ServiceMarq: Extracting Service Contributions from Call for Papers

Shi Tian
shitian@u.nus.edu
National University of Singapore
Singapore

Abhinav Ramesh Kashyap
abhinav@comp.nus.edu.sg
National University of Singapore
Singapore

Min-Yen Kan
knmnyn@nus.edu.sg
National University of Singapore
Singapore

Abstract

In an era, where large numbers of academic research papers are submitted to conferences and journals, the voluntary services of academicians to manage them, is indispensable. The call for contributions of research papers – through an e-mail or as a webpage, not only solicits research works from scientists, but also lists the names of the researchers and their roles in managing the conference. Tracking such information which showcases the researchers' leadership qualities is becoming increasingly important. Here we present *ServiceMarq* - a system which proactively tracks service contributions to conferences. It performs focused crawling for website-based call for papers, and integrates archival and natural language processing libraries to achieve both high precision and recall in extracting information. Our results indicate that aggregated service contribution gives an alternative but correlated picture of institutional quality compared against standard bibliometrics. In addition, we have developed a proof of concept website to track service contributions and is available at https://cfp-mining-fe.herokuapp.com and our github repo is available at https://github.com/shitian007/cfp-mining

CCS Concepts: • **Applied computing** → **Digital libraries and archives**; • **Information systems** → **Digital libraries and archives**; • **Computing methodologies** → **Information extraction**.

ACM Reference Format:
Shi Tian, Abhinav Ramesh Kashyap, and Min-Yen Kan. 2020. *ServiceMarq*: Extracting Service Contributions from Call for Papers. In *ACM Symposium on Document Engineering 2020 (DocEng '20), September 29-October 2, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3395027.3419596

1 Introduction

Academic journals and conferences announce *call for papers* (hereafter, CFPs) to solicit work for prospective publication. These calls are often disseminated by email, but especially in the case of conferences, take the form of dedicated CFP sub-pages within a conference website. CFP websites often acknowledge their committee members' efforts by listing names, affiliations and roles they play in the conference. These quality marks also help potential contributors judge the quality and impact of a venue. Quality venues attract quality scholars to not only contribute as authors, but also to serve in reviewing and organizing roles. Thus, being invited and serving as a committee member for a prominent conference is a marker of scholarly success.

While a scholar's reputation is traditionally measured using bibliometrics (i.e., citation count and H-index), service contributions become more important for middle and senior scholars to demonstrate their leadership within a community. This "serve or be skipped over" paradigm motivates the need to track service contributions, without which communities would be unable to execute meetings and peer review. For certain disciplines such as computer science, conferences serve as a major means of disseminating high impact work. Thus, service to conferences is a critical component of service that needs to be captured.

Information about the service contribution of researchers are widely dispersed and in many forms: in journal websites' editorial board listings, conference CFP websites, and on individual faculty pages. Currently, scholars have to argue their case for demonstrating service contributions for promotion, without help from any metric or centralised source. To the best of our knowledge, there is no system that extracts information from website CFPs and create a comprehensive database containing such information at scale.

We propose *ServiceMarq*, an automated system for the extraction of service contributions from CFP websites that addresses part of this problem, by automatically crawling relevant pages of a conference CFP website and extracting individuals' affiliation and role in the conference. It consolidates information over many CFPs, providing a holistic view of service contributions for a scholar. *ServiceMarq*'s aggregated scholar profile provides the magnitude and spread of contributions by different scholars and institutions (by aggregating scholars with the same affiliation).

DocEng '20, September 29-October 2, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-8000-3/20/09…$15.00
https://doi.org/10.1145/3395027.3419596

2 Related Work

Previous work that mined conference CFPs extracted only general information such as conference location and deadlines from email CFPs. These include formalizing the extraction as a token-level Begin, Interior, Other (BIO)-tagging task using a small set of email CFPs to train Conditional Random Fields [10], or employing the use of regular expression and manually-engineered patterns [1], or via a natural language processing toolkit such as GATE[1] [4] for Named Entity Recognition (NER). Li *et al.*'s system [5] extracted researchers and their affiliations from email CFPs, but disregards researcher roles, is probably the closest in spirit to our work, achieving a high level of performance of $0.89F_1$, but on a small dataset of 100 CFP emails.

Separately, databases documenting service contributions have also been proposed, focusing on the crowd-sourcing of information. OpenResearch is a semantic Wiki documenting conference service information mainly to inform researchers on the suitability of publication venues [11], but lacks sufficient coverage as it relies on user contributions, unlike proactive crawling which we do. Publons[2] aggregates peer review records from its source publisher datastream, relying on user-provided data. Neither provides an overview of the scholar service contribution landscape.

Literature on Information Extraction (IE) on semi-structured text is relevant to our approach. Prior work on Web text focused on structured web data records [6, 12]. Related work on extraction from semi-structured scholarly publications leveraged similar layout in academic publications [2] to extract authors and their affiliations. However, the assumption of a high degree of duplication in structured records of these proposed approaches make them less suited to our task of extracting information from webpages where <author, affiliation, role> tuples often manifest in *ad hoc* structures.

3 System Design

ServiceMarq tracks the service contributions of scholars, by extracting information from CFP webpages. In contrast to the prior work on email CFPs, we target CFP websites which present a more complete listing of service roles – sometimes listing entire programme committees – and because their semi-structured nature facilitates extraction of such information. We extract the names of relevant *Scholars*, their affiliated *Affiliations* and their *Roles* in managing the conference.

3.1 WikiCFP Focused Crawling with WaybackMachine Archive Fallback

We begin by conducting a focused crawl of relevant CFP webpages from WikiCFP[3], a user-contributed database of Computer Science conference CFPs. During the crawl, to overcome the issues of 1) dead links to sites of older conference iterations, and 2) duplicated URLs for multiple iterations of the same conference, we utilise *WaybackMachine Archive* (a digital archive of the Web over different timestamps) to retrieve legacy snapshots as a fallback. This improves coverage of conference CFPs by a significant 27.8%.

3.2 Word Embedding based Bi-LSTM Line Classification

Given the raw, transmitted Hypertext Markup (HTML) of each retrieved conference CFP page from the crawl, we make two important observations: 1) a majority of content on each webpage is irrelevant; 2) most of the relevant information exist as atomic items on a single, rendered HTML line in the form of either a *Scholar*, *Affiliation* or *Role Label*. To determine the relevant portions of the webpage, we first break the webpage up into individual lines of text (i.e., the text of an innermost HTML tag or those defined by < br > tags). The resulting lines are then attributed to a class of *Scholar*, *Affiliation*, *Role-Label*, *Complex* (lines having mixture of relevant data, which we process downstream.), or *Irrelevant*. We perform this as a supervised line classification task with a one-layer bi-directional long short-term memory (bi-LSTM) encoder followed by a one-layer feed-forward neural network. Each input token to the encoding layer is represented by the GloVe[8] embedding of the token, together with the additional features of the individual line's HTML tag(s) and the total number of tokens of the line[4]. The bi-LSTM line classification model is trained against 12,000 manually labelled lines, where two independent annotators achieve a Kappa of 0.83. Our model achieves an overall F_1 of 97%.

3.3 Scholar and Affiliation Entity Extraction and Role Attribution

For the extraction of the named entities of *Scholar* names and *Affiliation* entities, we use the flairNLP[5] library. In our manual verification of 500 individual extractions from the system, flairNLP achieves a satisfactory level of accuracy of *Scholar*s and *Affiliation*s at a precision of 98.8% and 94.6%, respectively. Given a majority of misclassifications to be of longer spans containing concatenated *Affiliations* and *Scholars*, we further improve the performance of extraction by training our own character-level bidirectional LSTM-CRF [3] with our own generated IOB[9] dataset. With the addition of the bidirectional LSTM-CRF, we achieve a 3.8% improvement in the recall of *Scholar* entities and a 1.2% improvement in the recall of *Affiliation* entities.

We observe that contributions in our CFPs mostly follow the pattern of a lead *Role-Label* followed by one or multiple

[1] https://gate.ac.uk/digilibs.html

[2] https://publons.com/about/home/

[3] http://wikicfp.com; WikiCFP, established in 2007, contains information only for conferences after 2007.

[4] Thus the feature vector of each word w_i in line l is $feat(w_i, l) = emb(w_i) \oplus emb(l_{tag}) \oplus length(l)$.

[5] https://github.com/flairNLP/flair

lines of either *Scholar*s *Affiliations* and/or *Complex* lines. We thus partition each webpage into sections based on the extracted *Role-Labels* and attribute extracted *Scholars* to each corresponding *Role-Label* header. *Scholar*s and *Affiliation*s are also matched based on the proximity of their lines. We provide a more detailed evaluation in the extraction of these entity pairs in section 3.5.

3.4 Affiliation and Scholar Disambiguation integrating External APIs

We first perform the auxiliary task of disambiguating *Affiliation*s. Using a Term Frequency × Inverse Document Frequency (TF·IDF) vector representation for each extracted *Affiliation*, we conduct Hiearchical Agglomerative Clustering (HAC) with an empirically determined distance threshold. Individual clusters denote individual affiliations.

For disambiguating scholar names, we utilize external dblp[6] search API. Inputting extracted names and cleaned affiliations, we retrieve the closest possible match to aid in name disambiguation. Futhermore, the APIs allow us to import affiliations for Scholars which we were not able to extract affiliations for, and handle naming inconsistencies (i.e., flipped ordering of first/last names, and omitted middle names).

3.5 System Performance

Recall. With an entire crawl through WikiCFP, our system finds meaningful extractions from 75% of accessible conference CFPs. A large percentage of errors are caused by CFPs in PDF format (not HTML), which we currently do not handle. The resultant dataset of 6,504 conferences with its corresponding <*Scholar*, *Affiliation*, *Role-labels*> tuples is well over a magnitude larger than ones in prior work.

Precision. We randomly sample 200 tuples to manually verify against the conference CFP sites. While only indicative, this held-out verification nets an accuracy of 92% with 16 errors, attributed as 2, 6 and 8 errors in extracting *Scholar*, *Affiliation* and *Role-label*s, respectively. With performance above the 90% level and moderate coverage, we believe *ServiceMarq*'s extractions serve as a strong starting point for measuring service contributions in conferences.

4 Front End Extension

As an extension to the automated crawling of service contributions through CFPs, we build a simple interface to facilitate end-user access. We seek for this service, as a proof-of-concept, to firstly encourage the exploration of our data and publicize our system. Second, in acknowledging imperfections in our extractions, such a service allows us to crowdsource efforts in the rectification of erroneous or omitted extractions.

The front-end service of our system would allow the user easy access to fine-grained details of service contributions, at the level of individual *Scholar* data as well as aggregated information on *Affiliation*. Figure 1 presents a snapshot of our web front end showcasing a researcher's contributions to various conferences.

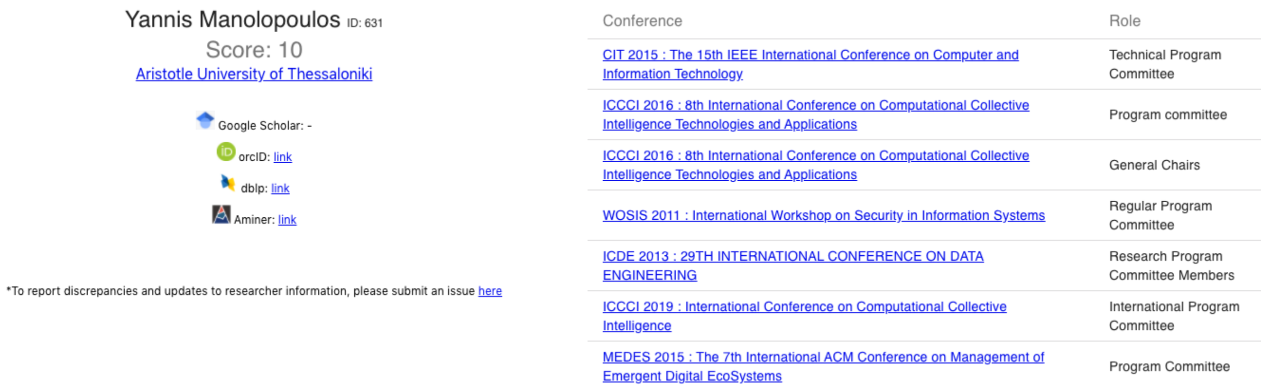
5 Discussion and Conclusion

Our hypothesis is that service contributions paint an alternative metric of research contribution somewhat orthogonal to traditional citation and bibliometric analysis. As we have demonstrated that *ServiceMarq* obtains satisfactory precision and recall, we wish to assess how well metrics compiled from its extractions correlate with existing metrics. More importantly, we seek to determine if our automated extraction system is able to paint a representative picture of the Computer Science service contributions landscape and provide a macro-analysis of the coverage of our extraction.

We first conduct a preliminary scoring of documented researchers in our database based on the number of service contributions, and aggregate their scores to obtain an overall scoring for different research institutions (denoted by SM-raw in Table 1). The reason for doing this is, 1) We acknowledge lapses in the comprehensiveness of the extracted data and hence accuracy of individual rankings might be affected, 2) The direct ranking of individual researchers remains controversial given our system is not yet mature and there is no established equivalent metric.

To improve on the comprehensiveness and reliability of our scoring, we perform an adaption of PageRank[7] in calculating *Scholar* service contributions. We denote *Scholar*s and *Conference*s as graph nodes and *Role-Label* as edges, and run the algorithm to convergence, before aggregating *Scholar* contributions for institutional scores (denoted as SM-pr). We observe a minor shuffling in the rankings of top institutions between SM-raw and SM-pr. We attribute the lack of a major change in ranking to the fact that prominent scholars in renowned institutions tend to produce research work in both quality and quantity. To quantify the improvement in representativeness we have gained in our calculation of SM-pr over SM-raw, we compare the ranking results of both methods to established ranking metrics. Table 1 shows a breakdown of the overlap between the top 20 research institutions in our system compared to those of other established holistic (QS, WorldU) and research only metrics (CSRanking). Here we safely assume a correlation between service contributions and overall research impact as measured by those on the established sites. We have also chosen to use overlap and not correlation statistics as a measurement of sameness since the established site rankings themselves have a high variability in terms of rankings.

[6] https://dblp.uni-trier.de/; dblp provides open bibliographic information on major computer science journals and proceedings



Yannis Manolopoulos ID: 631
Score: 10
Aristotle University of Thessaloniki

Google Scholar: -
orcID: link
dblp: link
Aminer: link

*To report discrepancies and updates to researcher information, please submit an issue here

Conference	Role
CIT 2015 : The 15th IEEE International Conference on Computer and Information Technology	Technical Program Committee
ICCCI 2016 : 8th International Conference on Computational Collective Intelligence Technologies and Applications	Program committee
ICCCI 2016 : 8th International Conference on Computational Collective Intelligence Technologies and Applications	General Chairs
WOSIS 2011 : International Workshop on Security in Information Systems	Regular Program Committee
ICDE 2013 : 29TH INTERNATIONAL CONFERENCE ON DATA ENGINEERING	Research Program Committee Members
ICCCI 2019 : International Conference on Computational Collective Intelligence	International Program Committee
MEDES 2015 : The 7th International ACM Conference on Management of Emergent Digital EcoSystems	Program Committee

Figure 1. An example of service contributions of an individual scholar as seen on our proof concept system.

	SM-raw	SM-pr	QS	WorldU	CSRankings
SM-raw	–	17	10	8	10
SM-pr	17	–	12	8	12
QS	10	12	–	11	12
WorldU	8	8	11	–	10
CSRankings	10	12	12	10	–
Average	9.33	10.7	11.5	10.5	11.0

Table 1. Counts and average (only with establised rankings like QS) of the number of common institutions, considering only the top 20 ranked by our and an established system.

We average over the number of overlaps only with established ranking metrics. We note that SM-pr achieves an overall higher degree of overlap as compared to SM-raw. In fact, SM-pr achieves an average on par with established metrics, indicative of a reasonable ranking. However, the rankings are notably different in a few aspects. First, the established sites base their rankings predominantly on traditional metrics such as publication data, while ours measures research impact through the alternative lens of service contributions of scholars within the institution. In addition, our system also assesses the impact of service from industrial research institutions that have remained largely undocumented. We omit them from our ranking as there is no comparable rankings for such institutions.

Finally, *ServiceMarq* is designed for production service: as an evolving system that ingests newly-crawled CFPs continuously. *ServiceMarq*'s data is provided in an easily accessibly manner within flat files in the public Github repo , which also provides transparent access to its data through suitable APIs and data dumps. Users can also update the data directly, by requesting to add or correct service records, available at https://cfp-mining-fe.herokuapp.com. In current work, we are also extending service coverage by extracting editorial boards from journal websites.

References

[1] Fábio L Correia, Rui FS Amaro, Luís Sarmento, and Rosaldo JF Rossetti. 2010. Allcall: An automated call for paper information extractor. In *5th Iberian Conference on Information Systems and Technologies*. IEEE, 1–4.

[2] Huy Hoang Nhat Do, Muthu Kumar Chandrasekaran, Philip S. Cho, and Min Yen Kan. 2013. Extracting and Matching Authors and Affiliations in Scholarly Documents. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (Indianapolis, Indiana, USA) *(JCDL '13)*. ACM, New York, NY, USA, 219–228. https://doi.org/10.1145/2467696.2467703

[3] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[4] Laurent Issertial and Hiroshi Tsuji. 2011. Information extraction and ontology model for a'call for paper'manager. In *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services*. ACM, 539–542.

[5] Xinyu Li, Roya Rastan, John Shepherd, and Hye Young Paik. 2013. Automatic affiliation extraction from calls-for-papers. In *Proceedings of the 2013 workshop on Automated knowledge base construction*. ACM, 97–102.

[6] Chunliang Lu, Lidong Bing, Wai Lam, Ki Chan, and Yuan Gu. 2013. Web entity detection for semi-structured text data records with unlabeled data. *International Journal of Computational Linguistics and Applications* 4, 2 (2013), 135–150.

[7] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The pagerank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[9] Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*. Springer, 157–176.

[10] Karl-Michael Schneider. 2006. Information extraction from calls for papers with conditional random fields and layout features. *Artificial Intelligence Review* 25, 1-2 (2006), 67–77.

[11] Sahar Vahdati, Natanael Arndt, Sören Auer, and Christoph Lange. 2016. OpenResearch: collaborative management of scholarly communication metadata. In *European Knowledge Acquisition Workshop*. Springer, 778–793.

[12] Yanhong Zhai and Bing Liu. 2005. Web data extraction based on partial tree alignment. 76–85. https://doi.org/10.1145/1060745.1060761