Towards a Public Cloud Services Registry

Ahmed Mohammed Ghamry, Asma Musabah Alkalbani, Vu Tran, Yi-Chan Tsai, My Ly Hoang, and Farookh Khadeer Hussain(✉)

Decision Support and e-Service Intelligence Lab Centre for Artificial Intelligence School of Software, University of Technology, Sydney, NSW 2007, Australia
{Asma.M.Alkalbani,Vu.Tran-1,Yi-Chan.Tsai}@student.uts.edu.au, Farookh.Hussain@uts.edu.au

Abstract. Cloud services registry is a cloud services datadase which contains thousands of records of cloud consumers' reviews and cloud services, such as Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). The data set is harvested from a web portal called www.serchen.com. Each record holds detail information about the service such as service name, service description, categories, key features, service provider link and review list. Each review contains reviewer name, review date and review content. This work is an extension of our previous work Blue Pages data set [6]. The data set is valuable for future research in cloud service identification, discovery, comparison and selection.

Keywords: Cloud services data set · Service discovery · Web harvesting

1 Introduction

With the modernization of IT technologies, cloud computing a critical component for organizations digital business by delivering computing resources as a service, such as Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Cloud computing with on-demand services brings many benefits to businesses by ensuring service availability and scalability, saving cost, improving efficiency and productivity [11]. In recent years, there has been an ongoing demand for businesses to adopt public cloud services. A forecast conducted by Cisco predicts that both public and private cloud markets will continue to grow between 2015 and 2020 [12]. However, the public cloud is expected to grow stronger than the private cloud. Service discovery and selection are challenging tasks for businesses as a result of the rapid growth in cloud services market. Using search engines, such as Bing and Google, is a traditional way to discover cloud services on the World Wide Web (WWW). This method enables consumers to find cloud services-related information. However, the consumers will need to peruse the content of each website content and compare service offerings manually to decide the most suitable cloud service for their needs, such as Database tools. Their search results may return irrelevant information, which makes the search effort more difficult.

A. Bouguettaya et al. (Eds.): WISE 2017, Part I, LNCS 10569, pp. 290–295, 2017.
DOI: 10.1007/978-3-319-68783-4_20

Another way to locate cloud services information is via web portals such as GetApp (https://www.getapp.com/). web portals offer a significant number of records of cloud service, but it is difficult to obtain up-to-date information when data changes. In recent years, therefore, cloud service discovery has attracted significant attention in the research community. Ontology-based cloud discovery approaches have been proposed in most of researches in cloud service discovery area. An ontology in semantic web technologies is often used to illustrate information and relationships. Cloud services ontology is usually based on the National Institution Standards and Technology (NIST) cloud computing standard and specifications [11]. For instance, a crawler engine is built to harvest web portals using cloud ontology concepts [13].

In [3,4] the authors carried out a categorization for cloud services based on cloud service ontology concepts. Other studies by [2,15] also categorized cloud services using business ontology concepts to help businesses find the most suitable cloud business service.

To identify the most appropriate solutions for cloud service discovery and selection, Alkalbani and Hussain [7] conducted a thorough review of current cloud discovery approaches and proposed eight criteria for comparing the current cloud discovery approaches. In general, a challenge for cloud service discovery approaches is to provide a consistent public registry with the ability to automatically identify and update service information in the registry [7,14].

A cloud services registry can offer a complete list of cloud services from multiple resources to give an insight into service information and facilitate the process of discovering and selecting cloud services with up-to-date information in the market. To address this need, the aim of this work is to provide a comprehensive data set which contains real world SaaS, PaaS and IaaS cloud service offerings. The data set is our future research extension to the Blue Pages data set [6].

To achieve our objectives, we used the latest web scraping technologies to harvest useful cloud service data from a cloud service web portal. web scraping is a well-known technique for harvesting data from the WWW by gathering web page data. Serchen Marketplace, www.serchen.com [1] is an excellent source of cloud services. This website contains thousands of records on cloud services in SaaS, PaaS and IaaS categories. It also has a significant number of reviews from customers about these services. In the developed harvesting tool, sample metadata are collected by tool users to provide a thorough analysis of the selected website. The collected sample metadata is analyzed by an algorithm to learn its structure via HTML page format, which serves as an the indicator to locate exactly where and how to harvest the information that the users expect. The metadata structure learned is the core for the harvesting process, in which thousands of web pages are extracted from the structure using scripting programming techniques. The data set holds all the useful services data obtained from the Internet portal. Each record stores service information such as service name, service description, categories, key features, service provider link and review list. Each review has information about reviewer name, review date and review content. At this stage,

we only harvest one portal with a view to developing a tool with the capacity to harvest future web portals. The data set contains more than 6000 cloud services and 5000 customer reviews. We plan to provide a mechanism for updating data to obtain up-to-date information from web portals. The rest of this paper is structured as follows: Sect. 2 discusses some related works on crawling cloud service data sets from the web. The results and analysis presented in Sect. 3. The conclusion and further improvements will be presented in Sect. 4.

2 Related Work

There is very little research in the literature that focuses on providing cloud service data sets. In [9] Han and Sim and [10] Kang and Sim, the authors designed and developed a crawler engine to discover cloud services via the web and provided statistics and details on available cloud services in their local repository. In practice, the authors introduced a cloud service ontology that supported the discovery and collection of cloud services web links via general search engines such as Google and Yahoo. The cloud ontology was also used to filter the collected data set to remove invalid service web links. The final result showed that there are around 5883 valid cloud services available on the web including 1552 cloud services implemented using Service Oriented Computing (SOC). However, the study does not provide a complete view of cloud services published on the web. The reason is that the authors used only the most general and abstract ontology concepts to search via general search engines for cloud services as IaaS, PaaS, SaaS, storage, communication, etc., which might provide only a cross-section of what is available on the web and therefore lead to inaccurate conclusions. Furthermore, the collected data sets lack main service information such as service provider name and service URL; also they include inadequate cloud services information that does not have semantic meaning.

Gong and Sim [8] presented a method to deal with the lack of cloud services information based on crawlers that are designed to collect information from several web resources. The authors targeted the most popular services providers, such as AWS, Rackspace, and Gogrid. A crawler is designed for each cloud service provider which gathers only service specification and service price for each service from the cloud provider's website. The data is stored locally, and a k-mean clustering algorithm performs is performed on them to detect the similarities and differences between cloud services. One of the limitations of this study is that it does relies only on popular service providers where it crawls the world wide web for cloud services is another potential concern because it provides a complete view of what cloud services are available through the web. Another problem with this approach is that it does not take into account the need to provide a publicly available clud market data set.

Another open source crawler engine providing a central repository for cloud services in SaaS was proposed by Alkalbani et al. [5]. This approach used the Nutch-Hadoop crawler to harvest cloud web portals and store data in SQL database. However, the drawback of this data set is the lack of QoS information and

Service Name	Service URL	Service Key Features	Service Categories	Service Description
Xero	www.xero.com	reconcilication	Accounting Software	offices in the United States,
Billy	www.billyapp.com	Accounts Receivable	Accounting Software	cloud-based accounting
Cantorix	www.cantorix.com	Recurring Invoicing	Accounting Software	new online invoicing
Bean Cruncher	beancruncher.com	Accounts Receivable	Accounting Software	Software is your best choice
AccountMate Software	www.accountmate.com/	Multiple Deployment	Accounting Software	AccountMate develops and
CosmoLex	www.cosmolex.com	Billing & Invoicing	Accounting Software	practice management
ZarMoney	www.zarmoney.com/	Accounts Payable	Accounting Software	software with invoicing,
Flare Cloud Accounting	www.flareapps.com	dashboard	Accounting Software	small business's and
docSTAR	www.docstar.com	Software	Accounts Payable	division of Astria Solutions
Nvoicepay	www.nvoicepay.com	International Payments	Accounts Payable	payments solutions to
expex	www.expexinc.com	Expense Management	Accounts Payable	managing your bills in the
Clear Pier	www.clearpier.com		Ad Networks	Premium Performance
Qadabra	www.qadabra.com		Ad Serving	got the magic.Earning more
Epom Ad Server	www.epom.com	Cross-channel ad	Ad Serving	multiplatform ad serving
Admixer	admixer.net	management	Ad Serving	Ltd. Group is a technology
Yext	www.yext.com		Advertising Software	with the award-winning
Onters	onters.com/	video	Affiliate Marketing	agency working in animation
TrackingDesk	www.trackingdesk.com	Analytics	Affiliate Marketing	media management, results
Appsee	www.appsee.com	Recordings	Analytics	analytics platform provides in-
Natero	www.natero.com	Lifetime Value	Analytics	success platform to merge
Altocloud	www.altocloud.com	Messaging	Analytics	and Digital Messaging: Real-
Logi Analytics	www.logianalytics.com		Analytics	self-service analytics,
Morningfame	www.morningfa.me	Analytics	Analytics	the web - like a blog post or a
Stytch	www.stytch.com		Analytics	platform that provides
Teamgraph	teamgraph.io	Analytics	Analytics	solution for CXOs to track
Attribution	attributionapp.com	attribution	Analytics	for all your marketing
Twitter Counter	twittercounter.com	Historical Data	Analytics	in 2008 and today tracks over
Snappii	www.snappii.com	Development	App Building Software	on the principles of
Bubble	bubble.is	mobile apps	App Building Software	more generally the need for
Shoutem	www.shoutem.com	Special Deals	App Building Software	comprehensive mobile app
Manektech	www.manektech.com/	software development	App Building Software	software & web development
Contus	www.contus.com	Software	App Building Software	on demand delivery software
FuGenX Technologies	fugenx.com	firms florida	App Building Software	Apps Development Company
Knack	www.knack.com	you need	App Building Software	easily build beautiful, data-
TechValens	www.techvalens.com		App Building Software	LLC is a leading provider in
CATS	www.catsone.com	Portal Branding	Applicant Tracking	based applicant tracking
HireGround	startdate.ca	Search candidate	Applicant Tracking	technology solutions
Bullhorn	www.bullhorn.com/		Applicant Tracking	CRM solutions for
RecruitBPM	www.recruitbpm.com	system	Applicant Tracking	applicant tracking system that
Jobularity	jobularity.com	job search	Applicant Tracking	Recruitment Platform.
Cloudwalk Hosting, LLC	www.cloudwalks.com	24x7 Free Chat, Email	Application Hosting	leading QuickBooks Hosting
Elucentra Cloud Services	www.elucentra.com	and Phone Support	Application Hosting	Elucentra provides an easy-to-
Cloudvara	www.cloudvara.com	99.99% Server Uptime	Application Hosting	way businesses work by
Solution	www.hitech-cloud.com	Voice Support	Application Hosting	in United States.We are the
Appointment-Plus	plus.com		Scheduling	worldwide leader in online
BookSteam	www.booksteam.com	Online Payments	Scheduling	leading Online Appointment
Timely	www.gettimely.com/	Scheduling	Scheduling	appointment scheduling for
BookedIN	bookedin.com	scheduling	Scheduling	phone? Tired of endless
OnShift	www.onshift.com	Software	Scheduling	software and proactive
SnapAppointments	snapappointments.com	Scheduling	Scheduling	winning online appointment
TimeTap	www.Timetap.com	Multiple Locations	Scheduling	scheduling platform built for
TrekkSoft	www.trekksoft.com	website	Scheduling	with an integrated booking
tuOtempO	www.tuotempo.com	Booking	Scheduling	PATIENT CRM that
Zirtual	www.zirtual.com	executive assistants	Scheduling	U.S. based, virtual assistants,

Fig. 1. List of some cloud service reviews

the fact that it only offers service name and URL [7]. To provide a more meaningful data set, another research in this area focuses on developing a web scraping tool to harvest SaaS offers. However, the data sets mentioned above contains only service name and service URL.

To address the need identified above, we propose to develop a new web scraping tool in this paper that can efficiently harvest cloud services from the cloud service marketplace to create a central cloud services registry. The new approach seeks to take advantage of the flexibility of the data structure displayed on web pages, enabling tool users to control the meaning, content and multi-level structure of the harvested data set. The tool also has the capacity for collecting data from future cloud service web portals using the same methodology, providing the registry with the potential to contain all the cloud services available on the web.

3 Results and Analysis

The primary objective of our work is to provide a cloud service data set to cloud consumers, comprised of IaaS, PaaS and SaaS. The significant level of information will enable users to use the dataset for multiple purposes; For instance, based on the reviews of consumers of cloud services providers, an organization can compare and decide the most suitable services for their functional needs. This study has collected information about 6000 services and 5000 service providers reviews from the Serchen Marketplace. The result shows that there are many cloud services providers offering a huge amount of services for specific areas such as dedicated hosting and web hosting. In contrast, there are fewer cloud services offered in some areas. For example, the web hosting category has the highest number of cloud services at 1500, where as there are zero cloud services offering Anti-Money Laundering Software. The review data set (consumer's comments) in this study is a valuable resource for potential future consumers to choose the service that meets their needs; it is also beneficial for service providers, to evaluate their customer reviews compared with those of their competitors.

4 Conclusion and Future Work

In this paper, we propose a central registry which includes a data set of real cloud services offerings from the web portal. To achieve our objective, we developed a harvesting tool which harvests cloud services information from the web. The data set holds thousands of cloud service records for SaaS, PaaS and IaaS categories, including 6000 cloud services and 5000 customer reviews. The data set is of benefit to cloud customers, service providers and the research community in the areas of cloud service discovery and selection. In future work, we plan to harvest more web portals and develop a comprehensive registry of cloud service listings.

References

1. serchen.com: Compare the best business software cloud services (2017). http://www.serchen.com/
2. Afify, Y.M., Moawad, I.F., Badr, N., Tolba, M.F.: A semantic-based software-as-a-service (saas) discovery and selection system. In: 2013 8th International Conference on Computer Engineering and Systems (ICCES), pp. 57–63. IEEE (2013)
3. Alfazi, A., Sheng, Q.Z., Qin, Y., Noor, T.H.: Ontology-based automatic cloud service categorization for enhancing cloud service discovery. In: 2015 IEEE 19th International on Enterprise Distributed Object Computing Conference (EDOC), pp. 151–158. IEEE (2015)
4. Alfazi, A., Sheng, Q.Z., Zhang, W.E., Yao, L., Noor, T.H.: Identification as a service: large-scale cloud service discovery over the world wide web. In: 2016 IEEE International Congress on Big Data (BigData Congress), pp. 485–492. IEEE (2016)
5. Alkalbani, A.M., Ghamry, A.M., Hussain, F.K., Hussain, O.K.: Sentiment analysis and classification for software as a service reviews. In: 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), pp. 53–58, March 2016
6. Alkalbani, A.M., Ghamry, A.M., Hussain, F.K., Hussain, O.K.: Blue pages: software as a service data set. In: 2015 10th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA), pp. 269–274. IEEE (2015)
7. Alkalbani, A.M., Hussain, F.K.: A comparative study and future research directions in cloud service discovery. In: 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), pp. 1049–1056. IEEE (2016)
8. Gong, S., Sim, K.M.: Cb-cloudle and cloud crawlers. In: 2014 5th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 9–12. IEEE (2014)
9. Han, T., Sim, K.M.: An ontology-enhanced cloud service discovery system. Proc. Int. MultiConf. Eng. Comput. Sci. **1**, 17–19 (2010)
10. Kang, J., Sim, K.M.: Towards agents and ontology for cloud service discovery. In: 2011 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), pp. 483–490. IEEE (2011)
11. Mell, P., Grance, T., et al.: The nist definition of cloud computing (2011)
12. Cisco Visual Networking. Cisco Global Cloud Index: forecast and methodology, 2014–2019. (white paper) (2013)
13. Noor, T.H., Sheng, Q.Z., Alfazi, A., Ngu, A.H., Law, J.: Csce: a crawler engine for cloud services discovery on the world wide web. In: 2013 IEEE 20th International Conference on Web Services (ICWS), pp. 443–450. IEEE (2013)
14. Sun, L., Dong, H., Hussain, F.K., Hussain, O.K., Chang, E.: Cloud service selection: state-of-the-art and future research directions. J. Netw. Comput. Appl. **45**, 134–150 (2014)
15. Tahamtan, A., Beheshti, S.A., Anjomshoaa, A., Tjoa, A.M.: A cloud repository and discovery framework based on a unified business and cloud service ontology. In: 2012 IEEE Eighth World Congress on Services (SERVICES), pp. 203–210. IEEE (2012)