Dipartimento di Elettronica, Informazione e Bioingegneria

POLITECNICO DI MILANO

August 16, 2014

To: STSM coordinator of Keystone COST Action

From: Marco Brambilla, Politecnico di Milano

STSM Scientific Report

STSM Visit by Marco Brambilla from Politecnico di Milano to Paris Dauphine University

1. Introduction

This report describes objectives and achievements of my visit of 5 days at the Dauphine University in Paris in July 2014, at the purpose of setting up a collaboration with the LAMSADE group. The main contact in the group was Khalid Belhajjame.

The visit has been completed successfully and has been crucial to start research collaborations between the two institutions on keyword-, knowledge- and crowd- based search over structured and rich data sources. A preliminary publication that was designed and submitted during the visit has been accepted at a workshop (see the results section for further details).

2. Purpose of the STSM

The objective of the visit was to combine semantic modeling and crowdsourcing at the purpose of improving keyword-based search. Concretely, the visit aimed to start a collaboration on keyword based search upon complex, structured data sources, by combining the expertise of Dauphine Paris on semantic modeling and querying of structured big data, with the expertise of my own group at Politecnico di Milano on crowdsourcing, media analysis, and structured Web content.

The two fields are largely complementary. Indeed, a combined research that aims at associating crowd based approaches with traditional keyword based techniques has the potential of significantly advance the knowledge in the fields of interest for Keystone.

3. Description of the work carried out during the STSM

According to the workplan attached to the request of the STSM, the following activities have been performed:

- Respective presentations of research of Dauphine University and of Politecnico di Milano
- Workshops for analysis of possible fields of collaboration
- Workshops on structured data sources formalization
- Workshop on crowdsourcing modeling and formalization
- Planning and preparation of a small, joint publication that proposes a vision on how to merge knowledge-based and crowd-based approaches to search and enrichment
- Planning of next steps of collaboration, including definition of possible experiments and publications

The work has been conducted mainly together with Khalid Belhajjame and Daniela Grigori, both affiliated with the LAMSADE group at Paris Dauphine.

4. Description of the main results obtained

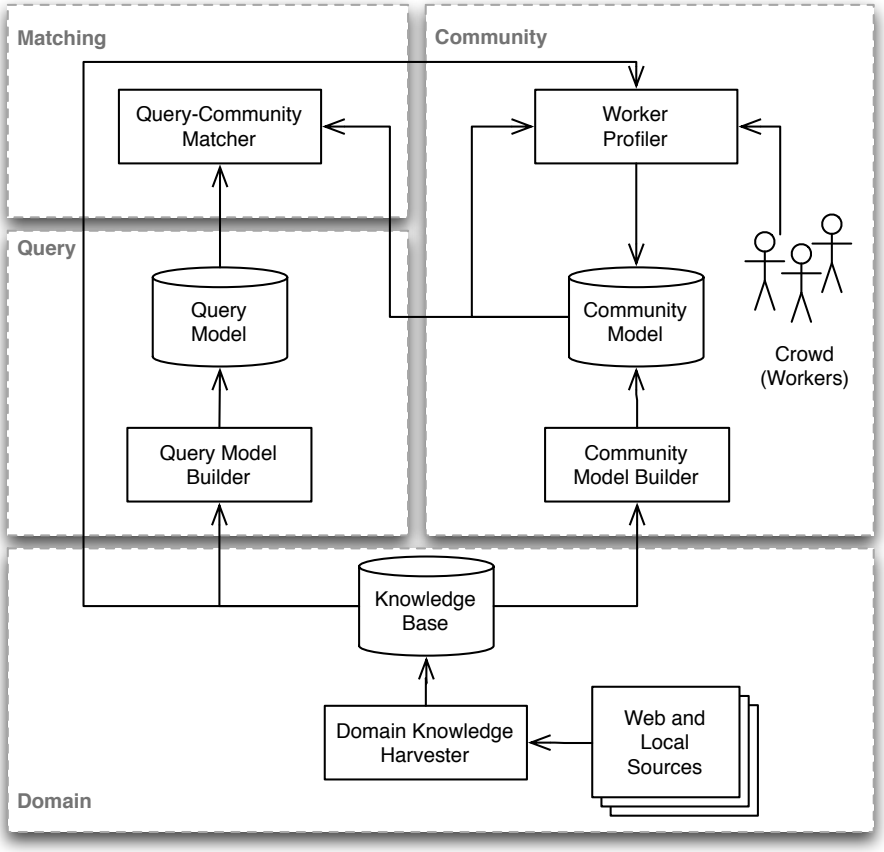
The main result of the visit is the conceptualization of **an approach for community profiling for improving the quality of query resolution through targeted crowdsourcing**.

Such result has been condensed **in a vision paper that has been submitted (and accepted already) to a French workshop on crowdsourcing**, namely: Crowdsourcing and human computation multidisciplinary workshop, organized in Paris by the CNRS MASTODONS challenge on September 15, 2014.

The objective of our work is to define in a rich way the concept of community of workers, defining a way to properly match crowdsourcing tasks to the communities based on expertise of workers and field/topic of tasks. We describe a set of conceptual models for queries, communities and workers, a high level architecture of the approach, and outline a set of possible strategies for addressing various aspects of expert-targeted crowdsourcing.

The figure below depicts the conceptual architecture we propose for addressing the problem of crowd targeting. The architecture covers four main aspects: Domain, Query (the question posed in the crowdsourcing task), Community and Matching. The harvesting component extracts information from existing web portals, social contents, etc., and stores it in a domain knowledge base to be used for describing communities and queries in the crowdsourcing

campaign. In what follows we outline the models underlying the Community and Query aspects, and outline strategies for matching queries with communities.



The **community model** is used to gather together workers characterized by an expertise potentially useful for answering given crowdsourcing queries. A community is defined intensionally in our system, meaning that it is not defined in terms of set of members, but instead in terms of its properties. Workers will be assigned to the community based on the matching of their profile with the community properties. In our model, we characterize a community using the following properties: Name, textual description, tags (semantic annotations coming from the knowledge base), type of community (explicit, meaning that it is known among workers; or implicit, i.e., determined purely by analyzing some domain knowledge harvested from the web), duration (statically defined or dynamic), grouping factor (interest, friendship, location, expertise, affiliation, etc.), communication channel (the means by which the members of the community can be solicited).

To define communities, the **community builder** can harvest multiple sources of information, such as the knowledge base itself, previous crowdsourced queries and worker profiles. The builder is also in charge of defining relationships between communities.

The **workers** that join (or are associated with) a given community are profiled using one of the following methods: i) Explicitly, i.e., by asking the workers to identify his/her expertise from a list of kinds of expertise or concepts; ii) Implicitly using their profile information on the crowd platform, when available; iii) Implicitly using work quality, i.e., by asking the worker a list of questions, and identifying his/her expertise by analyzing the quality of results (assuming to have the ground truth of the questions), possibly including tasks performed in the past, when available.

The **query model** describes the queries, i.e, the questions asked to the crowd within a tasks, for which executors are required to provide a response. We distinguish between query template and concrete query. A query template defines the structure of the question to be asked to the worker, without referring to specific entities, whereas a concrete query refers to particular entities. A query template is characterized by a textual description of the task it involves, the kind of operation that is requested to the worker, and the definition of expertise that may be needed for the worker to perform it (for instance, expertise on movies). Concrete queries can also include more precise description of needed expertise (for instance expertise on romance or thriller movies). Therefore, the whole query model is annotated with concepts from the domain knowledge base. Finally, a query may be also temporal-dependent, meaning that the correct response may change over time.

To assist the user in specifying a query model, the **query builder** model provides information in the knowledge base that may be of assistance, such as tags or keywords that are used and understood by the crowd and that can be matched with the worker profiles.

Given a query (be it template or concrete) and a collection of communities, the role of **matching** is to associate the query to the communities of workers that are best suited for the task implied by the query. We envisage the following two matching strategies.

- **Naive keyword-based matching.** Communities and queries are treated as bag of words. The bag of words are extracted from the properties characterizing the communities (name, description, etc.) and the query. Matching is calculated as the overlapping of the two bags of words.

- **Semantic matching.** Communities and queries are mapped to concepts (tags, taxonomies) described in the domain knowledge base, which capture the expertise they provide and require respectively. The matching is then performed based on such semantic annotations, also considering semantic associations between them.

5. Future collaboration with the host institution

We believe the field of expert-based crowdsourcing as a solution to keyword-based search is still largely unexplored and open to innovation. So far we elaborated a high-level description of our vision that aims at improving the quality of crowdsourcing results through expertise matching, without requiring increase of cost, thanks to automatic community building and matching techniques. However, the large part of the work is still to be done.

We plan to continue the collaboration between the two institutions and address the following problems:

- how to harvest a knowledge base that allows profiling communities and user queries in an optimal way.
- how to cope with the dynamics of both communities and work, due to changing needs, communities and expertise of workers over time.
- how to deal with query requiring multiple kinds of expertise, or expertise that is not explicitly defined within the community model.

Most of the collaboration will be achieved remotely and will involve further people (M.Sc. students, Ph.D. students and faculties).

6. Foreseen publications/articles resulting from the STSM

A first publication has been submitted and accepted at the Crowdsourcing and human computation multidisciplinary workshop, organized in Paris by the CNRS MASTODONS challenge on September 15, 2014. The paper is titled: **"Community Profiling for Crowdsourcing Queries"** and has been authored by: Khalid Belhajjame, Marco Brambilla, Daniela Grigori, and Andrea Mauri. See http://www.cnrs.fr/mi/spip.php?article355 for further details on the venue. See the attached program where our paper is highlighted in the last session of the day. We plan to attend the workshop and present our work there.

A set of experiements and implementations have been planned based on the described vision. Forthcoming joint publications will describe this research. We plan to submit further papers at conferences within the next year, reporting on our findings. Potential target venues are conferences such as WWW, ICWE, VLDB.

7. Confirmation by the host institution

See attached document by Paris Dauphine.

In faith,

Marco Brambilla

Politecnico di Milano