Service-Oriented Search: A Ranking and Retrieval Model based on Ontology of Service Relation

Hau-Wei Chang[1],Chiung-Wei Huang[1,2], and Hahn-Ming Lee[1,3]
Department of Computer Science and Information Engineering[1]
National Taiwan University of Science and Technology
Department of Electronic Engineering[2]
Ching Yun University
Institute of Information Science[3]
Academia Sinica, Taiwan
E-mail: hmlee@mail.ntust.edu.tw

ABSTRACT

Due to the general search engine might not work well on service-based search, a novel search approach, named Service-Oriented Search, based on applying the ontology of service is proposed in this paper. The ontology which reveals the knowledge of relationship among services is constructed by extracting the working flow of the Web Services and is employed to guide the search of Web services. At the ranking stage, we develop a simple but robust strategy to measure the distance between verb term and service term in order to find out the relevant and important Web pages that cover the E-Service information of interest. At last, the experimental results confirmed that the proposed approach not only is feasible, but also outperforms the search function of Taiwan's E-Government portal.

Keywords: Service-Oriented Search, ontology, service relation, E-Service, E-government.

1 : INTRODUCTION

The rapid growth of Web pages makes it difficult for a user to find out his or her relevant information from the Internet. Also, many governments and enterprises work hard to provide useful E-services on the Internet [1][2][3][4][5]. Thus, users can apply those services at home or at anywhere whenever they need. For example, when the user wants to lodgment or filing of taxation on the Internet (we called it "E-taxing" service here), the search results of most general search engines might return the Web pages with the term "E-taxing" but without the information about how to apply for the "E-taxing" service. Users have to confirm each link for finding their target web link. Also, there are two major issues in using current general search engines for a service-based search. First, **the Limits on Making Query:** Users may have difficulty in making an appropriate query to search engines because they are unfamiliar with the knowledge of service. Also, the keyword-matching method usually returns too many irrelevant search results to users [6][7]. Second, **the Lack of Domain Knowledge on relations：** In service domain, there exists valuable information behind the service, i.e., service relations. When applying for a service, users might need to apply for some other related services. For example, users need to download the dedicate software before they conduct the e-taxing. In addition, due to a service often includes many procedures; users need to complete some procedures by turns if they want to accomplish the service. Thus, we propose a new search method, named Service-Oriented Search. It first constructs the service ontology by extracting the working flow of the service guided by the service experts. Next, a novel ranking approach based on measuring the distance between verb term and service term is developed for finding out the relevant search results. Then according to ontology relation, the Web pages of search results are listed in a tree-form category which helps users discovering their Web pages of interest quickly. To conclude, we employ the service relation to guide the service-based search for Internet E-services. In addition, a prototype system, named Service-Oriented Search portal, is built for verifying our ideas. Furthermore, some experiments are conducted to evaluate the performance of our proposed method.

2 : SYSTEM ARCHITECTURE

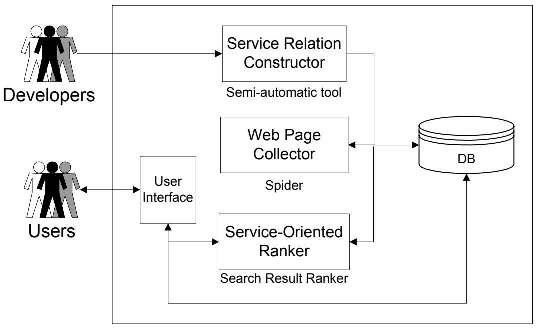


Figure 1. Architecture of the proposed system.

Figure 1 shows the architecture of the proposed system. Three major components are included in our system: (1) Service Relation Constructor, (2) Web Page Collector, and (3) Service-Oriented Ranker. The Service Relation Constructor is guided by the developers (experts) for constructing service relations and saving the ontology in frame structures. The Web Page Collector fetches the appropriate Web pages from the internet. Then, the Service-Oriented Ranker helps to rank the retrieved pages according to the service relations for providing a better recommendation on e-service Web pages. In what follows, we introduce them in detail.

2.1 : SERVICE RELATION CONSTRUCTOR

The architecture of Service Relation Constructor is shown in Figure 2. The major agents are the Developing Agent, CKIP Agent, Verb_Extraction Agent, and Service_Extraction Agent. The Developing Agent provides an interface for service developers to paste in the service workflow or procedures in text format and save them into the Service_Description DB. And the CKIP Agent parse the service workflow text into terms by using CKIP parser [8]. Then the parsed terms will be stored into Term DB. The Verb_Extraction Agent provides an interface for service developers to choose verb terms that are related to services. The Service_Extraction Agent allows service developers to select needed service terms from the Term DB. Finally, the frame of service relation is constructed and stored in Service_Relation DB. The template of service relation ontology is shown in Figure 3. As mentioned previously, there exists relationship between services in most situation. For example, when a user wants to apply for E-taxing, he should download E-taxing software first. Therefore, we can say that there exists "download" relationship between "E-taxing" and "E-taxing software". The frame of service relation about a service is a concrete form for representing the relationship between the service and other services. Figure 4 illusttrates the frame of service relation about "apply for E-taxing".

■ Developing Agent

For constructing a service ontology, the Developing Agent provides an interface for Service Developers to fill in the service information, i.e., service title, service thesaurus, workflow, etc as shown in Figure 5. After developers inserted all the necessary information, the developing agent stores the service thesaurus in Service_Thesaurus DB and the description information of service into Service_Description DB.

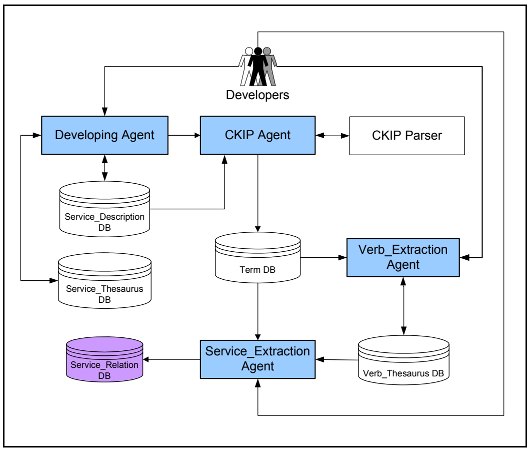


Figure 2. Architecture of the Service Relation Constructor.

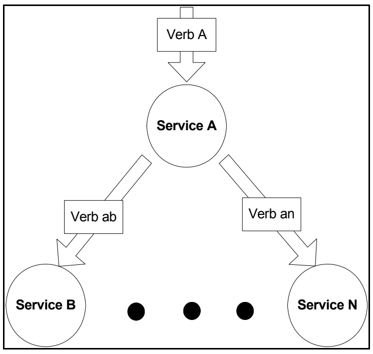


Figure 3. The template of service relation ontology.

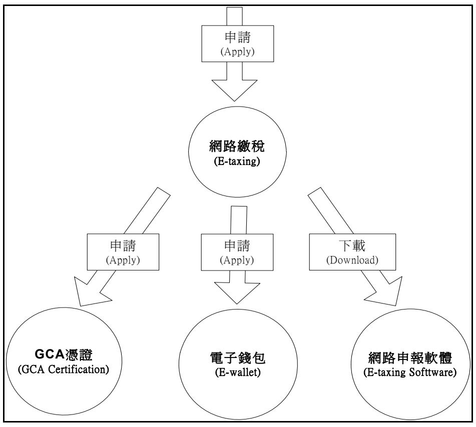


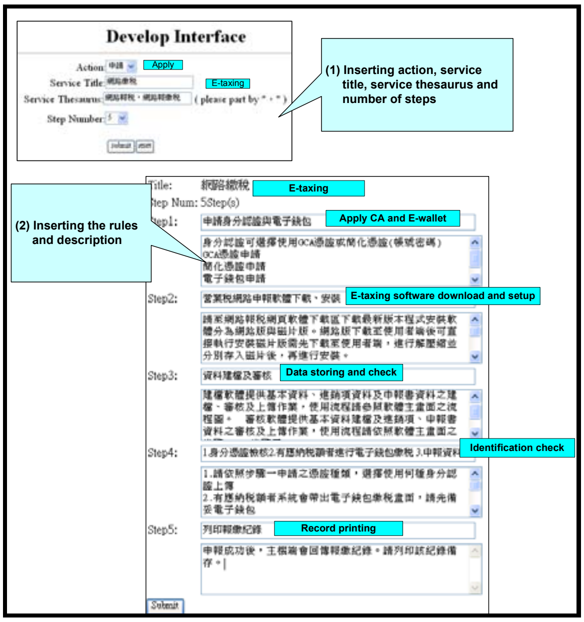
Figure 4. A Frame of service relation about "apply for E-taxing".

■ CKIP Agent

The CKIP Agent aims at extracting service workflow from Search_Description DB and parsing it into terms by invoking the CKIP (Chinese word segmentation) parser [8]. After parsing, the CKIP Agent stores parsed terms into the Term DB.

■ Verb_Extraction Agent

The Verb_Extraction Agent provides an interface for



Develop Interface

Action: 申請 Apply

Service Title: 網路繳稅 E-taxing

Service Thesaurus: 網路報稅，網路報繳稅 (please part by "，")

Step Number: 5

(1) Inserting action, service title, service thesaurus and number of steps

Title: 網路繳稅 E-taxing

Step Num: 5Step(s)

Step1: 申請身分認證與電子錢包 Apply CA and E-wallet

身分認證可選擇使用GCA憑證或簡化憑證(帳號密碼)
GCA憑證申請
簡化憑證申請
電子錢包申請

Step2: 營業稅網路申報軟體下載，安裝 E-taxing software download and setup

請至網路報稅網頁軟體下載區下載最新版本程式安裝軟體分為網路版與磁片版。網路版下載至使用者端後可直接執行安裝磁片版需先下載至使用者端，進行解壓縮並分別存入磁片後，再進行安裝。

Step3: 資料建檔及審核 Data storing and check

建檔軟體提供基本資料、進銷項資料及申報書資料之建檔、審核及上傳作業，使用流程請參照軟體主畫面之流程圖。 審核軟體提供基本資料建檔及進銷項、申報書資料之審核及上傳作業，使用流程請依照軟體主畫面之

Step4: 1.身分憑證檢核2.有應納稅額者進行電子錢包繳稅 3.申報資料 Identification check

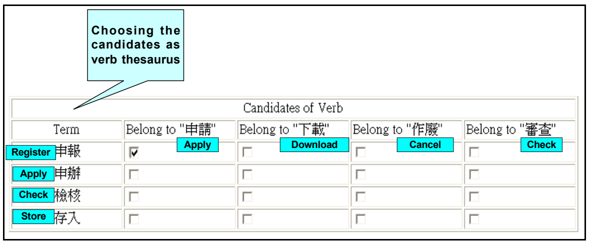
1.請依照步驟一申請之憑證種類，選擇使用何種身分認證上傳
2.有應納稅額者系統會帶出電子錢包繳稅畫面，請先備妥電子錢包

Step5: 列印報繳紀錄 Record printing

申報成功後，主機端會回傳報繳紀錄。請列印該紀錄備存。|

(2) Inserting the rules and description

Figure 5. Insert the workflow of a service.



Choosing the candidates as verb thesaurus

Candidates of Verb

Term	Belong to "申請" Apply	Belong to "下載" Download	Belong to "作廢" Cancel	Belong to "審查" Check
Register 申報	☑	□	□	□
Apply 申辦	□	□	□	□
Check 檢核	□	□	□	□
Store 存入	□	□	□	□

Figure 6. Interface for developers to choose terms as verb thesaurus.

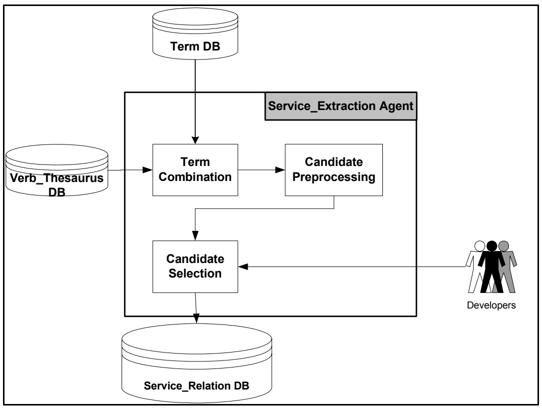


Figure 7. Architecture of Service_Extraction Agent.

service developers to select verbs and classify them into predefined verb classes. The predefined verb classes are "申請類(Apply class)", "作廢類(Cancel class)", "檢核類(Check class)", and "存入類(Store class)". After investigation, we found that the verb class is very important for a service because it implies what kinds of action the service needs to take. Figure 6 shows the interface of Verb_Extraction Agent.

■ **Service_Extraction Agent**

The Service_Extraction Agent has three main components: Term Combination, Candidate Preprocessing, and Candidate Selection. Figure 7 depicts the architecture of Service_Extraction Agent. The Term Combination Agent finds out terms that are possible to be part of the service title and then combines them appropriately into a service title. Due to the service title always appears around some verbs as mentioned previously, the agent fetches verbs from Verb_Thesaurus DB and check if they agree with this kind of situation. If so, those terms that appear around the verbs are regarded as the candidates of service title. But owing to the CKIP parser is a general purpose parser, the parsed terms are not necessary to meet the requirement of dealing with Web service terms. Some service terms, e.g., "國民身份證" (citizen ID), might be parsed into two terms, "國民" (citizen) and "身份證" (ID). Thus, for dealing with this problem, we also develop a strategy for combining this kind of terms appropriately.

Next, the Candidate Preprocessing component filters the noise of candidate terms. Two filtering strategies are listed as follows.

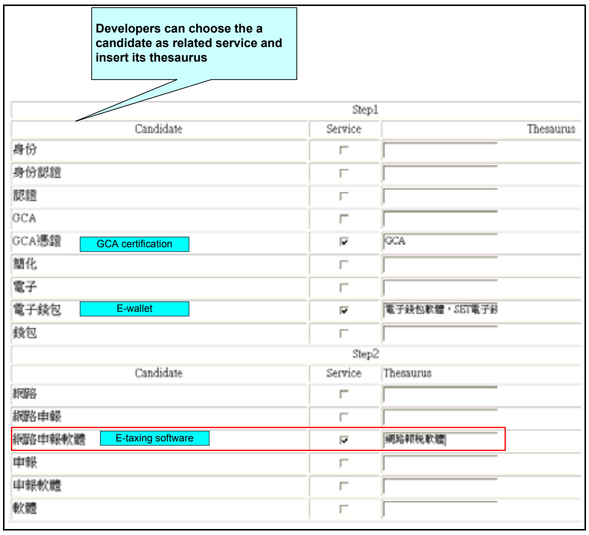
- **Removing the single word term:** We assume that a single word is meaningless because it is difficult for service developers to find out information in such a single term.

- **Removing candidates that contain symbols:** Candidate terms may contain symbols after Term combination. These kinds of candidate are meaningless in Chinese.

At last, the Candidate Selection provides an interface for service developers to select candidate terms. If they consider "網路申報軟體(E-taxing software)" is related to "申請網路繳稅 (apply for E-taxing)" service, he could click the checkbox in the Service field and enter the thesauruses in the Thesaurus field. After that, the Candidate Selection component stores the relation in Service_Relation DB and saves the thesaurus in Service_Thesaurus DB.

2.2 : WEB PAGE COLLECTOR

The main task of Spider is to crawling Web pages from the E-government portal in Taiwan [5]. For example, it crawls Web pages about E-taxing by querying the E-government portal with the service



Developers can choose the a candidate as related service and insert its thesaurus

Step1

Candidate	Service	Thesaurus
身份	□	
身份認證	□	
認證	□	
GCA	□	
GCA憑證	☑	GCA
簡化	□	
電子	□	
電子錢包	☑	電子錢包軟體，SET電子錢
錢包	□	

Step2

Candidate	Service	Thesaurus
網路	□	
網路申報	□	
網路申報軟體	☑	網路報稅軟體
申報	□	
申報軟體	□	
軟體	□	

Figure 8. Interface for developers to choose candidates.

thesaurus on E-taxing. The thesaurus of E-taxing is extracting from Service_Thesaurus DB. After crawling, it saves the Web pages in Metasearch_Result DB.

2.3 : Service-Oriented Ranker

The main task of Service-Oriented Ranker is to rank the extracted Web pages. Also according to some observations, two assumptions are made in our ranking strategy.

- If the title or content of a Web page contains both the service term and verb terms related to service, the page obtains higher ranking score.
- The less distance between the verb term and the service term in the title or content of a Web page, the page gets higher ranking score.

For example, if the verb term is adjacent to the service term such as ”申請網路繳稅(apply for E-taxing)” in the title/content of Web pages, we think that the Web pages are meaningful to users who want to apply for E-taxing. But in practice, the verb term and service term might not necessary adjacent such as “申請服務：1.網路報稅(apply service: 1.E-taxing).” Therefore, we rank the Web pages by the distance between the verb terms and service terms. The less distance between the verb terms and the service terms in the title/content of a Web page, the higher ranking score the page gets.

And the ranking score is calculated as:

$$Rank\_score_m(verb, service) = W_1 \times C\_dist_m^* + W_2 \times T\_dist_m^* + W_3 \times T\_ser_m^* \qquad (1)$$

where m denotes the Web page,
verb denotes the verb term,
service denotes the service term,
W$_1$, W$_2$, W$_3$ denote the weights.

$$C\_dist_m = \sum_{i=1}^{v\_num} \left[\frac{1}{\min_{j=1}^{s\_num} \left(|v_i - s_j| \right)} \right] \qquad (2)$$

$$T\_dist_m = \frac{relate\_len(t_m, VS)}{length(t_m)} \times \sum_{k=1}^{v\_num^*} \left[\frac{1}{\min_{l=1}^{s\_num^*} \left(|v_k^* - s_l^*| \right)} \right] \qquad (3)$$

where VS denotes V (set of v) Union S (set of s).
related_len() denotes the length of VS which appears in tm.

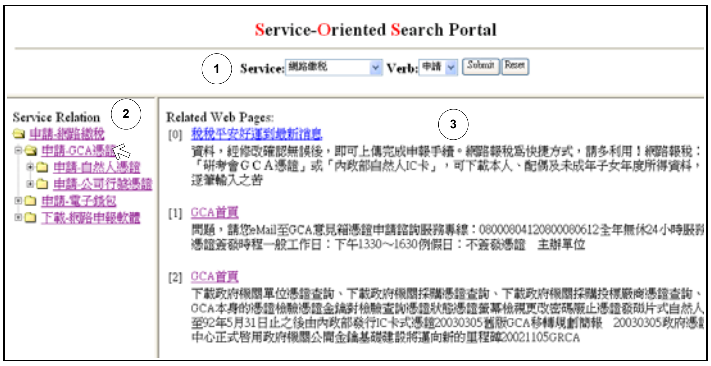
$$T\_ser_m = \frac{relate\_len(t_m, S)}{length(t_m)} \qquad (4)$$

C_dist$_m$ measures the distance between the thesaurus of *verb* term *(v)* and thesaurus of *service* term (*s*) in the content of a Web page as defined in equation (2). *T_dist*$_m$ calculates the distance between the thesaurus of *verb* term *(v)* and thesaurus of *service* term (*s*) in the title of a Web page as defined in equation (3). *T_ser*$_m$ is used to judge if *service* term occurs in the title of a Web page as defined in equation (4). An example of ranking calculation in Service-Oriented Ranker is shown in Figure 9.

3 : EXPERIMENTAL RESULTS

For evaluating the proposed approach, we conduct experiments focusing on the E-taxing service of Taiwan. The E-government portal [5] is the E-service portal of Taiwan government. It helps users to find out the Web pages related to the E-service of their interest. In addition, a prototype system, named Service-Oriented Search (SOS) portal, is constructed for verifying our proposed approach. Figure 10 shows a snapshot of our SOS portal. For comparison, we list the search results of our Service Oriented Service portal and that of E-government portal about the “申請網路繳稅(apply for E-taxing)” service.

Table 1 shows the top 5 URLs crawled form E-government portal. The “Service Node” in Table 1 is correspondent to the “Service Node” in Figure 11. It means that the top 5 URLs in node 1 are the top 5 URLs crawled from E-government portal with the query term “網路繳稅(E-taxing)”. Table 2 shows the top 5 URLs of our SOS portal. It means that the top 5 URLs in node 1 are top 5 URLs in the service “申請網路繳稅(apply for E-taxing)” in our system. Figure 12 shows the precision score of URLs that contain related information about E-taxing in top 5 URLs. Figure 13



Service-Oriented Search Portal

1 Service: 網路繳稅 Verb: 申請 Submit Reset

Service Relation 2

- 申請-網路繳稅
  - 申請-GCA憑證
    - 申請-自然人憑證
    - 申請-公司行號憑證
  - 申請-電子錢包
  - 下載-網路申報軟體

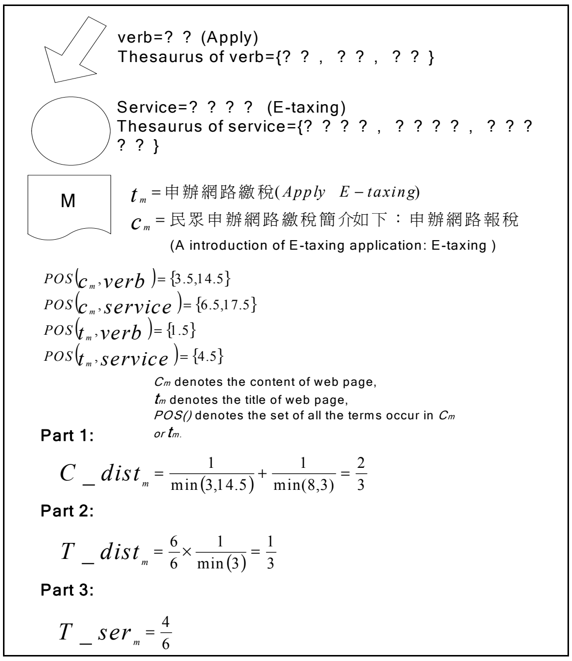
Related Web Pages: 3

[0] 稅稅平安好運到最新消息
資料，經修改確認無誤後，即可上傳完成申報手續。網路報稅為快捷方式，請多利用！網路報稅：「研考會ＧＣＡ憑證」或「內政部自然人IC卡」，可下載本人、配偶及未成年子女年度所得資料，逐筆輸入之苦

[1] GCA首頁
問題，請您eMail至GCA意見箱憑證申請諮詢服務專線：08000804120800080612全年無休24小時服
憑證簽發時程一般工作日：下午1330~1630例假日：不簽發憑證　主辦單位

[2] GCA首頁
下載政府機關單位憑證查詢、下載政府機關採購憑證查詢、下載政府機關採購投標廠商憑證查詢、GCA本身的憑證檢驗憑證金鑰對檢驗查詢憑證狀態憑證螢幕檢視更改密碼廢止憑證發碼片式自然人至92年5月31日止之後由內政部發行IC卡式憑證20030305舊版GCA移轉規劃簡報　20030305政府憑中心正式啓用政府機關公開金鑰基礎建設將邁向新的里程碑20021105GRCA

Figure 10. A snapshot of the SOS portal.



verb=? ? (Apply)
Thesaurus of verb={? ? , ? ? , ? ? }

Service=? ? ? ? (E-taxing)
Thesaurus of service={? ? ? ? , ? ? ? ? , ? ? ? ? ? }

t_m = 申辦網路繳稅$(Apply \quad E-taxing)$
c_m = 民眾申辦網路繳稅簡介如下：申辦網路報稅
(A introduction of E-taxing application: E-taxing)

$POS(c_m, verb) = \{3.5, 14.5\}$
$POS(c_m, service) = \{6.5, 17.5\}$
$POS(t_m, verb) = \{1.5\}$
$POS(t_m, service) = \{4.5\}$

Cm denotes the content of web page,
tm denotes the title of web page,
POS() denotes the set of all the terms occur in *Cm* or *tm*.

Part 1:

$$C\_dist_m = \frac{1}{\min(3,14.5)} + \frac{1}{\min(8,3)} = \frac{2}{3}$$

Part 2:

$$T\_dist_m = \frac{6}{6} \times \frac{1}{\min(3)} = \frac{1}{3}$$

Part 3:

$$T\_ser_m = \frac{4}{6}$$

Figure 9. An example of the ranking calculation in Service-Oriented Ranker.

Table 1. The number of URLs that contains related information (in Chinese)

Service Node	SOS portal	E-government portal
(1)網路繳稅	5	1
(2)GCA 憑證	4	2
(3)電子錢包	2	0
(4)網路申報軟體	3	2
(5)自然人憑證	4	3
(6)公司行號憑證	2	1
(7)二維條碼報稅軟體	0	0
(8)IRC 報稅軟體	5	0

presents the precision of URLs that contain related software about E-taxing in top 5 URLs. In Figure 12, for example, the precision score of node 1 is one (5/5) because there are five URLs in top 5 URLs containing the information about how to apply E-taxing.

In Figure 12, we can find that precision score in our system is higher than that in E-government portal. Therefore, it is reasonable that the top 5 URLs in our system contain more related information and related software about E-taxing than that in E-government portal. Also in Figure 12, precision score in node 7 of Service Oriented Search portal is zero. It is because all of top 5 URLs only contain the "二維報稅軟體" (2D E-taxing software) but not include the information about 2D E-taxing software. When a user wants to find something about 2D E-taxing software for service-based search, the best search results might be the URL that contains the software. In Figure 12 and Figure 13, the precision score in node 3 of each system is always low. It is because that the "電子錢包 (E-wallet)" service is not provided by the government. Therefore, we can not collect enough URLs about "電子錢包(E-wallet)" in E-government portal.

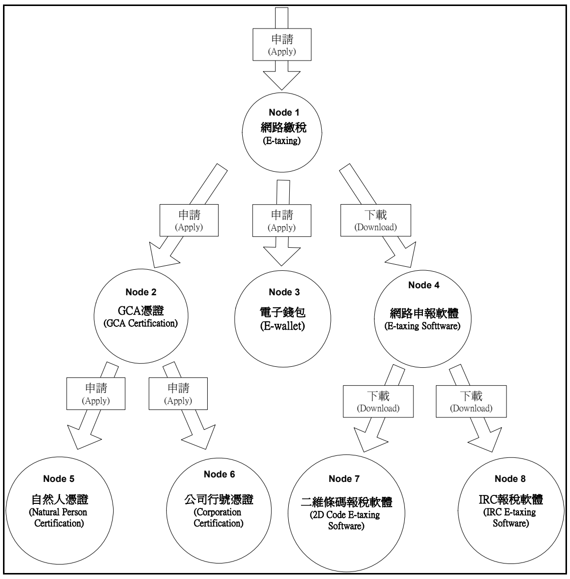


Figure 11. Frame of service relation about "apply for E-taxing".

Table 2. Top 5 URLs for two service nodes in Service-Oriented Service portal

Service Node	Order	Web Site Name	URL
(1)網路繳稅	1	網路報稅密碼申請 (E-taxing Password)	http://tax.nat.gov.tw/ca5main.htm
	2	財政部財稅資料中心 (Ministry of Financial)	http://web.mofdpc.gov.tw/page_1_3_3_5.htm
	3	網路繳稅(E-taxing)	http://www.kctax.gov.tw/02/02_04_02.htm
	4	網路報繳稅(E-taxing)	http://tax.nat.gov.tw/
	5	網路報繳稅(E-taxing)	http://www.kctax.gov.tw/02/02_04.htm
(2)GCA憑證	1	政府憑證管理中心 (GCA)	http://www.pki.gov.tw/
	2	GCA	http://gca.nat.gov.tw/
	3	電子公路監理網 (Electronic Motor Vehicle System)	http://www.mvdis.gov.tw/news/main_news.htm
	4	網路報繳稅(E-taxing)	http://tax.nat.gov.tw/blr/helpd.html
	5	GCA-檔案下載 (Software Download)	http://gca.nat.gov.tw/repository.htm

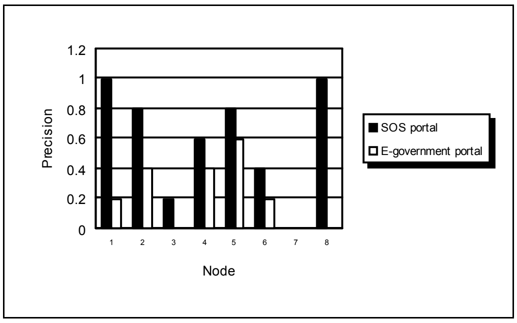


Figure 12. The precision score of URLs containing related information about E-taxing.

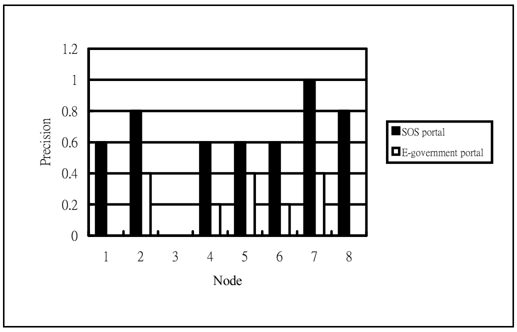


Figure 13. The precision score of URLs containing related software about E-taxing.

4 : CONCLUSION

We have proposed a novel search approach, named as Service-Oriented Search, which applies ontology of service relation for guiding the Web search. Through the help of service experts, the frames of service relation which reveals the knowledge of services are constructed. Then, the similarity measure on Web pages is conducted by a simple but robust strategy, i.e., we calculate the distance between verb term and service term to distill service-oriented Web pages.

Furthermore, according to the ontology relation, the search results are listing in a tree-form category which helps users discovering their Web pages of interest quickly. At last, the experimental results confirm that our method not only is feasible, but also outperforms the search function of Taiwan's E-Government portal.

Finally, there are some aspects that we can continue to improve our Service Oriented Search portal:

- **Constructing more frames of service relation.** Currently, we only construct the frame of service relation on "apply for E-taxing". Other service relations need to be built up for further investigations.
- **Broadening the crawling scope.** In this paper, our crawling strategy focuses on Web pages about government service. Therefore, we may broaden our crawling for enriching the service scope.
- **Applying the service relation to ranking method.** Experiments reveal that applying service relation in the search of service-oriented search works well. We would like to further apply the service ontology in the ranking of search results in the near future.

REFERENCES

[1]Benchmarking E-government: A Global Perspective, http://unpan1.un.org/intradoc/groups/public/documents/un/unpan008626.pdf

[2]E-Government for All, http://cmc.edc.org/library/ egov4all.html

[3]Global E-Government Survey, http://www.insidepolitics.org/egovt01int.html

[4]E-Government Strategy, http://www.whitehouse.gov/omb/inforeg/egovstrategy.pdf

[5]E-government Portal of Taiwan, http://www.gov.tw/.

[6]S. Chakrabarti, M. Berg, B. Dom, "Distributed Hypertext Resource Discovery Through Examples," *The VLDB Journal*, pp. 375-386, 1999.

[7]Q. Yang, H.F. Wang, J.R. Wen, G. Zhang, Y. Lu, K.F. Lee, H.J. Zang, "Toward a Next Generation Search Engine," *Proceedings of the 6th Pacific Rim Artificial Intelligence Conference*, 2000.

[8]CKIP, http://godel.iis.sinica.edu.tw/CKIP/.

[9]B. Yuwono, D.L. Lee, "Search and Ranking Algorithms for Locating Resources on the World Wide Web," *Proceedings of the 12th International Conference on Data Engineering*, pp. 164-171, 1996.