CroKnow: Structured Crowd Knowledge Creation

Jasper Oosterman(✉), Alessandro Bozzon, and Geert-Jan Houben

Delft University of Technology, P.O. Box 5031, 2600 GA Delft, Netherlands
{j.e.g.oosterman,a.bozzon,g.j.p.m.houben}@tudelft.nl

Abstract. This demo presents the *Crowd Knowledge Curator* (CroKnow), a novel web-based platform that streamlines the processes required to enrich existing knowledge bases (e.g. Wikis) by tapping on the latent knowledge of expert contributors in online platforms. The platform integrates a number of tools aimed at supporting the identification of missing data from existing structured resources, the specification of strategies to identify and invite candidate experts from open communities, and the visualisation of the knowledge creation process status. CroKnow will be demonstrated through a case study focusing on the enrichment of the Rijksmuseum Amsterdams digital collection.

Keywords: Crowd identification · Semantic representation · Knowledge creation

1 Introduction

Fuelled by the ever-growing need for open and semantically rich data sources, Knowledge Crowdsourcing (KC) is rapidly becoming a common tools for organisations to outsource knowledge creation to (possibly anonymous) individuals and communities willing to contribute with their domain-specific expertise. We refer to these contributors as *expert contributors* (or *experts*), so to stress their experience of, or their insight into, a targeted domain of knowledge. Artwork annotation is a known example of a KC task. There, successful contributors must be able to understand the abstract, symbolic, or allegorical interpretation of the reality depicted in the artwork, as well as to identify and recognise the occurrences of visual classes (e.g. plants, animals, objects) in the artwork [7,8]. Knowledge crowdsourcing is also fundamental to train expert systems (e.g. IBM Watson), or, in general, artificial intelligence methods focused on knowledge-related reasoning and prediction.

Previous work focusing on the identification of expert contributors for KC tasks, demonstrated how expert contributors could be approached and engaged to capitalise their familiarity with a domain of knowledge in order to execute activities such as content production, image annotation, etc. Candidate experts can be identified exploiting user modelling techniques relying on topic-based profiling [3], contextual properties (e.g. geographical location) [2], or Web content consumption [4]. In a recent work, we have shown how online social platforms

A. Bozzon et al. (Eds.): ICWE 2016, LNCS 9671, pp. 542–546, 2016.
DOI: 10.1007/978-3-319-38791-8_44

such as reddit are a viable source of contributors [5], and that carefully crafted expertise identification and invitation strategies can enable the interaction with large amount of expert contributors.

This paper presents the *Crowd Knowledge Curator* (CroKnow), a novel web-based platform that streamlines the processes required to enrich data sources by tapping on the latent knowledge of expert contributors in online platform[1]. The ultimate goal of CroKnow is to provide organisations with a tool that simplifies the crowdsourced creation and evolution of structured data sources (e.g. Wikis, or generic knowledge bases) by: (1) identifying missing data and specifying the knowledge to be created both at schema and instance level; (2) defining strategies for the identification and invitation of candidate experts from open communities; (3) supporting such experts in the knowledge creation task; and (4) keeping track of the knowledge creation process status.

The platform integrates and implements state-of-the-art methods and tools for each of these steps. Section 2 introduces the reader with the architecture of CroKnow, highlighting the components devoted to the identification and invitation of candidate experts, and to the extraction and quality assessment of knowledge from them. The demo will showcase the application of CroKnow to an artwork annotation problem, with a case study developed with the Rijksmuseum Amsterdam and their collection of 1M prints.

2 CroKnow Architecture

CroKnow has been developed with a modularity and customisation as main design goals. Figure 1 depicts its building blocks, each interacting with a centralised Orchestrator. Each component maps onto a process step in the crowd knowledge creation process, indicated with the numbers in the red circles.

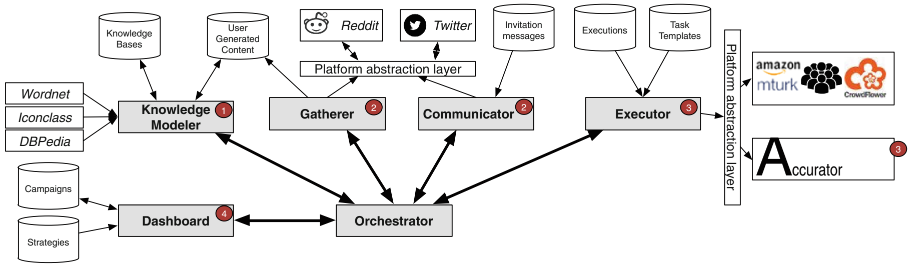


Fig. 1. CroKnow architecture (Color figure online)

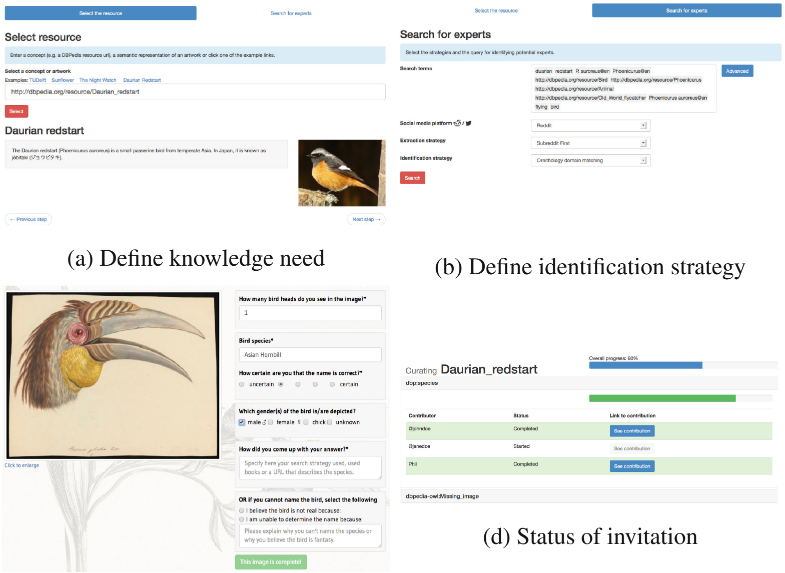
- The Knowledge Modeler ❶ handles the interaction with external knowledge bases to be enriched, and includes algorithms designed to assess the suitability of user generated content w.r.t. to a targeted knowledge model;

[1] A demonstration video is available at http://www.wis.ewi.tudelft.nl/ICWE2016_CroKnow.

- The Gatherer ❷ and Communicator ❷ respectively cater for the gathering of user generated content from online social platforms like reddit or Twitter, and the communication with users of such platforms;
- The Executor ❸ handles the creation and deployment of knowledge creation tasks by instantiating pre-defined templates. Tasks can be deployed on existing crowdsourcing platforms (e.g. Amazon Mechanical Turk), or on our own execution platform Accurator. The Executor is also responsible for the collection of runtime statistics about task executions.
- Finally, the Dashboard ❹ provides an easy-to-use interface where crowd knowledge campaigns can be specified and monitored by users. A campaign specification implies the specification, for each component, of suitable strategies.

Individual components of CroKnow have previously been instrumented, tested and used; Knowledge Modeler in [6], Gatherer and Communicator in [1,5], Executor and Accurator in [8]. The Dashboard component is a new addition resulting from the need to visualize the orchestration of the other components.

CroKnow had been implemented in Java and uses the web framework ERRAI for the front-ends of the dashboard and for the Accurator execution platform. Figure 2 depicts several instances of the user interfaces provided by the CroKnow Dashboard and by Accurator task execution interface.



Select resource

Enter a concept (e.g. a DBPedia resource url), a semantic representation of an artwork or click one of the example links.

Select a concept or artwork
Examples: TUDelft Sunflower The Night Watch Daurian Redstart

http://dbpedia.org/resource/Daurian_redstart

Daurian redstart

The Daurian redstart (Phoenicurus auroreus) is a small passerine bird from temperate Asia. In Japan, it is known as jōbitaki (ジョウビタキ).

(a) Define knowledge need

Search for experts

Select the strategies and the query for identifying potential experts.

Search terms: duarian redstart P. auroreus@en Phoenicurus@en http://dbpedia.org/resource/Bird http://dbpedia.org/resource/Phoenicurus http://dbpedia.org/resource/Animal http://dbpedia.org/resource/Old_World_flycatcher Phoenicurus auroreus@en flying bird

Social media platform / : Reddit

Extraction strategy: Subreddit First

Identification strategy: Ornithology domain matching

(b) Define identification strategy

How many bird heads do you see in the image?*
1

Bird species*
Asian Hornbill

How certain are you that the name is correct?*
○ uncertain ● ○ ○ ○ certain

Which gender(s) of the bird is/are depicted?
☑ male ♂ □ female ♀ □ chick □ unknown

How did you come up with your answer?*
Specify here your search strategy used, used books or a URL that describes the species.

OR if you cannot name the bird, select the following
○ I believe the bird is not real because:
○ I am unable to determine the name because:
Please explain why you can't name the species or why you believe the bird is fantasy.

(c) Share knowledge on Accurator

Overall progress: 60%

Curating Daurian_redstart

dbp:species

Contributor	Status	Link to contribution
@johndoe	Completed	See contribution
@janedoe	Started	See contribution
Phil	Completed	See contribution

dbpedia-owl:Missing_image

(d) Status of invitation

Fig. 2. Screenshots of CroKnow components

Figure 2a shows the selection of a resource in a semantic repository, e.g. a resource from DBPedia. The resource selection is the minimal knowledge need specification for the process, but it can also be extended by including an (automated) identification of missing properties w.r.t. to similar resources of the same type.

Figure 2b shows the definition of a search strategy (targeted online platform, queries for expert identification etc.) based on a knowledge need specified in the previous step. The knowledge need can require different levels of expertise. For example, the creation or retrieval of a descriptive image arguably requires a lower level of expertise, compared to the task of determining the name of a bird species. CroKnow supports the formulation of queries using (a combination of) keywords, properties from the structured resource and structured knowledge bases such as ontology's, taxonomies and vocabularies. The specificity of the query and the search strategy influences the amount of identified candidate contributors. CroKnow allows assessment and refinement of the chosen query and search strategy by providing feedback on the volume and relatedness of the identified candidates. The search strategy defines the target platform (reddit, Twitter etc.) and the method to extract user generated content from the target platform.

Figure 2c shows a task we have deployed for extracting domain specific knowledge (names of depicted bird species on artworks).

Lastly Fig. 2d shows the status of a campaign by visualising the status of the invitation and the executions. This feedback can serve as input to invite more candidates or to change the search strategy.

3 Demonstration Scenario - Artwork Annotation

We demonstrate CroKnow on the use case of artwork annotation, an actual use case developed together with Rijksmuseum Amsterdam. We will show how semantic resources can be used to define the knowledge need, and how the reddit and Twitter platform could be exploited to identify users with knowledge related the ornithology domain. The dashboard component (see Fig. 1) will provide users with an overview of the status of the crowd knowledge generation process.

References

1. Balduini, M., Bocconi, S., Bozzon, A., Valle, E.D., Huang, Y., Oosterman, J., Palpanas, T., Tsytsarau, M.: A case study of active, continuous and predictive social media analytics for smart city. In: Proceedings of the 5th SSC Workshop, Riva del Garda, Italy, 19 October 2014, pp. 31–46 (2014)
2. Cheng, Z., Caverlee, J., Barthwal, H., Bachani, V.: Who is the barbecue king of texas?: a geo-spatial approach to finding local experts on twitter. In: Proceedings of SIGIR2014, SIGIR 2014, pp. 335–344. ACM, New York (2014)
3. Difallah, D.E., Demartini, G., Cudré-Mauroux, P., Pick-a-crowd: tell me what you like, and i'll tell you what to do. In: Proceedings of the WWW 2013, pp. 367–374 (2013)

4. Ipeirotis, P.G., Gabrilovich, E.: Quizz: targeted crowdsourcing with a billion (potential) users. In: Proceedings of the WWW 2014, WWW 2014, pp. 143–154. ACM, New York (2014)
5. Kassing, S., Oosterman, J., Bozzon, A., Houben, G.: Locating domain-specific contents and experts on social bookmarking communities. In: Proceedings of the SAC 2015, Salamanca, Spain, 13–17 April 2015, pp. 747–752 (2015)
6. Nottamkandath, A., Oosterman, J., Ceolin, D., Fokkink, W.: Automated evaluation of crowdsourced annotations in the cultural heritage domain. In: Proceedings of the 10th URSW Workshop, Riva del Garda, Italy, 19 October 2014, pp. 25–36 (2014)
7. Oosterman, J., Nottamkandath, A., Dijkshoorn, C., Bozzon, A., Houben, G., Aroyo, L.: Crowdsourcing knowledge-intensive tasks in cultural heritage. In: ACM Web Science Conference, pp. 267–268 (2014)
8. Oosterman, J., Yang, J., Bozzon, A., Aroyo, L., Houben, G.-J.: On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks. Computer Networks **90**, 133–149 (2015)