the relevance of documents and user searches. We need to calculate the TF-IDF value of each word in the service description. This paper uses this method to reduce the dimension of high-dimensional vectors. The details are as follows:

$tf_{ij}$  is the frequency of feature words in the document, in order to improve the accuracy and precision of weight calculation, we will normalize it and normalize the  $tf_{ij}$  value of each text feature word to [0,1], then the standardized word frequency of the feature word  $t_i$  is as follows (3.1):

$$tf_{ij} = \frac{f_{ij}}{\max(f_{1j}, f_{2j}, \dots, f_{mj})} \quad (3.1)$$

where  $f_{ij}$  indicates the number of occurrences of the word in the document, and  $\max(f_{1j}, f_{2j}, \dots, f_{mj})$  indicates the number of occurrences of the word with the most occurrences in the document.

The frequency of inverse documents of feature words is as follows formula (3.2):

$$idf(i) = \log(N / 1 + df_i) \quad (3.2)$$

where N is the total number of documents in the corpus, and  $df_i$  is the number of documents containing the feature word  $t_i$ .

### 3.2.3 Word2Vec model

With the development of deep learning in recent years, feature word extraction based on neural network models, that is, the way that “word vectors” represent texts, has attracted more and more attention from academia. Each document is composed of many words. For the problem of how to effectively use word vectors to represent documents, literature [38] proposes to use TF-IDF to weight the word segmentation in each document based on word2vec. By calculating TF, select words with high frequency in the document, calculate IDF to filter out words without features, and the rest are used as feature words, excluding words with low frequency, and smoothing the case where the weight is 0. In this way, the high-dimensional vector is subjected to dimensionality reduction processing.

The Word2vec tool mainly includes two models: continuous bag of words (CBOW) and skip-gram [34-35], and two efficient training methods: negative sampling (negative sampling) And hierarchical softmax (hierarchical softmax) [36]. This article uses the CBOW model, optimized based on hierarchical softmax. The principle of CBOW is to predict what the target word is given the context of the target word, where the input layer is  $2w_{in}$  (window) word vectors in the context of the word  $w(t)$ , that is,  $w_{in}$  words of the target word context are selected as input, and the projection layer The vector is the summation of the  $w$  word vectors in the input layer, and the output layer corresponds to a Huffman tree [37]. We use the Huffman tree to find the probability of a word. The Huffman tree has  $N$  leaf nodes matching the words in the dictionary  $D$ , and  $N-1$  non-leaf nodes. When constructing a Huffman tree, each leaf node of the tree represents a word. Suppose that there are 3 branches from the root node to a leaf, each branch represents a binary classification, and each leaf has a unique path from the root to the leaf node. This path is used to estimate the probability of the word represented by the leaf node. Probability is defined as the probability of randomly walking from the root node to the leaf node. Each non-leaf node represents a logistic regression process. It can be stipulated that walking along the right subtree is a positive class, and left subtree is a negative class. The probability of going to the left and right subtrees is expressed by a logistic regression formula. The probability formula is as follows formula (3.3) and (3.4):

$$\lambda_1 = \frac{1}{1 + e^{-w_c \theta}} \quad (3.3)$$

$$\lambda_2 = 1 - \frac{1}{1 + e^{-w_c \theta}} \quad (3.4)$$

Where  $w_c$  is the word vector of the current internal node, and  $\theta$  is the model parameter that needs to be learned. Which word is finally output is determined by the  $\log N$  logistic regression process. The Huffman tree allows words with larger weights to be outp

ut in leaves with a smaller depth, so that a shorter encoding is obtained, and words with a higher frequency will be found at a lower cost.

The CBOW model shows in figure 3.1.

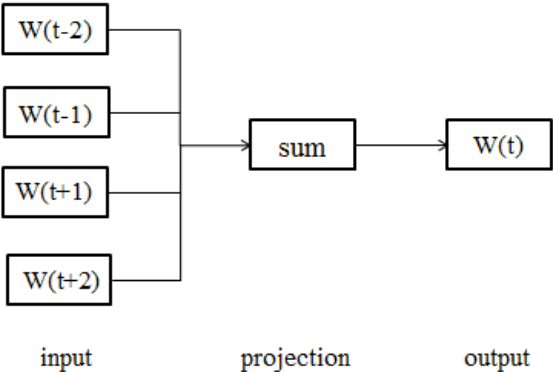


Figure 3.1 – Model CBOW

For the CBOW model,  $w_t$  is the vector after initializing the root node as the input word vector and adding and averaging. The parameters that the model needs to learn: the word vector of each word  $w(t)$  and the  $\theta$  of each internal node of the Huffman tree. Use context  $(w)$  to refer to  $2i$  words in the context of words. In order to maximize the value of  $p(w | \text{context}(w))$ , the result of  $w(t)$  is predicted by the stochastic gradient ascent algorithm. This problem can be modeled with conditional probability, which is given by softmax, so our model is to seek from (3.5) and (3.6):

$$P(w_t | w_{t-i} : w_{t+i}) = \frac{\exp(\bar{v} \cdot v_{w_t})}{\sum_{n=1}^N \exp(\bar{v} \cdot v_n)} \tag{3.5}$$

$$\bar{v} = \frac{1}{2i} \sum_{-i \leq j \leq i, j \neq 0} v_j \tag{3.6}$$

Where  $w_t$  represents the target word to be predicted and  $i$  represents the context size.

For a given sentence, the objective function of the model is to maximize the log-likelihood function of the above formula (3.7):

$$L = \frac{1}{T} \sum_{t=1}^T \log P(w_t | w_{t-i} : w_{t+i}) \tag{3.7}$$

Where  $T$  is the sentence length. Given the context "provides", "the", "best", "type", "produced", "by", the goal of the CBOW model is to predict the probability of the word "honey" (figure 3.2) To achieve such a goal, you need Find the maximum value of formula (3-7).



Figure 3.2 shows the principle of CBOW model is shown.

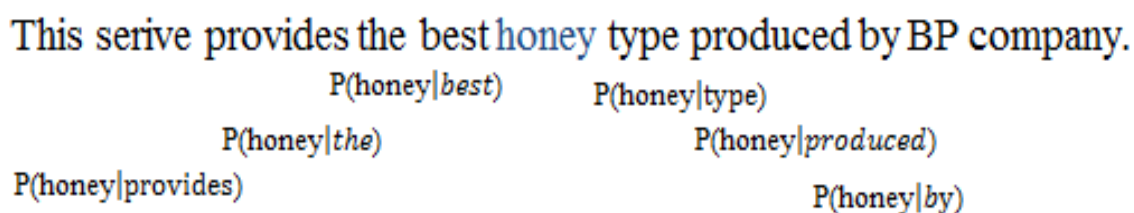


Figure 3.2- Principle of CBOW model

Implementation steps of CBOW model:

1. Construct a vocabulary from the training document;
2. One-hot encoding of vocabulary;
3. Construct a neural network based on the data and train to obtain the model;
4. The parameters (weight matrix) word vectors in the process of obtaining training data, that is, the ultimate goal of the hidden layer in the model is to learn the operations between the weight matrix of the hidden layer to obtain the output of the hidden layer, that is, each input word "Embedded word vector";
5. The final output layer is a softmax regression classifier, which gives a probability value between 0 and 1. The sum of the probabilities of all nodes in the output layer is 1;
6. When the neural network training is completed, the word vectors of all words can be obtained.

The resulting word vector has very good linear properties:  
 $\text{vectors}(\text{man}) - \text{vectors}(\text{woman}) + \text{vectors}(\text{daughters}) = \text{vectors}(\text{son})$ .

### 3.2.4 Semantic expansion

Word2vec can convert text words into vectors in a vector space, and the vector cosine similarity can represent the similarity of text meaning [38]. The description text of the Web service is usually short, and the description text of the Web service needs to be expanded. This article uses word2vec to train the word vector model of Wikipedia English corpus, the training tool is gensim, and the parameters are shown in table 3.1.

Table 3.1- Word2vec parameter setting table

Parameter	Value	Meaning
Size	200	Vector dimension
Window	5	Window size
min_count	5	Minimum word frequency
iter	10	Minimum word frequency
sg	0	CBOW

After completing the training on Wikipedia, we obtain a corpus, and then use the trained word vector model to semantically expand the description of the Web service. The specific method [39] is: in the word vector space, find the similarity to the word before the pre-processed word in the original Web service description document as the expansion, and obtain Web service description documents with different expansion levels. As shown in the figure, a few words with the closest cosine similarity to "author" are listed. Through the word2vec model training, the N-dimensional word vector  $w$  corresponding to each participle is obtained, where  $w = (v_1, v_2, \dots, v_n)$ .

Figure 3.3 shows the similar words to the target word.

```
[('Writer', 0.8938986901270063),
 ('Publisher', 0.8562425308037052),
 ('Novel', 0.8542623058682257),
 ('literature', 0.8443898147931245),
 ('dramatist', 0.8272116808220137),
 ('Copyright', 0.7606132531479470)]
```

Figure 3.3- Words similar to the target word

**3.3 LDA theme model**

LDA (Latent Dirichlet Allocation) is a theme model. It uses a probabilistic production model to model hidden topics in text. It is a document generation model. Each document has multiple topics, and the topic model can mine potential topic information, that is, given parameters, the model gives the topics of each document in the document set in the form of probability, which realizes Modeling is an unsupervised machine learning technique. At the same time, the LDA topic distribution is a bag of words distribution, that is, the document is composed of a group of phrases, and the order of the words has no effect.

**3.3.1 Multinomial distribution and Dirichlet distribution**

The LDA theme model is applied to the multinomial distribution and the Dirichlet distribution. The Dirichlet distribution is actually a higher-order *Beta* distribution. The function expression of the Dirichlet distribution is like (3.8):

$$Dirichlet(p|\alpha) = \frac{\Gamma(\sum_{k=1}^k \alpha_k)}{\prod_{k=1}^k \Gamma(\alpha_k)} \prod_{k=1}^k p_k^{\alpha_k - 1} \tag{3.8}$$

Where  $\sum_k p_k = 1, \vec{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$  is the parameter of Dirichlet distribution. Dirichlet distribution is the distribution of its corresponding posterior polynomial distri-

bution parameter  $\vec{p}$ , Dirichlet distribution is the conjugate distribution of polynomial distribution.

The function expression of the polynomial distribution is like (3.9):

$$Mult(n|\nu, N) = \binom{N}{n} \prod_{k=1}^k \nu_k^{n_k} \quad (3.9)$$

Where  $\vec{\nu} = (\nu_1, \nu_2, \dots, \nu_n)$  represents the probability that each value is selected.

### 3.3.2 Model introduction

The LDA model assumes that the topic distribution  $\theta$  and the word distribution  $\phi$  on the topic are both polynomial distributions, and the hyperparameter  $\alpha$ ,  $\beta$  represents the conjugate distribution of the polynomial distribution  $Dir(\alpha)$  Dirichlet distribution  $Dir(\beta)$ , the prior parameter of  $\theta$ , distribution of words on  $\phi$ .

The LDA model is shown in figure 3.4.

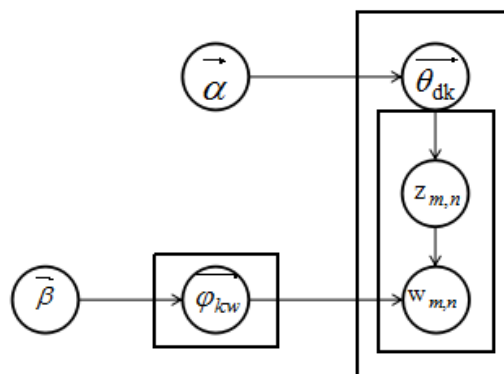


Figure 3.4- Illustration of LDA model

The process of generating text by the LDA theme model is as follows:

1. According to the Dirichlet prior distribution  $Dir(\alpha)$ , the distribution  $\vec{\theta}_{dk}$  of the document on the topic can be obtained.
2. For each document, according to the Dirichlet prior distribution  $Dir(\beta)$ , get the polynomial distribution  $\vec{\phi}_{kw}$  of each topic  $k$  on the word.
3. Obtain the theme  $Z_m, n$  through multiple distributions  $Mult(\theta_d)$ , and obtain the words through  $w_{m,n}$ .

From this generation process, the joint distribution of LDA after given hyperparameter  $\alpha$ ,  $\beta$  is obtained by formula (3.10):

$$P(w_d, z_d, \theta_d, \phi|\alpha, \beta) = P(\phi|\beta) \prod_{n=1}^{N_d} P(\theta_d|\alpha) P(w_{d,n}|\phi_{z_{d,n}}) P(z_{d,n}|\theta_d) \quad (3.10)$$

By eliminating the variable, the likelihood value of the text is calculated by (3.11):

$$P(w_d|\alpha, \beta) = \iint P(\theta_d|\alpha)P(\phi|\beta) \prod_{n=1}^{N_d} P(w_{d,n}|\theta_d, \phi_{z_{d,n}})P(z_{d,n}|\theta_d) d\phi d\theta_d \quad (3.11)$$

Then for the entire text set D, the likelihood value is calculated by (3.12):

$$P(D|\alpha, \beta) = \prod_{d=1}^{|D|} P(w_d|\alpha, \beta) \quad (3.12)$$

### 3.3.3 Parameter estimation

Gibbs sampling can generate data from a complex probability distribution. The posterior distribution calculation of latent variables is the core of LDA model parameter estimation. The key is to obtain the target probability distribution. According to the formula of Dirichlet parameter estimation, the document topic distribution  $\theta$  and topic word distribution  $\phi$  are obtained by formula (3.14) and (3.15):

$$\theta_{d_k} = \frac{n_d^{(k)} + \alpha_k}{\sum_{k=1}^k n_d^{(k)} + \alpha_k} \quad (3.14)$$

$$\phi_{k_w} = \frac{n_k^{(w)} + \beta_w}{\sum_{w=1}^v n_k^{(w)} + \beta_w} \quad (3.15)$$

Where  $n_k^{(w)}$  represents the number of words  $w$  assigned to the topic  $k$ ,  $n_d^{(k)}$  represents the number of words assigned to the subject  $d$  by the document  $k$ ,  $v$  represents the total number of words in the document,  $k$  represents the number of set topics,  $\beta_w$  represents the word  $w$  in the subject  $k$  Dirichlet prior parameters,  $\alpha_k$  represents the Dirichlet prior parameters of the subject  $d$  in the document  $k$ .

For better processing, make symmetric  $\phi_{z_d}$  and  $\theta_{d_k}$  prior probability assumptions for  $Dirichlet(\beta)$  and  $Dirichlet(\alpha)$ , and proceed from the posterior probability distribution  $w$  of the word  $P(w|Z)$  to the topic, and sample the distribution of the topic on the word from formula (3.16):

$$P(Z_n = k | Z_{-n}, w_n, x) \propto \frac{n_k^{(w)} + \beta_w}{\sum_{w=1}^v n_k^{(w)} + \beta_w} \cdot \frac{n_d^{(k)} + \alpha_k}{\sum_{k=1}^k n_d^{(k)} + \alpha_k} \quad (3.16)$$

Figure 3.5 shows the path probability.

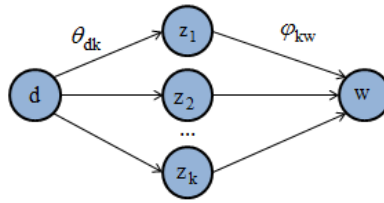


Figure 3.5- doc  $\rightarrow$  topic  $\rightarrow$  word Path probability

This probability is actually the path probability of doc  $\rightarrow$  topic  $\rightarrow$  word . The Gibbs sampling process is to gradually process each word in the document set. When the topic distribution of other words is known, the topic distribution  $P(Z_n = k | Z_{-n}, w_n, x)$  of the word is estimated, and according to this distribution, it is given to this Reselect the theme of the word, and count the theme distribution of each word in the document and the word distribution of each theme.

As we can see from the previous article, we get a document collection  $IDS = \{itd_1, itd_2, \dots, itd_n\}$ , through a series of preprocessing, and each document  $itd_i$  is characterized by a word vector  $w = (w_1, w_2, \dots, w_n)$  to form a document-word matrix, where n is the total number of documents and m is the total number of words, then the document- The word matrix is (3.17):

$$IDS = \begin{bmatrix} itd_1 \\ itd_2 \\ \dots \\ itd_n \end{bmatrix} = \begin{bmatrix} w_{11}, w_{12}, \dots, w_{1m} \\ w_{21}, w_{22}, \dots, w_{2m} \\ \dots \\ w_{n1}, w_{n2}, \dots, w_{nm} \end{bmatrix} \quad (3.17)$$

$w_{ij}$  represents the  $j$ th word under the  $i$ th document.

The LDA model is used to model the matrix IDS, and Gibbs iterative sampling is used to estimate the probability of  $w_{ij}$  generating topics, and the document-topic probability distribution matrix is finally obtained when converging like (3.18):

$$IDS = \begin{bmatrix} r_1 \\ r_2 \\ \dots \\ r_n \end{bmatrix} = \begin{matrix} \text{topic}_1, \text{topic}_2, \dots, \text{topic}_k \\ \begin{bmatrix} p_{11}, p_{12}, \dots, p_{1k} \\ p_{21}, p_{22}, \dots, p_{2k} \\ \dots \\ p_{n1}, p_{n2}, \dots, p_{nk} \end{bmatrix} \end{matrix} \quad (3.18)$$

Where  $k$  is the number of topics, and  $p_{ij}$  is the distribution probability of the  $i$ -th document on the  $j$ -th topic. For example, we set the number of topics  $k$  to 20, the number of iterations to 1800, and print the topic probability distribution of the first sample like figure 3.6.

The first sample:

```
[3.62505347e-04 3.82173513e-02 4.94784609e-04 8.13143578e-03  
1.52206232e-03 5.14782509e-04 8.23053476e-03 3.07695682e-02  
4.04792609e-01 4.41782609e-04 4.41782609e-04 4.36782609e-04  
1.52206232e-03 2.18826087e-01 4.41782609e-04 4.41782609e-04  
4.14582639e-04 3.90632174e-02 4.41782609e-04 1.17916463e-01]
```

Figure 3.6-Theme probability distribution

Through this step of operation, we get the document-topic probability distribution matrix described by the service text, ready for the next cluster analysis.

### 3.4 Summary of this chapter

This chapter mainly implements a pre-processing of documents before clustering to obtain information such as service name and service description. Then use Word2Vec to train the vector model to expand the semantics of the service description, and through TF-IDF weighting processing, followed by the extended description document using the topic model for feature modeling, and the short text topic modeling into long text topic To improve the accuracy of the theme expression of service content. This is the basis for clustering services using the clustering algorithm in the next step, preparing for the next chapter for clustering services.

## 4 WEB SERVICE DISCOVERY METHOD BASED ON CLUSTERING

The traditional Web service discovery method only searches each service in turn in the service registration center, and needs to match each service, which is inefficient. In order to improve the performance of Web service search, the massive services of the service registration center can be optimized. Service clustering is an effective method. By clustering services with similar functions or other attributes, you can quickly locate the service cluster that meets the user's needs. This limits the scope of the search to a specific service cluster and is performing services. During matching and discovery, reducing the service search space reduces the number of service searches, avoids matching calculations with unrelated services, and reduces matching and search time. In this way, Web services of similar structure and similar functions can be better managed and identified from the massive services, and the efficiency of service discovery is greatly improved.

### 4.1 Clustering algorithm analysis

Cluster analysis originated from taxonomy and is a very important field in unsupervised learning. Statistical data analysis technology is widely used in many fields. The so-called unsupervised learning is to explore certain rules from the given data and discover the potential internal patterns and structures. The data is not labeled with categories, which is also called unsupervised learning. Cluster analysis is to divide the samples from the data set into different classes and discover the hidden knowledge in each subset, and each subset is called a "cluster". By describing information and characteristics, make the data points in the group similar. Given a set of data points, we can use a clustering algorithm to divide data points with similar attributes or features into a cluster, the data in the cluster has a high similarity, and the data points in different clusters have a high dissimilarity, should have highly different attributes or characteristics.

#### 4.1.1 Evaluation of clustering effectiveness

Cluster validity evaluation is an objective indicator to measure the validity of classification results. For the same data set, the clustering results obtained by different clustering algorithms may also be different. In order to evaluate the quality of the clustering results, that is, the accuracy and effectiveness, it is necessary to select appropriate validity indicators to evaluate the clustering effect. The effectiveness of clustering results and the number of optimal clusters obtained are the main criteria for evaluating the effectiveness of clustering. The smaller the distance between the objects in the cluster, that is, the cluster distribution satisfies compact independence, the greater the distance between the clusters, it means that this is a good clustering.

At present, the commonly used clustering effectiveness evaluation is mainly aimed at two aspects, one is the comparison of clustering results of the same clustering algorithm under different parameters, and the other is the comparison of different clustering

algorithms. Specific evaluation methods include external evaluation method and internal evaluation method. Among them, the external evaluation method is a supervised method. An external model is established based on the actual clustering distribution, and the matching degree of the clustering result and the real distribution is compared. Using the external model as a reference benchmark, a certain measure is needed to judge the degree of conformity between the clustering result and the benchmark data, and the benchmark data is needed. Commonly used external evaluation methods include Jaccard coefficient, FM index, Rand index parameter and B3 (bcubed index). The internal indicators are based on unsupervised methods and do not use the original information of the original data distribution. The clustering results are measured based on the closeness of the objects in the cluster and the separation between clusters. Method without benchmark data. Commonly used internal indicators include Cophenetic correlation coefficient, DB index, contour coefficient, and BWP indicator (between within proportion).

#### 4.1.2 Classification of clustering algorithms

Web service clustering provides important support and guarantee for Web service discovery. They can be divided into the following categories: division method, hierarchical method, density algorithm, grid-based method and model-based method, these methods are suitable for different fields according to their unique characteristics.

The comparative analysis of the five clustering algorithms is shown in table 4.1 below.

Table 4.1- Comparison of clustering algorithms

Clustering algorithm	Advantages	disadvantages
The K-means algorithm [41]	Fast and easy to implement.	Sensitive to the initial clustering center and must be specified in advance
Hierarchical clustering algorithm[42]	Does not need to determine the value	The calculation amount is particularly large and the processing speed is slow
DBSCAN[43]	Relatively anti-noise, it can be found that samples of any shape are computationally complex	High-dimensional data is not easy to define density
GMM[44]	Understandable and fast	Initialization sensitivity needs to specify K value
Spectral clustering[45]	Can be found that non-spherical samples	Need to specify the K value



Although each algorithm uses its own unique method to define the distance metric, they all require  $O(n^2)$  or more calculation time to calculate their distance metric. The K-means algorithm is a classic algorithm in data mining. Due to the low calculation time of  $O(n_k)$ , it is often used for clustering high-dimensional large document vectors. This paper proposes a new clustering method based on K-means clustering and hierarchical clustering, based on the advantages and disadvantages of the two clustering methods. However, because K-means clustering needs to set the number of clusters by itself, it cannot automatically find the optimal number of clustering categories. Too many or too few settings may affect the clustering effect and may result in unstable quality of results. Hierarchical clustering can process categorical data and quantitative data, but the processing speed is relatively slow. Usually, it is necessary to combine the relevant results to subjectively judge the number of clustering categories.

### 4.1.3 K-means clustering

K-means clustering algorithm is an iterative solution cluster analysis algorithm. Its central idea is to divide data objects into different clusters by iteration, so as to minimize the required objective function, so that the generated clusters are as compact as possible. And independence. By minimizing the distance between vectors within each cluster (as shown in Equation 4.1), the k-means algorithm represents each cluster by its centroid vector.

Input: the number of classification clusters  $K$  and the data set containing  $n$  data objects

Output:  $K$  clusters completed by clustering

Steps:

1. First, randomly select  $k$  samples as the center  $c = c_1, c_2, c_3 \dots c_k$  of the initial clustering;
2. For each sample  $x_i$  in the data set, the distribution calculates the distance to the  $k$  cluster centers. Then divide it into the clusters with the smallest cluster center, according to the Euclidean distance formula (4.1):

$$d(x_i, c_j) = \left( \sum (x_i - c_j)^2 \right)^{\frac{1}{2}} \quad (4.1)$$

3. For each category  $c_j$ , recalculate its clustering center (4.2):

$$c_j = \frac{1}{|a_i|} \sum_{x \in a_i} x ; \quad (4.2)$$

4. Repeat steps 2 and 3 above until a certain termination condition is reached.

The termination condition can be set to:

1. The cluster center no longer changes
2. All data have been allocated

Fake code:

Get data n m-dimensional data

    Create K points as starting centroids

    while (t)

        for (int i = 0; i <n; i ++)

            for (int j = 0; j <k; j ++)

                Calculate the distance from centroid j to data point i

    for (int i = 0; i <k; i ++)

        1. Assign the data point to the nearest cluster

        2. For each cluster, calculate the mean of all points in the cluster as the centroid

    End

#### **4.1.4 Hierarchical clustering**

Hierarchical clustering, also called systematic clustering, the basic idea is to use multiple samples as a class, calculate the distance between two samples, merge the two types with the closest distance into a new type, then calculate the distance, and then merge Until there is only one category.

    Input: sample set, number of clusters, or a termination condition

    Output: clustering results

    Steps:

    1. We treat each sample in the sample set as a separate cluster and calculate the similarity between each cluster class to obtain a similarity matrix. The distance for the similarity between clusters used here is formula (4.3):

$$\text{dist}(C1, C2) = \max_{p_i \in C1, p_j \in C2} \text{dist}(p_i, p_j) \quad (4.3)$$

    Where C1 and C2 represent two cluster classes, and  $p_i, p_j$  is any two samples;

    2. When the similarity between the two clusters is the highest, merge into a cluster, continue to calculate the similarity between each cluster, and update the similarity matrix;

    3. Repeat the above two steps 2 and 3 until a certain termination condition is reached to obtain the preset K clusters.

    The termination condition can be set to:

        1. When the function converges

        2. The maximum number of iterations max or the termination threshold is reached

$\mu$

    Fake code:

    Get n m-dimensional data as n clusters, and finally divide into k clusters

    while (n > k)

    Calculate the distance between two clusters

    Find the two clusters with the smallest distance

N-= 1  
end

K-means is our most commonly used clustering algorithm based on Euclidean distance. It believes that the closer the distance between two targets, the greater the similarity. Its biggest feature is that it can quickly process a large amount of data. However, because K-means clustering needs to set the number of clusters by itself, it cannot automatically find the optimal number of clustering categories. Too many or too few settings may affect the clustering effect and may result in unstable quality of results. Hierarchical clustering can process categorical data and quantitative data, but the processing speed is relatively slow. Usually, it is necessary to combine the relevant results to subjectively judge the number of clustering categories. We made improvements on this basis, and proposed a KMHC Clustering clustering algorithm for the advantages and disadvantages of the two clustering algorithms.

## 4.2KMHC clustering

Chen et al. [46] considered the advantages and disadvantages of K-means clustering and hierarchical clustering in dealing with cluster analysis problems, and proposed a traditional HK clustering algorithm, which combines them organically according to the characteristics of the two. -means algorithm needs to realize the determination of the number of clusters, which is prone to errors due to artificial regulations, and hierarchical clustering does not need to realize the determination of K value. However, when the number of clusters K is uncertain in advance, it is necessary to repeatedly execute the entire algorithm flow to obtain the optimal clustering results from different clusters. To this end, we introduced the contour coefficient, so the basic idea of our algorithm is: first set different algorithm termination conditions, use the hierarchical clustering algorithm to obtain the number of clusters and cluster center, calculate the contour coefficient, and then use the K-means algorithm.

### 4.2.1 Contour coefficient

The contour coefficient [47] (Silhouette Coefficient) can quantify the similarity between the objects in the data set and other objects in the cluster and the objects in other clusters, and somehow combine the two quantized similarities Sex, get the pros and cons of clustering evaluation. The effectiveness of the clustering effect can be evaluated using the profile coefficient  $s(k)$  value through formula (4.4):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (4.4)$$

Where  $a(i)$  represents the average distance from sample  $x(i)$  to other samples in the same cluster  $C_i$ , and  $b(i)$  represents the evaluation distance from sample  $x(i)$  to all samples in other clusters  $C_j$ . And calculated  $s(i)$  by (4.5):

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & a(i) < b(i) \\ 0, & a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & a(i) > b(i) \end{cases} \quad (4.5)$$

The closer the  $s(i)$  is to 1, the more reasonable the sample clustering is, the closer to -1 is, the sample should be divided into other clusters, and the closer to 0, the sample should be on the boundary of the cluster. The contour coefficient after clustering can be expressed as formula (4.6):

$$s(k) = \frac{1}{m} \sum_{i=1}^m s(i) \quad (4.6)$$

Where  $k$  is the number of clusters and  $m$  is the number of all data objects in the data set.

#### 4.2.2 KMHC Clustering

KMHC Clustering clustering algorithm first adopts split-level hierarchical clustering algorithm to select the two nearest clusters for merging, meanwhile calculate the average of two cluster centroids as the center of the new cluster after merging, repeat the above steps, and calculate the cluster Contour factor. Set the termination conditions of multiple hierarchical clustering algorithms of different degrees. Once the termination condition of a preset hierarchical clustering algorithm is reached, select the  $K$  value and clustering center obtained by the hierarchical clustering algorithm with the largest contour coefficient as the next step. The  $K$  value and initial center position of the  $K$ -means algorithm will be carried out soon. The adopted  $K$ -means algorithm can obtain different clustering results. Finally, the clustering effectiveness is used to evaluate each clustering result, and the cluster with the highest quality is selected as the final result clustering.

When using the Euclidean distance for calculation, because of the large influence of dimension, the large order will affect the manifestation of small order, and the Mahalanobis distance [48] has nothing to do with dimension, and can eliminate the correlation interference between variables.

The definition of Mahalanobis distance is given by formula (4.7):

$$D_M(x, u) = \sqrt{(x-u)^T \Sigma^{-1} (x-u)} \quad (4.7)$$

Where  $x = (x_1, x_2, \dots, x_m)^T$ ,  $u = (u_1, u_2, \dots, u_n)^T$ , T means transpose,  $\sum$  means sample covariance matrix.

The process of KMHC clustering algorithm is as follows:

Input: IDS theme matrix

Output: optimal K clusters

Steps:

1. We treat each document in the sample set IDS as an independent cluster, calculate the similarity between each cluster class, and obtain the similarity matrix.
2. For the clusters obtained in step (1), calculate the centroid of each cluster separately. When the distance between the two clusters is the smallest, merge them into a cluster and update the similarity matrix.
3. Repeat the above two steps to obtain the silhouette when the algorithm terminates under different termination conditions, and use the K value corresponding to the largest silhouette value as the input of the K-means algorithm.
4. Use the center of the obtained clusters as the clustering center of the K-means algorithm (here, the Mahalanobis distance is used) to perform clustering until the entire cluster object no longer changes, and obtain K clusters.

After clustering the web services, we get a cluster cluster, each cluster has its own cluster center, when performing web service discovery, first select the cluster where the web service is located according to the cluster center, and then Identify specific services that meet user needs in the cluster, so that the scope of service queries can be greatly reduced and the efficiency of service discovery can be improved. For example, we get a travel service cluster, including {AccommodationInfoService, City2CityRouteFinderService, CityLuxuryHotelService, CountryLightning Service, CountryWeatherProcessService ...}, after obtaining this service cluster, we next perform function matching and calculate the requested service based on the semantic similarity of ontology The similarity to each service in the service cluster returns the services that meet the requirements to the user, and obtains services that meet the user's needs.

### 4.3 Semantic Web service discovery

In the previous section, we obtained candidate service sets through KMHC Clustering. In this section, we will find services that meet user needs from the obtained candidate service clusters and give the corresponding service discovery framework.

#### 4.3.1 Service matching

The core of the Semantic Web Service is to obtain the best matching service from the massive service based on the information requested by the service, that is, service matching. Function-based semantic Web service similarity matching includes input semantics and output semantics similarity matching. In this paper, the semantic similarity method based on the ontology concept is used to perform similar calculations on the IO para

meters of the Web service, that is, the input and output semantics. First use Xpath technology to map the service information described by OWL-S into the ontology tree, and use the idea of ontology to calculate the similarity of functional attributes. The semantic distance is expressed by calculating the geometric distance between two concepts. The concept calculates the path sum to the nearest common parent node. The calculation model based on the semantic similarity of ontology concepts is like formula (4.8) and (4.9):

$$Sim(c_1, c_2) = \frac{1}{Dis(c_1, c_2) + 1} \quad (4.8)$$

$$Dis(c_1, c_2) = mp(c_1, r(c_1, c_2)) + mp(c_2, r(c_1, c_2)) \quad (4.9)$$

Among them,  $mp(c_1, r(c_1, c_2))$  represents the shortest path from the concept  $c_1$  to the nearest common parent node of  $c_1$  and  $c_2$ , and  $mp(c_2, r(c_1, c_2))$  represents the shortest path from the concept  $c_2$  to the nearest common parent node of  $c_1$  and  $c_2$ . Reference [31] considers that the lower the level of concepts in the ontology tree, the smaller the similarity between concepts. In the specific calculation, the depth of the closest common parent node of the two concepts in the ontology tree should be considered. In addition to the shortest path between the concept pairs, the maximum depth of the ontology tree where it is located is also considered. The similarity calculation formula based on the semantic distance of ontology concept is given by formula (4.10):

$$Sim(c_i, c_j) = \frac{dp(RCPN(c_i, c_j), c_i) + dp(RCPN(c_i, c_j), c_j)}{(Dis(c_i, c_j) + 1) \times (\max(dp(c_i), dp(c_j)))} \quad (4.10)$$

Among them,  $dp(RCPN(c_i, c_j), c_i)$  represents the depth of the closest common parent node of the concept to  $c_1$  and  $c_2$  in the ontology tree where the concept  $c_1$  is located, and  $\max(dp(c_i), dp(c_j))$  represents the maximum depth in the ontology tree where the concept  $c_1$  and  $c_2$  is located. Figure 4.1 is an ontology concept tree.

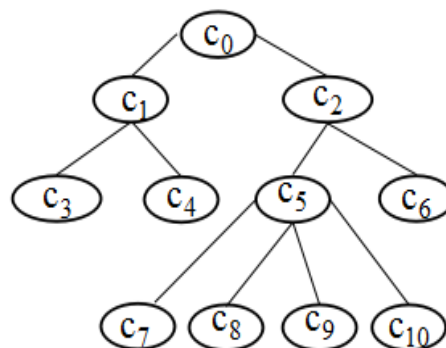


Figure 4.1- Ontology concept tree

The data is shown in table 4.2.

Table 4.2 -Concept node similarity result set

Concept node	Semantic distance similarity	Concept node	Semantic distance similarity
$c_0 / c_0$	1	$c_3 / c_4$	0.44
$c_0 / c_1$	0.67	$c_2 / c_5$	0.67
$c_0 / c_2$	0.67	$c_2 / c_6$	0.67
$c_1 / c_2$	0.33	$c_5 / c_6$	0.44
$c_1 / c_3$	0.67	$c_5 / c_7$	0.75
$c_1 / c_4$	0.67	$c_7 / c_8$	0.75

From the data in the above table, it can be analyzed that as the ontology level increases, the similarity between concepts becomes higher and higher. In practical applications, as the classification becomes finer, the deeper the hierarchy in the ontology tree, the sub-concept node The similarity is higher than that of the upper layer. However, we also found that the similarity of the sibling nodes is the same, so that in the final service matching, multiple services that meet the service requirements may be returned, which requires further work to optimize, such as setting the weight of the node. How to set appropriate weights is also a problem we need to study in the future. Here we can calculate the available formula of IO similarity of two services through formula (4.11):

$$Sim(s_1, s_2) = Sim_{in}(s_1, s_2) + Sim_{out}(s_1, s_2) \quad (4.11)$$

In the formula,  $Sim_{in}$ ,  $Sim_{out}$  represents the input concept similarity and output concept similarity of the two services, and we return the services that meet the requirements from the obtained clusters through the semantic similarity method based on the ontology concept.

### 4.3.2 Service discovery process

Service providers register services in the service registration center. In the face of massive services, the service registration center performs clustering preprocessing based on the service text description information. After clustering, the service registration center forms different service clusters, and then calculates the conceptual semantic similarity of the input and output information to further obtain services that meet the user's needs from the clustering, and finally the service caller is on the platform according to their own needs. Make a query to find the set of Web services that you need. The steps of service discovery are as follows.

1. Establish a recommendation service library for all types of users;
2. Preprocessing of service description;

3. Use clustering algorithms to cluster services and establish a user cluster library;
4. On the basis of satisfying the first layer of matching, perform input and output matching with functional attributes. Sort services and return results to users;
5. Update the user cluster library and recommended service library according to the update mechanism.

Figure 4.2 shows the service discovery framework based on KMHC clustering and concept similarity.

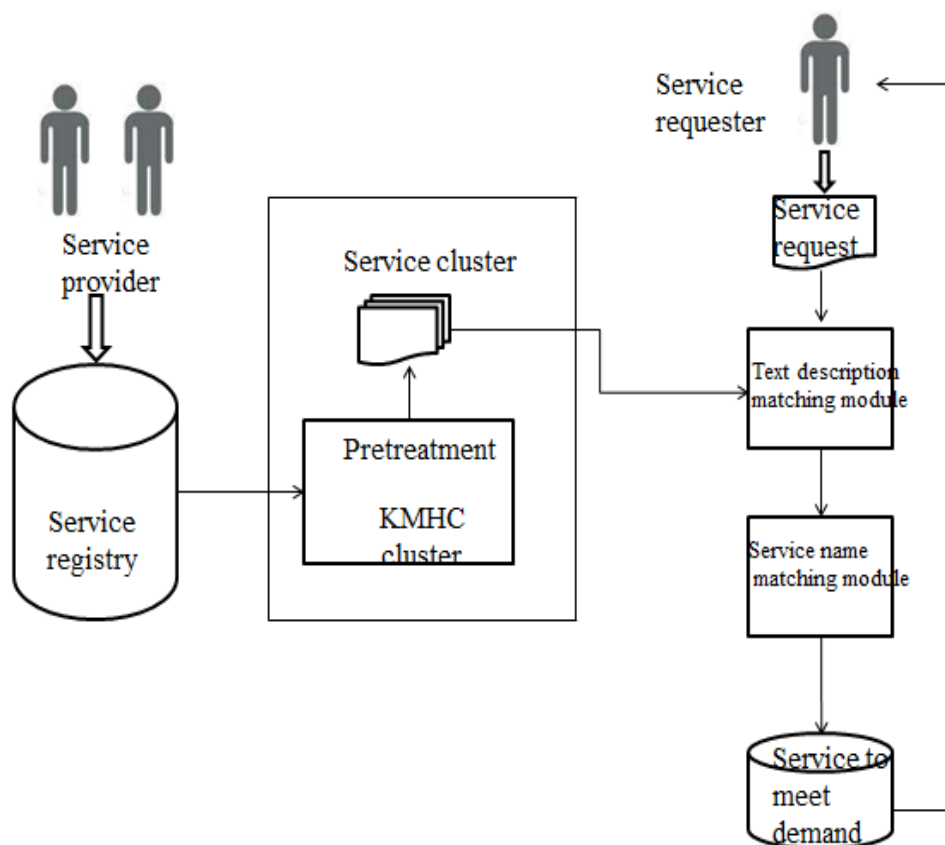


Figure 4.2 -Service discovery framework

## 4.4 Experimental analysis

### 4.4.1 Experimental settings

The experimental data set used in this paper is OWLS-TC4 [49], which is the fourth version of the OWL-S service retrieval test set. This set supports the evaluation of the performance of the OWL-S service pairing algorithm. It provides 9 different areas (education, medical, food, travel, communications, economics, weapons, geography and simulation), covering all aspects of life. It includes 1083 semantic Web services, and also provides 42 related sets of test queries for performance evaluation experiments. The programming language uses Python and the compilation environment JetBrains



PyCharm 2017. The hardware environment and operating system used in this experiment are shown in table 4.3.

Table 4.3 – Experimental hardware parameters

parameter	value
processor	Intel i5
RAM	32G
operating system	Win7

#### 4.4.2 Results analysis

Experiment 1: Comparison of clustering efficiency of Web services under different topics.

In this section, an experiment is designed to determine the number  $K$  of LDA topics. In the experiment, the hyperparameter  $\alpha$  in the LDA model is set according to the number of topics  $K$ . Let  $\alpha = 50/K$ ,  $\beta = 0.01$ , calculate the required parameters through Gibbs, and set the maximum number of iterations to 1000. The document subject vectors generated by LDA are clustered using KMHC Clustering algorithm. In order to evaluate the experimental results more intuitively and comprehensively, this paper uses Accuracy (4.12), Recall(4.13), and F (4.14) values to represent the clustering results.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (4.12)$$

$$Recall = \frac{TP}{TP+FN} \quad (4.13)$$

$$F-measure = \frac{2A * R}{A + R} \quad (4.14)$$

The meaning of the relevant values is shown in table 4.4.

Table 4.4- Meanings of related numerical values of evaluation criteria

	Predict positive	Predict negative	Total
Forecast positive	TP	TP	TP+FN
Forecast negative	FP	FP	FP+TN
Total	TP+FP	TP+FP	TP+FP+TN+FN

Accuracy represents the percentage of correctly tested samples in the total test samples; Recall represents the percentage of correctly identified positive samples in the total positive samples; F-measure is a combination of Accuracy and Recall. The clustering results are shown in the figure 4.3.

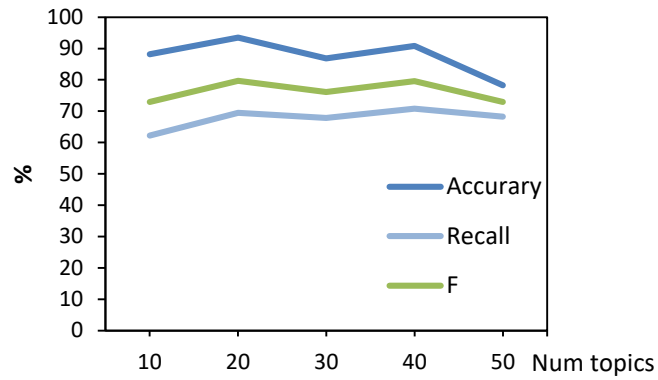


Figure 4.3- Clustering effect under different themes

Statistical language models usually use the degree of confusion to judge the performance of a model. The smaller the degree of confusion, the better the model's ability to predict data [50]. The calculation formula of perplexity is as follows (4.15):

$$perplexity(D) = \exp\left(-\frac{\sum_{d=1}^M \log p(w)}{\sum_{d=1}^M N_d}\right) \quad (4.15)$$

Where  $N_d$  represents the number of words in document  $d$ ,  $M$  represents the number of documents in the data set, and  $p(w)$  represents the probability of each word in the data set. The puzzle-theme curve of the number of different topics in the LDA model is shown in the figure 4.4.

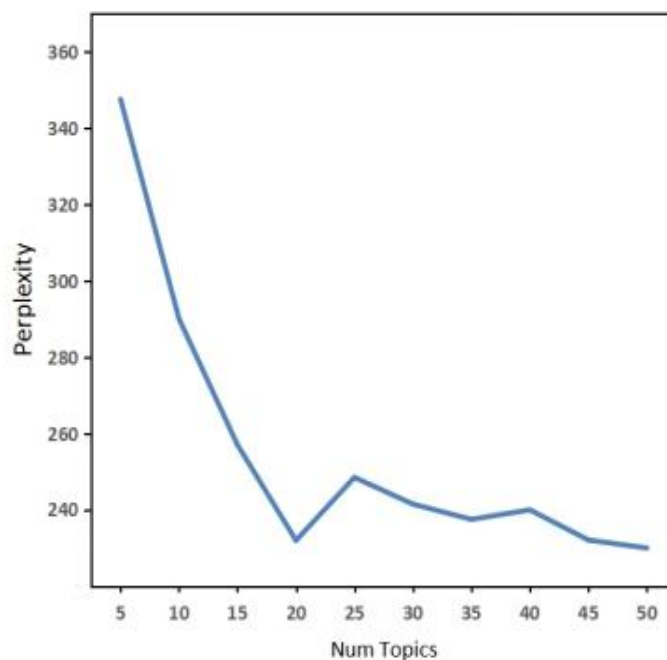


Figure 4.4-LDA model perplexity

It can be seen from the figure that when the number of topics is 20, the F-measure reaches the maximum. As the number of topics increases, the F-measure fluctuates and reaches a peak when the number of topics is 40. If the number is too large, the interpretability will be reduced, and the confusion index will show a clear inflection point when the number of topics is 20. Therefore, this paper believes that when the number of topics is set to 20, the topic modeling effect is the best.

Experiment 2: verify the effectiveness of the service discovery algorithm proposed in this paper

In service matching, precision and recall are usually used to measure the quality of a Web service matching algorithm. Precision is used to measure the search accuracy rate of the system algorithm; Recall is used to measure the range of the system algorithm search. The higher the precision and recall, the better the service matching algorithm. Relevant represents the total number of Web services returned by the query. Retrieved represents the number of Web services that meet the query conditions in the test sample set. The formula is as follows (4.16) and (4.17):

$$\text{Precision} = \frac{\text{relevant} \cap \text{retrieved}}{\text{relevant}} \tag{4.16}$$

$$\text{Recall} = \frac{\text{relevant} \cap \text{retrieved}}{\text{retrieved}} \tag{4.17}$$

By observing the F value corresponding to the number of different topics, we can see that when  $K = 20$ , the training effect of LDA is the best. Choose the method in this paper and compare it with the service discovery method proposed in [17].

Through experimental comparison, we can see that the discovery method proposed in this paper has improved the recall rate and precision rate compared with the method proposed in [17]. This is because the text description part of the Web service is generally short. After a series of operations such as removing stop words, the TF-IDF value of the feature word is calculated, and the vocabulary with low frequency is eliminated. Since there are not many remaining feature words in the text, For this problem, we use word2vec to train Wikipedia, semantically expand the feature vectors described in the text, and use the LDA topic model for modeling to achieve the topic expression of Web service content. On this basis, service clustering greatly reduces the scope of the query, makes up for the short description text, lack of co-occurrence of word frequency and sparse semantics, and improves efficiency. On this basis, the KMHC clustering algorithm is used for service clustering, and the service matching is performed through semantic similarity based on the ontology concept. Compared with the algorithm proposed in [17], the Web service discovery method proposed in this paper has a lot of advantages in accuracy. Figure 4.5 shows the comparison of the effects of the two methods.

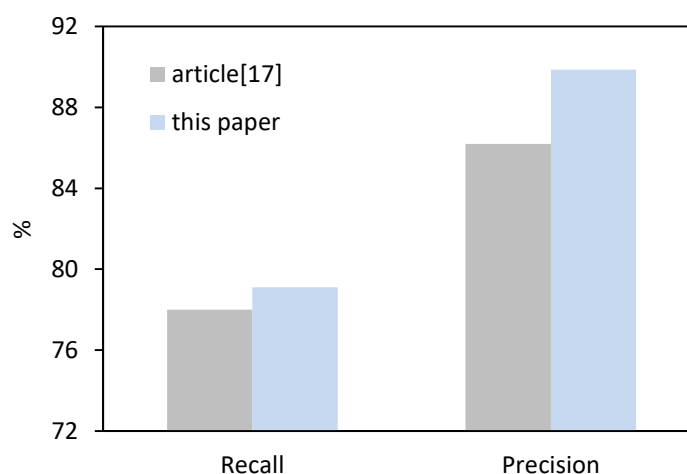


Figure 4.5 -Comparison of the effects of the two methods

According to the method of this paper, after performing clustering of service domains, function matching is based on conceptual similarity. Compared with sequential matching, it can be seen that the difference is not large when the number of services is relatively small, but as the number of services increases, the function matching takes longer than the sequential matching. The increase is slower, and in the case of the same number of services, the execution time of this method is shorter. Experimental results show that the method proposed in this paper improves the matching efficiency. The results shows in the table 4.5.

Table 4.5 – Service matching efficiency

Number of services	Service matching time/s	
	Sequential matching	This article
50	11	11
200	54	52
500	157	150
700	298	285

#### 4.5 Summary of this chapter

This chapter implements service discovery. First, similar services are aggregated into a service cluster, and then the services required by the service requester are obtained through calculation based on similarity. This chapter first introduces the advantages and disadvantages of K-means and hierarchical clustering algorithm in detail. On this basis, an improved clustering algorithm KMHC Clustering clustering algorithm is proposed. Concept similarity, get the service required by the service requester, and give a service discovery framework. Through experiments, we can prove that the method proposed in this paper has certain advantages.

## **5 USE OF TECHNOLOGY OF THE SEMANTIC WEB SERVICES IN WORK OF THERMAL POWER PLANT**

### **5.1 The analysis strong and weaknesses of technology based on the semantic Web services, opportunities and threats of its application use of technology in work of thermal power plant**

With the rapid development of information technology, thermal power units urgently need to solve the problems that various types of data cannot be shared and applied, information islands among multiple systems, and on-site operation safety [51]. Utilizing technologies such as the Internet of things, artificial intelligence, and big data, and adopting multiple forms of communication, combined with geographic information technology, a highly intelligent production emergency command center is built. To realize the integration of the information resources of the whole plant, it can centrally display and comprehensively apply on-site video, voice, security, and production process data. Through the work of command and dispatch, integrated communications, plan management and daily external information analysis, it provides managers with a basis for auxiliary decision-making, and at the same time, remote command and decision-making for major safety production activities.

Thermal power, as an important source of electricity in China, is related to national energy security [52]. From the perspective of sustainable development, the traditional extensive development model of thermal power enterprises urgently needs to transform to a green development model. In the context of the deep integration of informatization and industrialization, in response to the development of new technologies in the Internet, big data, cloud computing and other information fields, the power industry is facing intelligent transformation and upgrading, and the construction of smart power plants will become an inevitable trend.

However, at present, there is still a problem that it cannot be shared and applied in the production and management of thermal power units, and major safety events and maintenance operation events cannot be visually displayed and cannot provide decision guidance [53]. Therefore, it is especially important to build an intelligent production emergency command center that can centrally display the information of the entire plant, manage and analyze the data of the entire plant, and provide emergency plans and command decisions in a timely manner. Due to the popularity and scale of web applications, the types of services have become more and more abundant. As a standard for remote access, Web services are also changing user needs. Web service discovery refers to a method for quickly and efficiently selecting the correct matching method from multiple services according to user needs (such as functional requirements) to find the service they need. According to the previous analysis, a unified Web service registry that can register the entire plant information can manage all the information.

Informatization can play a major auxiliary role in the process of handling emergency incidents. Various departments can establish disposal plans for various emergencies in the intelligent plan system and form a plan database. The system is associated with pro-

duction and fire protection, and has a built-in expert database and various information resources, which can be quickly linked to the emergency response department. When an emergency occurs, the intelligent plan system can retrieve a list of eligible plans based on the nature of the event, the scale of casualties, and the severity. These plans can provide a reference for the command staff when dealing with emergencies, in order to speed up the handling of emergencies and improve the effectiveness of handling emergencies. Therefore, the requirements for information technology will be relatively high. The current informatization construction is not yet perfect, and it needs a long way to complete.

For the application of semantic-based Web services in thermal power, we use SWOT diagram to analyze the actual situation of the application.

SWOT analysis is an analysis method proposed by Kenneth R. Andrews, used to examine how an enterprise obtains (sustainable) competitive advantage in the market. SWOT comes from the abbreviation of 4 English word letters, which are strength, weakness, opportunity and threat. The S, W analysis mainly focuses on the comparison of the company's own strength and competitors, while the O, T analysis mainly focuses on the changing external SWOT analysis. It is pointed out that companies should develop their internal advantages through the implementation of strategies to take advantage of the opportunities brought by the external environment Avoid the threats brought by the external environment and avoid the disadvantages within the enterprise, so as to obtain the advantages of continuous competition.

The SWOT analysis method is used to analyze the internal advantages and disadvantages of thermal power development and external opportunities and challenges, thus forming a SWOT analysis matrix, synthesizing the four major factors, and formulating SO, ST, WO, WT strategies for semantic-based Web services Provide a reference for how to better develop in China.

SO: Usually, this development strategy is chosen when the organization faces external opportunities and has internal and internal advantages. It encourages to fully seize external opportunities to seek new development paths, which is in line with the advantages of urban economic and social development.

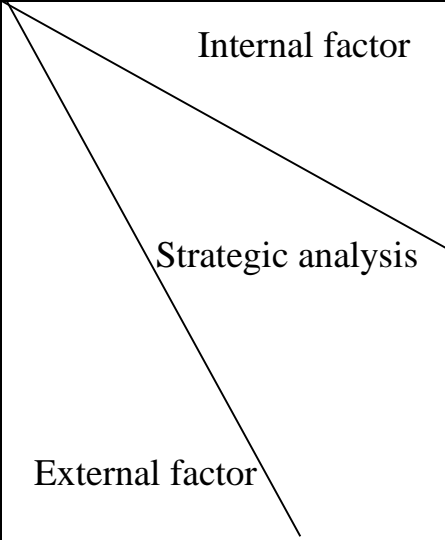
WO: Usually, this development strategy is chosen when the organization faces external opportunities and lacks obvious internal and internal advantages. It encourages to fully seize external opportunities to effectively improve its own deficiencies, standardize internal management and improve quality, and establish a good brand image.

ST: This development strategy aims to use the internal advantages of the organization to resist the external threats of the royal family, encourage the full play of the internal advantages of the organization, and seek new development opportunities to avoid or mitigate possible external threats.

WT: In the case of serious external threats and serious external threats, in order to overcome or improve their own deficiencies, face and resolve external threats, usually apply WT strategy.

We draw SWOT matrix analysis of application of semantic Web service in thermal power plant to show clearly as table 5.1.

Table 5.1–SWOT matrix analysis of application of semantic Web service in thermal power plant

<p style="text-align: center;">Internal factor</p>  <p style="text-align: center;">Strategic analysis</p> <p style="text-align: center;">External factor</p>	Strength	Weakness
	<ol style="list-style-type: none"> <li>1. Sharing apps</li> <li>2. Achieve the integration of information resources of the whole plant</li> <li>3. Conduct remote command and decision</li> <li>4. Provide a basis for managers to assist in decision-making</li> </ol>	<ol style="list-style-type: none"> <li>1. Relatively high cost</li> <li>2. The management model is not perfect</li> <li>3. Disconnection between power generation and operation management</li> </ol>
Opportunities	SO	WO
<ol style="list-style-type: none"> <li>1. In the context of "Made in China 2025", it will be more helpful to promote the wisdom of thermal power enterprises and the construction of smart power plants.</li> </ol>	<ol style="list-style-type: none"> <li>1. Actively respond to the current power system reform and market situation, adapt to the new economic normal, and accurately judge the power reform situation</li> </ol>	<ol style="list-style-type: none"> <li>1. Strengthen communication and cooperation with the government</li> <li>2. Strengthen the joint construction of local enterprises and commit to coordinated development</li> </ol>
Threats	ST	WT
<ol style="list-style-type: none"> <li>1. The current rise in coal prices has led to an increase in the cost of power generation, and energy conservation and consumption reduction in thermal power plants is very important</li> </ol>	<ol style="list-style-type: none"> <li>1. Coordinate the relationship between the plan and the market</li> <li>2. Strengthen mutual trust communication between the same industry</li> <li>3. Reserve talents who meet the requirements of the new reform situation</li> </ol>	<ol style="list-style-type: none"> <li>1. Optimize fuel management</li> <li>2. Building an intelligent production emergency command center</li> <li>3. Combining big data technology with the traditional power generation industry to improve the production management model of the power generation industry</li> </ol>

From the SWOT form we can draw that we must seek new opportunities in reform, maintain technological innovation, and adapt to the new economic normal. Focus on scientific and technological innovation.

## 5.2 Gantt's schedule of actions for implementation of technology based on the semantic Web services in work of thermal power plant

Gantt chart is a management tool invented by American management scientist Henry Laurence Gantt to indicate the progress of project work with lines, which is mostly used in project management [54]. In this section, we use the Gantt chart to show the work of our thesis as follows in the table 5.2.

Table 5.2–Gantt's schedule

Research and Project stages	Performers	Period of implementation of the project 2019 - 2020, month									
		9	10	11	12	1	2	3	4	5	
1	2	3									
1. Development of introduction.	Y.Sang										
2. Search on semantic Web	Y.Sang										
3. Project to determine	Y.Sang										
4. Model design	Y.Sang										
5. The analysis model	Y.Sang										
6. Calculate data	Y.Sang										
7. Analyze data	Y.Sang										
8. Collect SWOT materials	Y.Sang Professor A. Alabugin. Senior lecturer R. Alabugina.										
9. SWOT analysis	Y.Sang, Professor A. Alabugin. Senior lecturer R. Alabugina.										
10. Gantt's schedule	Y.Sang, Professor A. Alabugin. Senior lecturer R. Alabugina.										

We give the progress of the work through the Gantt chart. The work schedule can be seen from the Gantt chart.



### 5.3 Summary of this chapter

The advantages, disadvantages, opportunities and challenges of thermal power generation are introduced, and the SWOT table is used to analyze the application of semantic Web service-based technology in thermal power plants. Finally, a Gantt chart is displayed for our thesis workflow.

						<i>page</i>
0						
<i>Изм</i>	<i>Page</i>	<i>Document #</i>	<i>Signat.</i>	<i>Date</i>	<i>13.04.01.2020. 290.06 EN</i>	50



vice is divided into K cluster clusters. Each cluster cluster has a service as a cluster center. By calculating the distance between the requested service and this service, the cluster to which the requested service belongs can be determined. Finally, through the semantic similarity method based on ontology concept, the function matching is performed, and the services that meet the requirements are returned from the obtained clusters, thereby improving the efficiency of service discovery.

4. Introduce the use of semantic Web service technology in thermal power plants, and use SWOT table to analyze the advantages and disadvantages of semantic Web service-based technology, analyze the opportunities and threats of applying technology in the work of thermal power plants, and finally target our thesis workflow made a Gantt chart display.

## 6.2 Outlook

Due to objective factors such as ability and time, there are many aspects that can be improved in this paper. The method proposed in this paper only considers the function of Web services, and different service requesters have different requirements for the requested service quality. This paper does not include service quality in the measurement scope. Service quality attribute matching can meet the special needs of users, thereby meeting the higher-level needs of users. The next step is to combine service price, reliability and response time with service discovery, calculate the quality of service, and provide users with more reliable services. At the same time, due to the limitation of service data, it has not been able to carry out experimental verification on sufficient data sets. The next step will expand the field of application of the method in this paper.

						<i>page</i>
2						
<i>Изм</i>	<i>Page</i>	<i>Document #</i>	<i>Signat.</i>	<i>Date</i>	<i>13.04.01.2020. 290.06 EN</i>	52

## CONCLUSIONS

The main research content of this paper is to face the problem of semantic Web service discovery, and it is expected to improve the clustering algorithm service by combining the document-topic model to perform service discovery. This article is divided into five chapters for presentation. The specific content and organizational structure are as follows.

**Chapter 1** is the introduction. It mainly introduces the background of service discovery, research purpose and research significance. At the same time, it introduces the research status at home and abroad. It summarizes the achievements and shortcomings in the process of service discovery.

**Chapter 2** is related theory of semantic Web services. This chapter focuses on the theoretical knowledge related to semantic Web services. First, the related theories of Web services are introduced. Web services are software modules that run on the network, are service-oriented, and are based on distributed programs. They are extensions of existing applications to the Internet. Secondly, it introduces the Semantic Web Service. Semantic Web Service is a combination of Semantic Web technology and Web service technology, which enriches the semantic information of Web services and facilitates computer understanding and storage. Finally, the semantic web service similarity is introduced based on semantic distance, information content, and attribute, which provides a theoretical basis for the following specific applications.

**Chapter 3** mainly implements a pre-processing of the service description document, mainly realizes the service name, service description, input, output and other information of the OWL-S document obtained by Jena, and performs text pre-processing. Then use Word2Vec to expand the semantics of the service description, and then establish a document-topic model based on LDA to prepare for the next chapter for service clustering.

**Chapter 4** mainly proposes a KMHC Clustering clustering algorithm combined with the document-topic model on the basis of the original clustering algorithm. Similarity calculation, further screening out services that meet the requirements, giving a framework for service discovery, and doing experimental analysis.

**Chapter 5** analyzes strong and weaknesses of technology based on the semantic WEB services, opportunities and threats of its application use of technology in work of thermal power plant. Gantt's schedule of actions for implementation of technology based on the semantic web services in work of thermal power plant.

**Chapter 6** gives summary and prospect. Summarize the work of the full text, and put forward ideas and prospects for further research in the future.

## REFERENCES

- 1 Shi, L. Analysis of WSDL document structure in Web services. Software. – 2012. – Issue 10. – P.142-143.
- 2 Chen, W. WSDL term tokenization methods for IR-style Web services discovery. Science of Computer Programming. –2012. –Issue 77. –No. 3. – P.355-374.
- 3 Wang, L. Web service composition based on extended WSDL behavior description. /L. Wang and L. Zhang //Computer Engineering. – 2014. – Issue 40. –No.1. – P.88-92.
- 4 Liu, Z. Research and application of SOAP protocol security. / Z. Liu, S.Jia and S. Zhan //Computer Engineering. –2008. –Issue 5. – P.142-145.
- 5 Al-shammary. SOAP Web Services Compression using Variable and Fixed Length Coding. /D. Khalil, I//Proceedings 2010 Ninth IEEE International Symposium on Network Computing and Applications (NCA). – 2010. – P.84-91.
- 6 Zang, T. The Research and Realization of UDDI in Service-Oriented Architecture. / T. Zang, M. Tian and L. Zhang//Journal of Shenyang University of Science and Technology. –2014. – No.6. – P.87-91.
- 7 Yang, X. Network software system integration mode and implementation design based on UDDI. / X. Yang and X. Deng //Journal of Chongqing University of Technology (Natural Science). – 2016. –Issue 30. –No.12. – P.135-139.
- 8 Yuan, H. Research on the core technology of Web service discovery based on semantics. / H. Yuan, F. Ye, X. Li and W. Peng //Computer Applications. – 2006. – No.11. – P. 2661-2663.
- 9 Zhong, M. A multi-level matching method for semantic Web services. / M. Zhong and S. Song //Computer Applications. –2007. – No.01. – P.199-201.
- 10 Martin D. The OWL Services Coalition. OWL-S 1.0 Release. [Http://www.daml.org/services/owl-s/1.0](http://www.daml.org/services/owl-s/1.0), 2007.
- 11 Sambasivam, G. An QoS based multifaceted matchmaking framework for Web services discovery / G. Sambasivam , J. Amudhavel and T. Vengattaraman //Future Computing and Informatics Journal 3- 371e383,2018.
- 12 Xu, G. Improved semantic Web service discovery method based on QoS constraints. / G. Xu, L. Ma and Z. Feng //Computer and Digital Engineering. –2018. –Issue 46. No.06. – P.1178-1181.
- 13 Li, S. Semantic Web Service Discovery Based on QoS and Fuzzy Particle Swarm Optimization. Computer Applications. –2012 – Issue 32. – No. 05. – P.1347-1350.
- 14 Cao, G. A Web service discovery algorithm that supports fuzzy QoS attributes. / G. Cao and S. Liu //Computer Applications and Software. –2011– Issue 28. – No. 12. – P. 119-121.
- 15 Hui, X. An efficient social-like semantic-aware service discovery mechanism for large-scale Internet of Things. / X. Hui, C. Hu and X. Fu //Computer Networks. – 2019. – No.152 – P.210–220.
- 16 Ma, X. Semantic Web service discovery based on IO and information content. / J. Chen and K. Li //Computer System Application. –2016–Issue 25.– No.02.– P.141-145.
- 17 Ou, W. Web service matching algorithm based on semantic similarity. / W. Ou, C.

Zeng, D. Han, Z. Peng and Y. Liu //Computer Science. – 2012. – Issue 39. – No.01. – P.92-95.

18 Farrag. Semantic Web Services Matchmaking: Semantic Distance-Based Approach. / Farrag, T. A, Saleh, A. I, Ali. H. A //Computers and Electrical Engineering. – 2013. –Issue 39. – No.02. – P.497-511.

19 Tian, H. Research on Web Service Discovery Based on User Community Relations. / H. Tian, H. Fan and W. Du // Journal of Communications. – 2015. –Issue 36. – No.10. – P.28-36.

20 Zhu, Z. A learning service discovery algorithm based on situational awareness [J]. Computer Science. – 2012. – Issue 39. – No.02. – P.132-135.

21 Yang, Y. Research on Web Service Discovery Method Based on User Context Clustering. / Y. Yang, L. Chen and B. Xie //Computer Engineering and Design – 2012. – Issue 33 – No.04. – P.1442-1446.

22 Ding, M. Accurate recommendation of knowledge discovery service based on user interest measurement. / M. Ding, Q. Bi, P. Xu and J. Li // Library and Information Service – 2019. – Issue 63. – No.03. – P.21-29.

23 Zhang, X. Multifunctional clustering of Web services based on LDA and fuzzy C-means. / X. Zhang, J. Liu, Q. Xiao and M. Shi //Journal of Central South University (Natural Science Edition). – 2018. – Issue 49. – No.12. – P.2986-2992.

24 Li, L. A QoS Management Framework for Web Services Based on WSDL. / /Computer Knowledge and Technology. – 2014. – Issue 10. – No.22. – P.5189-5191.

25 Zhang, X. One of the core technologies of Web services: SOAP protocol. / X. Zhang and J. Zhang //Electronic Technology Journal. – 2010. – No.03. – P.93.

26 Object Management Group.CORBA to WSDL / SOAP Interworking Specification. USA: Object Management Group, 2003.

27 Xun, E. Calculating the similarity of English words based on the Semantic Web. / E. Xun, W. Yan //Journal of Information. – 2006. – No.25. – P.01.

28 Zhang, S. A hybrid semantic similarity calculation method based on WordNet. / S. Zhang, W. Xing and Y. Cai //Computer Engineering and Science. – 2017 – Issue 39. – No.05 – P.971-977.

29 Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. // Proc of the 14th International Joint Conference on Artificial Intelligence. – 1995– P. 448-452.

30 Liu, X. Measuring semantic similarity in WordNet. / X. Liu and Y. Zhou // Proceedings of the Sixth International Conference on Machine Learning and Cybernetics. Hongkong, China: IEEE. – 2007 – P.3431-3435.

31 Xu, F. Comprehensive calculation of semantic similarity of ontology concepts based on SA-BP algorithm. / F. Xu, X. Ye, L. Li and J. Cao //Computer Science, 2020.

32 David Martin. OWL-S: Semantic Markup for Web Services. /Mark Burstein, Jerry Hobbs//Http://www.w3.org/Submission/OWL-S/, 2004.

33 Bird Steven Edward Loper and Ewan Klein (2009)//Natural Language Processing with Python. O'Reilly Media Inc.

34 Farrag. Sentiment classification of online consumer reviews using word vector

										<i>page</i>
5										
<i>Изм</i>	<i>Page</i>	<i>Document #</i>	<i>Signat.</i>	<i>Date</i>					13.04.01.2020. 290.06 EN	55

representations. / Farrag and Sangeet Srivastava//Procedia Computer Science. – 2018 – Issue 132. – P.1147-1153.

35 Sam Henry McInnes. Vector representations of multi-word terms for semantic relatedness. / Sam Henry McInnes, Clint Cuffy, Bridget T//Journal of Biomedical Informatics. – 2018 – No.77. – P.111-119.

36 Simone Totaro. A non-parametric softmax for improving neural attention in time-series forecasting. / Simone Totaro, Amir Hussain and Simone //Scardapane Neurocomputing. – 2020. – No.381. – P.177-185.

37 Zhang, G. Generation and application of optimal binary tree. Modern Electronic Technology. – 2008– No.10. – P.112-113.

38 Huang, R. Study on sentiment analyzing of internet commodities review based on Word2vec. / R. Huang and W. Zhang //Computer Science. – 2016. – No.43. – P.387-389.

39 Xiao, Q. Web service clustering method based on Word2Vec and LDA topic model. / Q. Xiao, B. Cao, X. Zhang, J. Liu and Y. Li //Journal of Central South University (Natural Science Edition). –2018. – Issue 49. – No.12 – P.2979- 2985.

40 Chen, K. Research on word weight calculation method based on entropy in text classification. / K. Chen, Z. Zhang and J.Long //Computer Science and Exploration.– 2016. – Issue 10. – No.9 – P.1299-1309.

41 R. Jothi. DK-means: A deterministic k-means clustering algorithm for gene expression analysis. / R. Jothi, SK Mohanty and A. Ojha //Pattern Analysis and Applications. – 2019. – Issue 22. – No.2. – P.649- 667.

42 Nicola Maffei. Hierarchical clustering applied to automatic atlas based segmentation of 25 cardiac sub-structures. / Nicola Maffei, Luca Fiorini, Giovanni Aluisio //Physica Medica. –2020. – No.69. – P.70-80.

43 Severino F Galan. Comparative evaluation of region query strategies for DBSCAN clustering. Information Sciences. – 2019. – No.502 – P.76-90.

44 Mehdi Hassan. Robust spatial fuzzy GMM based MRI segmentation and carotid artery plaque detection in ultrasound images. /Iqbal Murtza, Aysha Hira//Computer Methods and Programs in Biomedicine. – 2019. – No.175 – P.179-192.

45 Xie, J. Unsupervised feature selection algorithm based on spectral clustering. / J.Xie, L. Ding and M. Wang // Journal of Software. [https:// doi.org/10.13328/ j.cnki.jos.005927](https://doi.org/10.13328/j.cnki.jos.005927).

46 Chen, T. A Combined K-Means and Hierarchical Clustering Method for Improving the Clustering Efficiency of Microarray/ T.Chen and Y.Chen // Proceeding of 2005 International Symposium on Intelligence Signal Processing and Communication System, 2005 .

47 P.Anitha. RFM model for customer purchase behavior using K-Means algorithm. / P.Anitha and Malini M.Patil //Computer and Information Sciences, 2019.

48 Zhang, C. Nearest neighbor filling algorithm for missing data based on cluster analysis. / C. Zhang, H. Feng, Kai. Jin and T. Yang // Computer Applications and Software. – 2014. – Issue 31. – No.05. – P.282-284.

49 Klusch M. OWLS-TC4: OWL-S service retrieval test collection version 4 [EB /

OLJ/ Khalid M, Kapahnke P. // [http://pro-jecs.semWebcentral.org/frs/download.php/487/OWLS-TC4\\_PDDL.zip](http://pro-jecs.semWebcentral.org/frs/download.php/487/OWLS-TC4_PDDL.zip).

50 Qu, J. Subject topic evolution analysis based on topic filtering and topic correlation. / J. Qu and S. Ou // Data Analysis and Knowledge Discovery. – 2018. – Issue 2. – No.01 – P.64-75.

51 Ma, Y. Design and Implementation of Emergency Command Center for Thermal Power Unit Production Based on Internet Technology. / Y. Ma, H. Han, Y. Huang, W. Du and, Y. Ni // Journal of Shenyang Institute of Engineering (Natural Science Edition). – 2020. – Issue 16. – No.01–. P.79-82.

52 Gong, W. Research on the regulation of Xinjiang thermal power industry from the perspective of low-carbon economy. Jilin University, 2011.

53 Xie, J. Research on Theory and Method of Information Sharing and Interoperability in Heterogeneous Application of Power System. Huazhong University of Science and Technology, 2008.

54 Zhao, J. Analysis of the application of Gantt chart in human resources management. Commercial News. –2019. – Issue 15. – P.195-196.

										<i>page</i>
7										
<i>Изм</i>	<i>Page</i>	<i>Document #</i>	<i>Signat.</i>	<i>Date</i>	<i>13.04.01.2020. 290.06 EN</i>					57