Semantics Driven Human-Machine Computation Framework for Linked Islamic Knowledge Engineering

Amna Basharat(✉)

Department of Computer Science, University of Georgia, Athens, GA 30602, USA
amnabash@uga.edu

Abstract. Formalized knowledge engineering activities including semantic annotation and linked data management tasks in specialized domains suffer from considerable knowledge acquisition bottleneck - owing to the lack of availability of experts and in-efficacy of computational approaches. Human Computation & Crowdsourcing (HC&C) methods successfully advocate leveraging the human processing power to solve problems that are still difficult to be solved computationally. Contextualized to the domain of Islamic Knowledge, my research investigates the synergistic interplay of these HC&C methods and the semantic web and will seek to devise a semantics driven human-machine computation framework for knowledge engineering in specialized and knowledge intensive domains. The overall objective is to augment the process of automated knowledge extraction and text mining methods using a hybrid approach for combining collective intelligence of the crowds with that of experts to facilitate activities in formalized knowledge engineering - thus overcoming the so-called knowledge acquisition bottleneck.

Keywords: Human computation · Semantic web · Task profiles · Crowdsourcing · Islamic knowledge · Quran · Hadith

1 Introduction

Challenges associated with large-scale adoption of semantic web technologies continue to confront the researchers in the field. Researchers have recognized the need for human intelligence in the process of semantic content creation and analytics, which forms the backbone of any semantic application [1,2]. Realizing the potential that human computation, collective intelligence and the fields of the like such as crowdsourcing and social computation have offered, semantic web researchers have effectively taken up the synergy to solve the bottlenecks of human experts and the needed human contribution in the semantic web development processes. This paper presents a novel contribution towards this intersection of the semantic web and human computation paradigm.

H. Sack et al. (Eds.): ESWC 2016, LNCS 9678, pp. 793–802, 2016.
DOI: 10.1007/978-3-319-34129-3_48

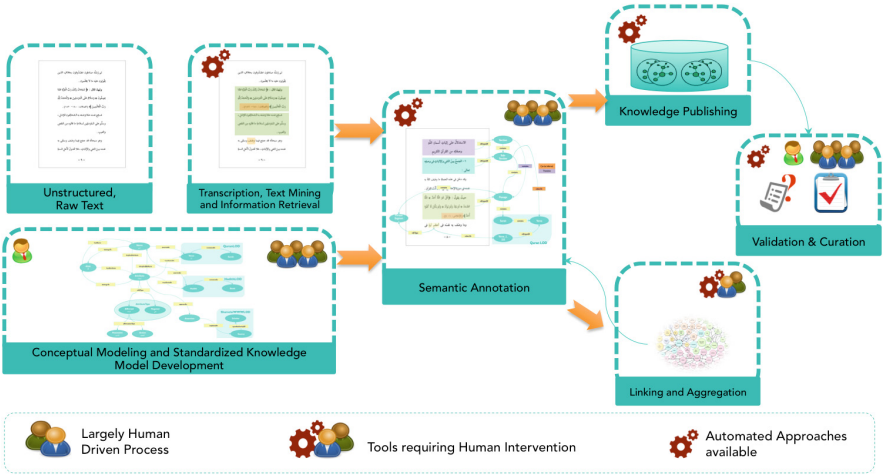

Fig. 1. Challenges in knowledge engineering processes

1.1 Background and Motivation

One of the major challenges hindering successful application of ontology-based approaches to data organization and integration in specialized domains is the so-called *'knowledge acquisition bottleneck'* [3]- that is, the large amount of time and money needed to develop and maintain the formal ontologies. This also includes ontology population and semantic annotation using well established vocabularies. My research primarily is motivated to overcome the inherent knowledge acquisition bottleneck in creating semantic content in semantic applications. We have established how this is particularly true for knowledge intensive domains such as the domain of Islamic Knowledge, which has failed to cache upon the promised potential of the semantic web and the linked data technology; standardized web-scale integration of the available knowledge resources is currently not facilitated at a large scale [4]. To date, only one dataset on the Linked Open Data (LOD) cloud in the domain exists [5].

To understand the knowledge acquisition bottleneck encountered in this domain (and others), consider the knowledge engineering processes illustrated in Fig. 1. Existing methods towards semantic annotation and linked knowledge generation are either (a) computationally driven, employing on text mining and information extraction methods or, (b) expert driven (such as conceptual modelling, annotation and validation). While the computational methods may assist in large-scale knowledge acquisition, however, the lack of formalized and agreed upon knowledge models and sensitivity of the knowledge at hand- primarily obtained from unstructured and multilingual data- makes the knowledge engineering process far from trivial. Islamic domain suffers a great deal from a lack of suitable training data and gold standards. This only adds to the challenge of ensuring the reliability and scalability of these methods. Expert driven methods involve subject specialists however,

these are often not scalable (time or cost). It is no wonder that the efforts towards the vision of standardization and formalization of Islamic Knowledge as proposed by [6] have remained futile.

1.2 Research Context and Problem Statement

To address these and similar challenges, researchers have recognized that the realization of the semantic and linked data technologies will require not only computation but also significant human contribution [1,2]. Humans are simply considered indispensable [7] for the semantic web to realize its full potential. Emerging research is advocating the use of Human Computation & Crowdsourcing (HC&C) methods to leverage human processing power to harness the collective intelligence and the wisdom of the crowds by engaging large number of online contributors to accomplish tasks that cannot yet be automated. There has been growing interest to use crowdsourcing methods to support the semantic web and linked open data research by providing means to efficiently create research relevant data and potentially solve the bottleneck of knowledge experts and annotators needed for the large-scale deployment of semantic web and linked data technologies. Recent framework called CrowdTruth proposed by [8] is a step forward that recognizes the challenges in gathering distributed data in crowdsourcing.

Within the realm of (semantic web based) knowledge engineering tasks, several levels of complexity may be encountered. Some tasks are simple, while others are more knowledge intensive. While some may reasonably be amenable to computational approaches, others need domain specific expert annotations and judgements. Therefore, not all tasks are fit for general purpose crowdsourcing. This is specifically true for the domain of islamic knowledge, owing to the specialized nature of the learning needs presented by diverse users, and heterogenous, multilingual knowledge resources. Therefore, I emphasize that specialized domains such as the one that forms the basis of my research, i.e. the Islamic knowledge domain, needs more than just faceless crowdsourcing. Emerging research paradigm recognizes this in the form of nichesourcing [9,10] or Expert-Sourcing [11], as a natural step in the evolution of the crowdsourcing to address the need of solving complex and knowledge intensive tasks, and as means to improve the quality of crowdsourced input.

Based on these ideas, I propose the design and development of a hybrid workflow framework that combines human-machine computation in a manner that allows the ability to compose tasks to a varying degree of granularity and delegating them to generic and specialized crowds depending on the needs of the task fulfillment requirements. In order to achieve this, I propose the utilization of semantics based representations of tasks, workflows and worker profiles.

2 State of the Art

The potential of HC&C has been leveraged by semantic web researches such as Noy et al. [12], Sarsua et al. [13] and others in attempting to solve the bottlenecks

of human experts and the needed human contribution in the semantic web development (ontology engineering) processes. Some early efforts that led to the evolution of this approach include Ontogame [14] and inPho [15]. Two major genres of research may be seen emerging in the last few years, in an attempt to bring human computation methods to the semantic web: (1) Mechanized Labour and (2) Games with a Purpose for the Semantic Web. Several recent research prototypes have attempted to use micro-task crowdsourcing for solving semantic web tasks e.g. ontology engineering [12,13] and linked data management [16]. Recent work by Hanika, Wohlgenannt and collegues [17,18] have attempted to provide tool support for integrating crowdsourcing into ontology engineering processes by providing a plugin for the popular ontology development tool Protégé. Other approaches such as [19] and [20] adopted the Lui von Ahn's "games with a purpose" [21] paradigm for creating the next generation of the semantic web. The idea is to tap on the wisdom of the crowds by providing motivation in terms of fun and intellectual challenge.

The evidence of semantic web techniques applied to improve the state of human computation systems is also emerging. Sabou et al. [22] propose the notion of hybrid genre workflows to overcome the limitations of traditional workflows in the crowdsourcing settings. Research also suggests the use of ontologies to improve the human-computation process [23], which provides the motivation for proposing a semantics driven model of human computation for this research.

3 Research Proposition

In this research, I seek to develop a semantics driven, generic and reusable human computation based framework for semantic annotation, knowledge acquisition and generation of Linked Islamic knowledge. The framework utilizes HC&C paradigm incorporated into hybrid knowledge engineering (ontology development and linked data management) workflows to produce semantics based multi-lingual, linked knowledge resources. A hybrid model of human computation, that leverages both generic workers and experts, is utilized to not only validate and verify the findings obtained through computational approaches such as text mining techniques, but also to perform higher level conceptual problem solving, integrative analysis and inter-linking of knowledge sources at hand. The results of the crowd contributions are used to create formalized, shared knowledge models and benchmarks for distributed knowledge sources and to meaningfully link them in primarily the Islamic Knowledge domain.

The proposed framework is validated through contextualized application to the domain of Islamic and Religious texts, where the overall vision is to provide means to enable efficient and reliable knowledge discovery for religious knowledge seekers for a more meaningful knowledge seeking and learning experience. Some of the key research questions are addressed as part of the process are as follows:

RQ-1: What is the amenability of crowdsourcing ontology engineering tasks in knowledge intensive domains? Can the tasks of thematic disambiguation, semantic annotation and thematic interlinking be reliably crowdsourced in the specialized and knowledge intensive domains such as the Islamic knowledge?

RQ-2: Is there significant performance gap between experts and crowd workers when performing such tasks?

RQ-3: Can the contributions from crowd workers and experts be reliably and efficiently combined for the purpose of knowledge engineering, using semantics driven, hybrid and iterative workflows? What methodological considerations would enable such an effective synergy?

4 Research Methodology and Approach

In order to realize the research vision, and to answer the research questions, I undertake the research methodology, a high level view of which is presented in Fig. 2.

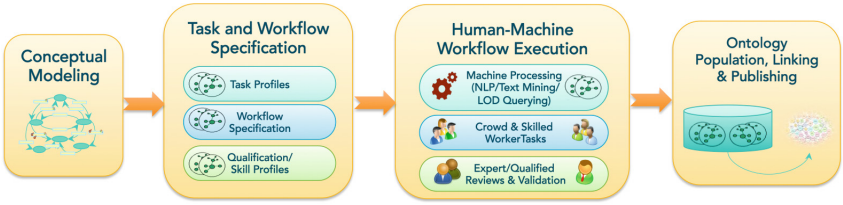

Fig. 2. High level constituent stages of framework design

Conceptual Modeling: The initial step in our approach consists of defining a conceptual model of the domain i.e. an ontology schema. For the initial development, the scope is limited to ontology population tasks. The entities and relations in this conceptual model will then become the candidates for annotation with task profiles, which will define how the instances of the entities and relations will be populated through an iterative workflow of human and machine tasks.

Task and Workflow Specification: As part of this research, I introduce the notion of *Semantics Driven Task Management*, enabled by semantically represented *task profiles*. In this stage, I map the semantic annotation tasks onto associated *task profiles*, which define the data sources, the nature of annotation and disambiguation required, the output mappings and other essential task design parameters. This facilitates modeling tasks to a varying degree of granularity. The provision for customizable and generic task templates is made to encourage reusable tasks templates. The *task profiles* primarily pertain to various common knowledge engineering tasks such as creation of concepts, relations, hierarchies, entity links, annotation and curation to name a few. A key consideration for effective task management is the representation of simple vs. knowledge intensive tasks.

Another key aspect of the framework is enabling *Semantics Driven Workflows*. This is achieved using *workflow profiles*, which contain descriptions of

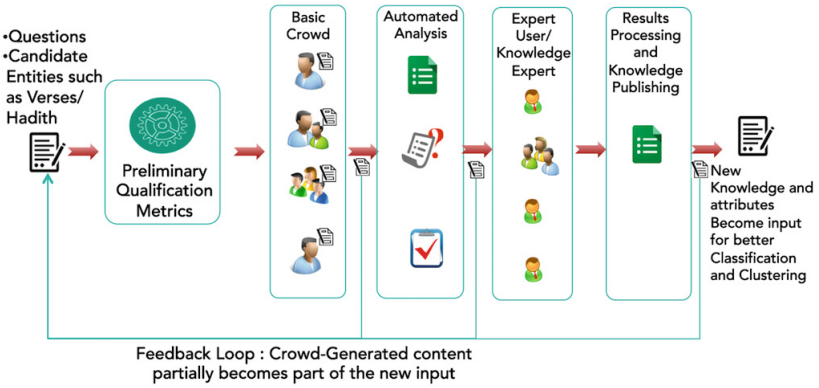

Fig. 3. Illustration of hybrid, iterative workflows

how the tasks that are subject to human vs. machine computation are composed together. The proposed framework utilizes semantics based representations for *workflow specification and management*. This is a unique and novel feature of the framework. Semantic representation of workflows provides the means to dynamically route and delegate suitable tasks to the suitable set of workers based on the skills and knowledge requirements. A workflow may be a composition of machine-task, crowd-task, niche(expert)-task or more depending upon the nature and the complexity of the task subject to the case study. The *workflow profile* captures the composition of tasks, their requirements, processing outcomes and any dependencies amongst other parameters. In relation to the conceptual model, a *workflow profile* for each task will specify the steps required to be performed in order for the entities and relations in a particular triple to be populated. A machine based task in the workflow may include an automated Information Retrieval (IR) task, or a task based on Natural Language Processing (NLP), such as extracting morphological variations of a given word (in Arabic). Text mining techniques are also employed to find portions of semantically relevant texts. Existing knowledge is also retrieved from available LOD sources. For the human driven tasks, a hybrid approach is utilized.

Human-Machine Workflow Execution: Based on the task and workflow specification designed in the previous stage, the system is implemented for executing these workflows. The proposed framework enables *Skill Driven Crowd Management* by utilizing the idea of niche or expert sourcing, managing task allocation and response collection to and from generic and specialized crowds. In the case of Islamic knowledge modeling, niches or experts would essentially be the teachers, scholars/subject-specialists in the domain. As per this idea, distinction is made between the contributors who can only perform simple, atomic tasks, vs. those to whom the knowledge intensive tasks are best delegated to. In addition, an important contribution is the notion of *Iterative and Dynamic*

Workflows, which ensures designing tasks and dynamic workflows, that execute in an iterative manner, suited to the knowledge and the skills of the crowd. Figure 3 shows an illustration of such a workflow, whereby preliminary contributions for a task are gathered from a generic crowd, and based on some analysis and findings, a followup task requires experts or super users to validate or advance upon the contributions obtained to deliver concrete results. Initial efforts aim to bridge and combine the contributions from platforms such as Amazon Mechanical Turk[1] (AMT) and the custom web framework, however, as the framework matures, I consider completely relying on my own framework.

Ontology Population, Linking and Publishing: Once the results are aggregated and validated, the last stage of the framework is designed to automatically populate the ontology under consideration, link with available linked data sources and publish the knowledge base.

5 Ongoing Work and Preliminary Results

My initial work focused on developing a reusable and generic crowdsourcing workflow using the AMT. I have developed and tested this in the context of linked data management tasks, and it has been recently published in our research entitled CrowdLink [24]. Although, this work was limited to linked data management tasks, however based on the experience and findings, I envision that a more mature human-machine computation framework, with hybrid properties, to solve the inherent challenges of knowledge engineering in specialized and knowledge intensive domains may be achieved.

My dwellings into the domain of Islamic knowledge has shown great potential for further undertaking this domain. I have conceptualized a macro-knowledge structure for Islamic knowledge and define macro and micro level links within the Islamic knowledge [4]. I have developed two key case studies to validate the different aspects of the framework. The central theme of the case studies focus on the two primary sources of Islamic knowledge namely: *The Qur'an* and the *Hadith*. The key aspects and the work status of each is summarized:

Case Study A - Thematic Disambiguation, Annotation and Linking of the Quranic Verses: The main focus of this case study includes four tasks: (1) Thematic Disambiguation, (2) Semantic Annotation, (3) Thematic Classification and (4) Semantic Relation Identification for the verses of the Qur'an.

For illustration consider the *Thematic Disambiguation* task. The Arabic is a rich morphological language and a phrase often needs to be disambiguated for its meaning. A typical workflow involves the extraction of the candidate verses that may or may not menifest a particular theme, based on some NLP or information extraction technique. These candidate verses then become input to a micro-task on AMT. I have also implemented a similarity computation framework for the Qur'anic verses [25], for retrieving highly similar and relevant verses. The crowd

[1] www.mturk.com.

disambiguates the theme's occurrence in the verse. The responses from the crowd are aggregated and analyzed.

To capture some of this knowledge and links, I have conducted some initial experimentation using the AMT. The results from the AMT have been promising for the simpler tasks, even though the tasks required the workers to have knowledge of the arabic language. However, some tasks such as identifying internal relationships within the verses of the Quran, which require deeper knowledge are not as straight forward.

My ongoing work is focusing on developing decision metrics and getting expert reviews and validations for a subset of the tasks which have been crowdsourced. For this purpose, I have developed a prototype web application, which takes as input the aggregated results of the tasks from the AMT and recommends them to suitable experts for review and validation.

Case Study B - Thematic and Inter-Contextual Linking of the Quran and Hadith: The foremost contribution of this case study is to use the learnersourcing methodology, a specialize form of crowdsourcing, using the custom developed web application to gather annotations, classifications and relationships not only for the verses of the Qur'an but most importantly, the relationships between the Qur'anic verses and the Hadith. This is by far the most knowledge intensive and complex task, since the process may not be automated. The conceptual modeling and the task designs for this case study has been completed. This case study combines all the various aspects of the framework that have been outlined and is currently in progress.

6 Evalution Plan

I evaluate the methodology for my research by contextualizing the domain of application to Islamic knowledge, given its sensitive and knowledge intensive nature. I define task execution workflows pertaining to some key knowledge engineering tasks in the domain primarily focusing on thematic disambiguation and annotation. The purpose of these experiments is two fold. Firstly, to evaluate crowdsourced thematic annotation data in comparison with ground truth obtained from experts. Secondly, to establish the amenability of crowdsourcing for obtaining ground truths in contexts where no prior baseline annotations exist.

In addition, I also plan to design, experiment and analyze other tasks such as semantic relation identification for identifying internal relationships between knowledge units, whereby, a considerable degree of knowledge expertise is required. The aim is to establish the right balance between crowd and expert contribution in achieving reasonably confidence in the acquisition of knowledge.

I will also attempt to carry out an analysis of the extent to which the semantic profiling of tasks and crowds improves the task annotation quality. I expect that some reasonable metrics will evolve in the process. I also aim to investigate to what extent our methodology will be generalizable to other tasks and domains. The results of this analysis will help establish the scalability of my approach.

7 Conclusions

In this paper, I propose a semantics driven framework for combining human and machine computation for the purpose of knowledge engineering. The framework utilizes semantics based tasks, workflow and worker profiles and allows for design and execution of iterative and dynamic workflows based on these profiles. The contextualization that my research aims to tackle, of engineering formalized knowledge models for the purpose of an enhanced knowledge seeking experience in the Islamic knowledge domain, is a high-impact problem, however a non-trivial one. The knowledge at hand is sensitive, and ensuring the credibility and authenticity of knowledge sources is challenging. I therefore believe that a hybrid approach, whereby contributions from crowds and experts, based on skills and knowledge background, combined with automated approaches, will considerably improve the efficiency and reliability of semantic annotation tasks in specialized domains. My initial experiments show favorable results for thematic disambiguation and annotation tasks. For more knowledge intensive tasks, work is underway for obtaining expert contributions and inputs based on aggregated results from crowd inputs. I expect to perform well defined evaluation of our decision metrics, for analyzing and aggregating these annotations. I also envision making the framework generalizable for other knowledge intensive domains.

Acknowledgments. I would like to acknowledge the contributions from my supervisors Dr. Khaled Rasheed and Dr. I. Budak Arpinar for their support and advice.

References

1. Simperl, E., Acosta, M., Flöck, F.: Knowledge engineering via human computation. In: Handbook of Human Computation, pp. 131–151. Springer, New York (2013)
2. Siorpaes, K., Simperl, E.: Human intelligence in the process of semantic content creation. World Wide Web **13**(1–2), 33–59 (2010)
3. Sabou, M., Scharl, A., Michael, F.: Crowdsourced knowledge acquisition: Towards hybrid-genre workflows. Int. J. Semant. Web Inf. Syst. **9**(3), 14–41 (2013)
4. Basharat, A., Rasheed, K., Arpinar, I.B.: Towards linked open islamic knowledge using human computation and crowdsourcing. In: Proceedings of the International Conference on Islamic Applications in Computer Science And Technology (2015)
5. Sherif, M.A., Ngomo, A.C.N.: Semantic Quran - a multilingual resource for natural-language processing. Semant. Web **6**(4), 339–345 (2015)
6. Atwell, E., Brierley, C., Dukes, K., Sawalha, M., Sharaf, A.B.: An artificial intelligence approach to arabic and islamic content on the internet. In: Proceedings of NITS 3rd National Information Technology Symposium (2011)
7. DiFranzo, D., Hendler, J.: The semantic web and the next generation of human computation. In: Handbook of Human Computation, pp. 523–530. Springer, New York (2013)
8. Inel, O., et al.: CrowdTruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In: Mika, P., et al. (eds.) ISWC 2014, Part II. LNCS, vol. 8797, pp. 486–504. Springer, Heidelberg (2014)

9. de Boer, V., Hildebrand, M., Aroyo, L., De Leenheer, P., Dijkshoorn, C., Tesfa, B., Schreiber, G.: Nichesourcing: Harnessing the power of crowds of experts. In: ten Teije, A., et al. (eds.) EKAW 2012. LNCS, vol. 7603, pp. 16–20. Springer, Heidelberg (2012)
10. Oosterman, J., Bozzon, A., Houben, G.J., et al.: Crowd vs. experts: nichesourcing for knowledge intensive tasks in cultural heritage. In: International WWW Conferences Steering Committee, pp. 567–568 (2014)
11. Retelny, D., Robaszkiewicz, S., To, A., Lasecki, W.S., Patel, J., Rahmati, N., Doshi, T., Valentine, M., Bernstein, M.S.: Expert crowdsourcing with flash teams. In: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, pp. 75–85. ACM (2014)
12. Noy, N.F., Mortensen, J., Musen, M.A., Alexander, P.R.: Mechanical turk as an ontology engineer? Using microtasks as a component of an ontology-engineering workflow. In: Proceedings of the 5th Annual ACM Web Science Conference, WebSci 2013, pp. 262–271. ACM, New York (2013)
13. Sarasua, C., Simperl, E., Noy, N.F.: CrowdMap: Crowdsourcing ontology alignment with microtasks. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012, Part I. LNCS, vol. 7649, pp. 525–541. Springer, Heidelberg (2012)
14. Siorpaes, K., Hepp, M.: OntoGame: Weaving the Semantic Web by Online Games. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 751–766. Springer, Heidelberg (2008)
15. Niepert, M., Buckner, C., Allen, C.: Working the crowd: Design principles and early lessons from the social-semantic web. In: Proceedings of the Workshop on Web 3.0: Merging Semantic Web and Social Web at ACM Hypertext (2009)
16. Simperl, E., Acosta, M., Norton, B.: A semantically enabled architecture for crowdsourced linked data management. In: CrowdSearch, pp. 9–14. Citeseer (2012)
17. Hanika, F., Wohlgenannt, G., Sabou, M.: The uComp Protégé plugin: crowdsourcing enabled ontology engineering. In: Janowicz, K., Schlobach, S., Lambrix, P., Hyvönen, E. (eds.) EKAW 2014. LNCS, vol. 8876, pp. 181–196. Springer, Heidelberg (2014)
18. Adams, B., Wohlgenannt, G., Sabou, M., Hanika, F., et al.: Crowd-based ontology engineering with the ucomp protégé plugin. Semantic Web (Preprint), pp. 1–20
19. Siorpaes, K., Hepp, M.: Games with a purpose for the semantic web. IEEE Intell. Syst. **23**(3), 50–60 (2008)
20. Simko, J., Bielikov, M.: Semantic acquisition games, pp. 35–50 (2014)
21. Von Ahn, L., Dabbish, L.: Designing games with a purpose. Commun. ACM **51**(8), 58–67 (2008)
22. Sabou, M., Scharl, A., Fols, M.: Crowdsourced knowledge acquisition: Towards hybrid-genre workflows. Int. J. Semant. Web Inf. Syst. **9**(3), 14–41 (2013)
23. Luz, N., Silva, N., Novais, P.: Generating human-computer micro-task workflows from domain ontologies. In: Kurosu, M. (ed.) HCI 2014, Part I. LNCS, vol. 8510, pp. 98–109. Springer, Heidelberg (2014)
24. Basharat, A., Arpinar, I.B., Dastgheib, S., Kursuncu, U., Kochut, K., Dogdu, E.: Semantically enriched task and workflow automation in crowdsourcing for linked data management. Int. J. Semant. Comput. **8**(04), 415–439 (2014)
25. Basharat, A., Yasdansepas, D., Rasheed, K.: Comparative study of verse similarity for multi-lingual representations of the qur'an. In: Proceedings on the International Conference on Artificial Intelligence (ICAI), pp. 336–343 (2015)