Harvesting multiple resources for Software as Service offers: a big data study

Asma Musabah Alkalbani*, Ahmed Mohamed Ghamry**, Farookh Khadeer Hussain*, and Omar Khadeer Hussain**

*Decision Support and e-Service Intelligence Lab,
Center of Quantum Computation and Intelligent Systems,
School of Software, University of Technology, Sydney, NSW 2007, Australia
Asma.M.Alkalbani@student.uts.edu.au,Farookh.Hussain@uts.edu.au
**School of Business,
University of New South Wales Canberra (UNSW Canberra),
Australian Defence Force Academy,
Canberra, ACT, 2602,
a.ghamry@unsw.edu.au,O.Hussain@adfa.edu.au

Abstract. Currently, the World Wide Web (WWW) is the primary resource for cloud services information, including offers and providers. Cloud applications (Software as a Service), such as Google App, are one of the most popular and commonly used types of cloud services. Having access to a large amount of information on SaaS offers is critical for the potential cloud client to select and purchase an appropriate service. Web harvesting has become a primary tool for discovering knowledge from the Web source. This paper describes the design and development of Web scraper to collect information on SaaS offers from target Digital cloud services advertisement portals, namely www.getApp.com, and www.cloudreviews.com. The collected data were used to establish two datasets: a SaaS provider's dataset and a SaaS reviews/feedback dataset. Further, we applied sentiment analysis on the reviews dataset to establish a third dataset called the SaaS sentiment polarity dataset. The significance of this study is that the first work focuses on Web harvesting for cloud computing domain, and it also establishes the first SaaS services datasets. Furthermore, we present statistical data that can be helpful to determine the current status of SaaS services and the number of services offered on the Web. In our conclusion, we provide further insight into improving Web scraping for SaaS service information. Our datasets are available online through www.bluepagesdataset.com

Keywords: Software as a Service, Service offer, Web harvesting, SaaS dataset

1 Introduction

Over the past few years, with the continuous and rapid growth of cloud computing technologies, Software-as-a-Service (SaaS) has become one of the world's

largest digital business industries. SaaS shows a hybrid year-to-year increase, and several reports indicate that SaaS is becoming widely accepted. For instance, Gartner stated that in 2014, SaaS achieved 48.8 billion dollars in revenue [1]. One prediction indicates that by 2020, sales of SaaS will be more than 132 billion dollars. Another prediction by the International Data Corporation (IDC) is that by 2017, the SaaS market will be worth $107 billion, more than twice as much as its 2013 estimate of $47.4 billion [2]. Hence, the SaaS market has become highly competitive for SaaS service providers all over the world.

The Internet is the primary resource and the only distribution channel for the SaaS industry, transforming the Internet into a global SaaS marketplace. For example, there is a vast amount of SaaS information provided by SaaS-related websites containing SaaS offers, SaaS prices and details on SaaS providers. In addition, there has been a growth in Web-based portals, such as cloud reviews [3] and getApp [4], which provide a list of service offers collected from multiple sources on the Web.

Generally, publicly available search engines, such as Google, Yahoo, and Bing, are used to search for SaaS service offers on the World Wide Web (WWW). The results of these search engines show the potential that exists for extracting SaaS offers from the Web. However, the key issues lie within the quality of the results, as these search engines do not recognize SaaS offers. Usually, the obtained results comprise both relevant and irrelevant web sources. Consequently, accessing information on SaaS offers remains a problem as there is a lack of an available and efficient searching and information retrieval tools to find SaaS offers on the Web.

Therefore, our research concept is to utilize the existing content and structures of SaaS offers used from multiple sources to investigate SaaS offers on the Web in order to provide a complete view of the available SaaS offers. In other words, this study attempts to discover the SaaS offers which are available on the Web today.

The majority of the research to date, however, has focused on enhance SaaS discovery by using semantic technology to enhance Cloud information retrieval such as in [5][6]. The results of these research studies have shown the potential that lies in using semantic technologies to enhance data retrieval results from existing text-based search engines. Research by [7] proposed semantic information filtering of a search engine's results. Basically, the filter identifies the similarities between the cloud ontology concepts and the search engine's results, and then based on the specific threshold, it identifies if the information retrieved is relevant or irrelevant to the cloud domain. The cloud domain ontology comprises 424 concepts, which present the information.The semantic filter has been evaluated using virtual websites with up to 15700 web pages, including irrelevant and relevant cloud service providers' virtual sites.

In 2013 Noor et al. [8] consulted a cloud ontology to crawl Web-based resources, and then stored the crawling result in a local repository in order to obtain a cloud dataset. This study is considered to be the first effort toward obtaining a cloud dataset, but the dataset has several limitations, including a

lack of primary service information, such as service name and service URLs, and the data values do not have the semantic meaning associated with them. Even though there have been numerous efforts to enhance the discovery of cloud services over the Web, the main limitation of these studies is that they fail to address the issue of investigating and discovering cloud service offers across multiple Web resources.

Therefore, to address this issue, this work introduces a framework for harvesting multiple SaaS offer resources to build the first SaaS repository. In this paper, we propose a SaaS Web scraper which crawls across several publicly available web portals to establish SaaS datasets. Our proposed method shows better results compared with existing approaches in regard to the provision of details on how many SaaS services are available today on the Web. We successfully collected around 5294 existing SaaS offers accessible on the Web today. Our dataset on SaaS offers comprises the main attributes that are needed for service selection. Moreover, this dataset assists in drawing a statistical distribution of SaaS offers, therefore providing accurate conclusions.

Web harvesting (Web scraping) is a computer technique to extract information and data from Web sources [9] In other words, it is the transformation of unstructured data (HTML format) into structured data, also called Web data extraction. Although there has been much research on the subject of Web harvesting, no previous study so far has used the Web harvesting technique to investigate, collect, and gather information about cloud computing from the Web. In this work, we apply the Web harvesting task that targets Web sources to extract data about SaaS offers and SaaS consumers' reviews. This paper aims to obtain SaaS data from multiple sources. For this study, the target is restricted: to extract SaaS offers and consumers' reviews and feedback on the services from multiple sources on the WWW.

In this work, we establish three datasets: a SaaS offers dataset, a SaaS reviews dataset, and a SaaS polarity dataset, which can be potentially used as a resource for SaaS service discovery, selection, and composition. Moreover, this data could be used for SaaS knowledge discovery which plays a vital role in the construction of a SaaS knowledge base. This research makes the following contributions: we examine the potential of using the Web harvesting technique to extract information on SaaS services offers and SaaS consumers' reviews from multiple sources on the WWW; we introduce the notion of a SaaS services dataset to collect SaaS service offers that can be potentially used as a base for SaaS service discovery, selection, and composition; a SaaS dataset can also be used for SaaS knowledge discovery in order to construct a SaaS knowledge base; by continually scraping the existing SaaS service sources available on the Web, the dataset is capable of providing up-to-date data on SaaS services, hence this dataset is effective for service discovery, we collect and analyse the results and present various statistics including how many SaaS are accessible and what different categories of SaaS are accessible and we apply sentiment analysis and run several machine learning experiments on the SaaS reviews dataset containing the SaaS polarity dataset that can be accessed on the Web today.

To the best of our knowledge, this is the first study to do the following :

1. to investigate the Web to discover the amount of SaaS offers available today on the Web;
2. to establish a SaaS offers dataset;
3. to collect SaaS consumers' reviews to analyse and investigate overall satisfaction of SaaS consumers. Such analysis could provide useful information to improve the quality of SaaS provided;
4. to establish a SaaS polarity dataset that can possibly be used for applying some machine learning prediction techniques and for deep learning as well;
5. to provide ongoing research which aims at establishing the largest cloud services dataset and knowledge base.

The rest of this paper is organized as follows: section 2 describes some of the related work; section 3 discusses resources to find SaaS service offers; section 4 describes the architecture of the dataset; section 5 describes the methodology of our research; section 6 discusses the results and the evaluation of harvesting SaaS web sources; section 7 describes some of the challenges in the discovery of services; section 8 discussed conclusion and future work.

2 Related work

Recently, researchers have shown an increased interest in the discovery of cloud service issues, whereas previously many had focused on the discovery of cloud services on the Web using semantic technologies. A considerable amount of literature has been published on building a cloud ontology to enhance the dynamic discovery of cloud services over the WWW. A recent study by Afify et al. [5] developed a system for cloud service discovery containing a business ontology that assists service registry, service discovery, filtering and ranking the final result. This study does not support the dynamic discovery of cloud services, hence the information of the service offers need to be provided manually.

Research by Magesh et al. [10] proposed a semantic description for cloud service offers including service name, service level of agreement, service price, and service features. The study suggested representing each cloud offer as a single ontology and then combining all of them to construct a global ontology. The constructed global ontology has 64 entities and 128 properties. Unfortunately, this study neglects the need for quality of service information and rating attributes in selecting the services. In another effort by Kang et al[11] a cloud service ontology was introduced to enhance the dynamic discovery of cloud services on the Web. This research used several reasoning methods to find semantic similarities between the user's request and the search engine results. The selected services are ranked, based on the price in the time slots that were determined by the consumers.

A different approach taken by Tahamtan et al.[12] is to assist a business organization to find an appropriate cloud business service, and a cloud business functions ontology was proposed to achieve this goal. The ontology includes

most of the business function concepts and classifications outlined in [13]. In order to locate the right service with the right provider, the ontology is designed to map between the cloud service concepts and business concepts. In addition, the ontology includes some other important service attributes, such as service characteristics and service delivery model. Unfortunately, as with other existing work, their work, too, fails to account for the model's QoS parameters.

Although semantic searching methods may partially support the discovery of cloud services, they do not provide users with efficient ways to find proper services. Additionally, a scarcity of contributions in the current literature to determining the current status and distribution of cloud services. Very little research has been conducted on investigating cloud services on the Web. In [8] the authors' details on cloud services were collected throughout the Web by crawling Web sources. However, the study does not provide a complete view of the cloud services on the Web, and also the dataset provided in this study lacks primary service information, such as service name and service URL, and the data values do not have the semantic meaning associated with it. Therefore, it may provide inaccurate or misleading conclusions.

A recent study by Alkalbani et al. [14] proposed establishing a central repository for SaaS services. The study makes use of an open source, namely the Nutch-Hadoop crawler, to crawl details on SaaS offer from the Web and then stores the result in a local repository. The key shortcoming of this research is that the study only provides a service URL and service name.

3 SaaS services offer resources

Finding information on SaaS service offers is not an easy task, especially since SaaS offers do not have standards to support service publishing, service description, discovery of service providers and their services offered, as is the case for Web services. For Web services, a service registry has been developed which plays a vital role as a publicly available, central access point to describe and publish Web services using semantic annotations. However, generally speaking, cloud service discovery is strictly tied to publicly available general search engines, such as Google, Yahoo, and Bing. These engines search text to find information related to service offers. However, they usually retrieve both relevant and irrelevant information. The following briefly describes the range of possible Web resources for finding SaaS services offers on the Web.

3.1 Cloud Service Portals or Directories

Web-based service directories or portals, such as getApp, cloudreviews, and others, is one possible method for finding SaaS services. The majority of services listed in these portals and directories have been collected from different cloud providers. Capturing SaaS services from these portals requires access to their repositories which are not publicly accessible. Another way to capture service offer data is by building a custom web scraper designed to capture and collect

the service offer data from each portal independently, which is the focus of our re- search.

4 SaaS Datasets

Service offer information is a business structure for publishing service and business information which should be carefully considered when selecting the service. In the case of SaaS service offers, the service offer is simply meta-information or in other words, a HTML document. Therefore, for the purpose of this study, we target only meta-information for each service offer, which usually includes: service name, service description, service provider, service URL, service rating, service price, etc. In addition, this research considers collecting the services' rating as a part of the service offer that needs to be known when making a service selection decision. Also, this study considers collecting service reviews and feedback which could provide a useful summary about SaaS users' satisfaction. Thus, this research establishes three SaaS datasets: (1) SaaS offers dataset, (2) SaaS reviews dataset, and (3) SaaS sentiment polarity dataset. The next section details our procedure to construct these datasets.

5 Methodology

The mechanism implemented to achieve our research objectives is as follows: (1) defining the accessible Web-based sources from which SaaS information can be obtained, including publicly available web portals such as getApp, (2) designing and building a Web scraper to automatically crawl and collect information about SaaS including SaaS offers and SaaS consumers' reviews/feedback, and finally storing the results in a local repository to establish SaaS datasets, which is explained in more detail in the results section. Our research framework, as shown in Figure 1, consists of the following stages:

- **Meta-collector:** To collect the meta-resources, first we download the Web source for each Web portal home page (HTML document), and then we extract the service offer links (URLs) from the home page sources. Then, we obtain the meta-information for each service link (URL).
- **Meta-validator:** We verify and validate the collected URLs and ensure that we retrieve all the offered URLs.
- **Meta-storage:** We store the meta-information on each URL (service offer page source) as a *"Meta-Source object"*.
- **Meta-parser:** In this step, we define the targeted information that needs to be harvested for all services including: service offer template (service name, service price, service provider), and service reviews.
- **Meta-database:** Finally, we store the extracted information for each service offer as a *"Service Object"*, and lastly we establish a meta-database which has around 5294 SaaS service offers.

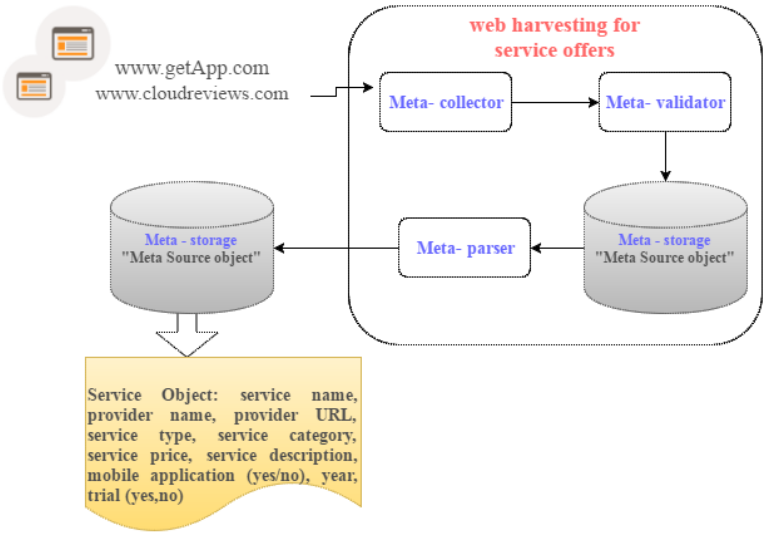

Fig. 1. Research Framework

6 Results and Statistics

In this section, we present the results and details of the SaaS datasets and the statistics on the harvested data. The harvested data is distributed among two datasets as follows:

1 **SaaS offers dataset:** contains the primary information on SaaS offers harvested from the targeted Web resources. The SaaS offers data collection shows that each offer comprises the following attributes: service name, service provider name, provider URL, service rate, service description, year founded, mobile application (yes/no), starting price, service type, service category, free trial (yes/no).

The data harvesting took place between February 2015 and August 2015, and the total number of SaaS offers harvested was 5294. Our constructed dataset illustrates that the majority of harvested offers are from getApp, which provides around 5146 service offers. Table 1 presents details on the SaaS dataset with respect to the resources used to collect the SaaS offers. Also, the results of this study, as illustrated in Table 2, indicate that the total number of unique SaaS offers is 3184, and surprisingly, we found that around 2110 are duplicated. This result may be explained by the fact that some services be- long to more than one category (Table 2 shows that around 1512 ser- vice offers belong to more than one category). Also, the study found that, so far, the maximum number of service offers per category is five. In addition, Table 3 shows that there is no feedback data recorded in our constructed dataset from www.cloudreviews.com , whereas around 6343

Table 1. Summary of harvested SaaS offers per web resources

	www.getApp.com	www.cloudreviews.com
Harvested offers	5146	148
Execution time	3 minutes	1 minutes
Total harvested offers	**5294**	

Table 2. Summary of Unique/Duplicated harvested SaaS offers per web resource

	www.getApp.com	www.cloudreviews.com	Total
Unique offers	3038	146	3184
Duplicated offers	2108	2	2110
Total harvested offers			5294

were collected from www.getApp.com. This table shows the distribution of service re- views by service category. The data in the table indicates that the operations management application received the highest feedback from users, fol- lowed by customer management. Moreover, as can be seen from Figure 1, only 14% of offers provide a URL. The data collected shows that the majority of service offers published do not have a provider link, which accounts for 86% of the constructed dataset.

2 **SaaS reviews/feedbacks dataset:** contains the collected reviews/feedback that have been made by services' users.

3 **SaaS sentiment polarity dataset:**
Sentiment analysis was applied on the Saas reviews/feedback dataset to determine the tone of each SaaS post/review as being either positive, negative, or neutral. As a result of this analysis, we have another dataset, namely "the SaaS sentiment polarity dataset". The results obtained from the sentiment analysis on the SaaS reviews are shown in Table 4. More details on this analysis can be found in [15].This dataset is a very useful for training machine learning algorithms and for fur- ther study. To conclude, all these can be accessed online through www.bluepagesdataset.com as the first publicly available datasets for SaaS offer information, SaaS reviews and feedback, and the SaaS sentiment polarity dataset.

7 SaaS harvesting challenges

At any point, Web portals may update service offers, therefore the Web scraper needs to be able to update or revisit Web resources to identify the changes that have taken place and update the downloaded data. Additionally, the number of SaaS services increases as well as the number of web portals, therefore it is an ongoing process to keep data up-to-date. From our experience with Web harvesting in this work, to achieve database availability, the challenges are:

Table 3. Number of reviews/feedback per service type

	Service types	Reviews/Feedbacks
T1	Finance & accounting	1055
T2	Marketing	797
T3	Communications applications	389
T4	collaboration applications	1448
T5	Sales	1263
T6	Project management	1203
T7	Customer management	1655
T8	IT management	997
T9	Customer service & support	763
T10	Operations management	1822
T11	Business Intelligence & Analytics	208
T12	HR & Employee management	931

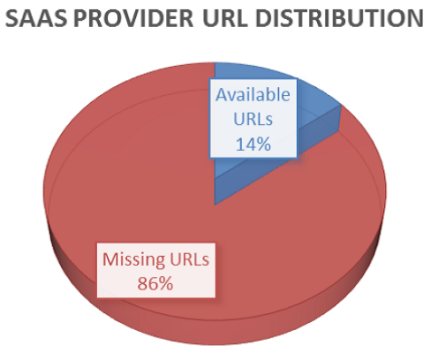

Fig. 2. SaaS provider URL distributions

Table 4. Summary of sentiment analysis

Polarity of reviews	Number of Reviewers
Positive	2487
Neutral	1312
Negative	201
Total	4000

– the customized Web scraper needs to monitor sites/pages for changes and up-dates.
– adding more repositories to our dataset requires designing and adding more code to our customized scraper.
– building a dynamic scraper that can handle the addition of more repositories and page source changes.

8 Conclusions

Finding and selecting relevant Software as a Service (SaaS) offers is mainly done manually by scanning through a number of suggestions from general search engines, such as Google or Bing. The dynamic discovery of SaaS service offers is necessary, especially when the number of services on the Web and the number of Web portals continues to significantly increase. Our study presented the implementation of a "Web scraper" to discover and investigate the number of SaaS offers available on the Web. We harvested SaaS offers from targeted web portals. The results provide an overview of the current status of SaaS offers and knowledge on the Web. An interesting result shows that some SaaS services are categorized according to business function, and some services belong to more than one category. For future work, our objective is to construct a large SaaS knowledge base and we will continue harvesting more Web sources as well as develop effective tools to dynamically update our datasets.

References

1. Gartner says worldwide it spending is forecast to grow 0.6 percent in 2016, `http://www.gartner.com/newsroom/id/3186517`, (Accessed on 08/08/2016).
2. Worldwide saas and cloud software 20152019 forecast and 2014 vendor shares - 257397, `https://www.idc.com/getdoc.jsp?containerId=257397`, (Accessed on 08/08/2016).
3. Cloud reviews — cloud hosting — managed cloud — cloud storage & apps, `http://www.cloudreviews.com/`, (Accessed on 08/08/2016).
4. Business software reviews, saas & cloud applications directory — getapp, `https://www.getapp.com/`, (Accessed on 08/08/2016).
5. Y. Afify, I. Moawad, N. Badr, M. Tolba, A semantic-based software-as-a-service (saas) discovery and selection system, in: Computer Engineering & Systems (ICCES), 2013 8th International Conference on, IEEE, 2013, pp. 57–63.
6. T. Han, K. M. Sim, An ontology-enhanced cloud service discovery system, in: Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1, 2010, pp. 17–19.
7. J. Kang, K. M. Sim, Cloudle: an agent-based cloud search engine that consults a cloud ontology, in: Proc. Intl Conf. Cloud Computing and Virtualization, Citeseer, 2010, pp. 312–318.
8. T. H. Noor, Q. Z. Sheng, A. Alfazi, A. H. Ngu, J. Law, Csce: a crawler engine for cloud services discovery on the world wide web, in: Web Services (ICWS), 2013 IEEE 20th International Conference on, IEEE, 2013, pp. 443–450.
9. G. Weikum, M. Theobald, From information to knowledge: harvesting entities and relationships from web sources, in: Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM, 2010, pp. 65–76.
10. M. Vasudevan, P. Haleema, N. C. S. Iyengar, Semantic discovery of cloud service catalog published over resource description framework, International Journal of Grid and Distributed Computing 7 (6) (2014) 211–220.
11. J. Kang, K. M. Sim, Cloudle: a multi-criteria cloud service search engine, in: Services Computing Conference (APSCC), 2010 IEEE Asia-Pacific, IEEE, 2010, pp. 339–346.

12. A. Tahamtan, S. A. Beheshti, A. Anjomshoaa, A. M. Tjoa, A cloud repository and discovery framework based on a unified business and cloud service ontology, in: 2012 IEEE Eighth World Congress on Services, IEEE, 2012, pp. 203–210.
13. M. Kerrigan, A. Mocan, M. Tanler, D. Fensel, The web service modeling toolkit-an integrated development environment for semantic web services, in: European Semantic Web Conference, Springer, 2007, pp. 789–798.
14. A. Alkalbani, A. Shenoy, F. K. Hussain, O. K. Hussain, Y. Xiang, Design and implementation of the hadoop-based crawler for saas service discovery, in: 2015 IEEE 29th International Conference on Advanced Information Networking and Applications, IEEE, 2015, pp. 785–790.
15. A. M. Alkalbani, A. M. Ghamry, F. K. Hussain, O. K. Hussain, Sentiment analysis and classification for software as a service reviews, in: 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), IEEE, 2016, pp. 53–58.