Open Innovation and Semantic Web : Problem Solver Search on Linked Data

Milan Stankovic

Hypios Research, 187 rue du Temple, 75003 Paris, France
STIH, Université Paris-Sorbonne, 28 rue de Serpente, 75006 Paris, France

milan.stankovic@hypios.com

Abstract. The novel practice of Open Innovation on the Web has imposed new challenges to known expert search approaches. At the same time, many potential sources of evidence about users' interest and expertise (e.g., research papers, blogs, activities) are becoming ubiquitously present as Linked Data. In this paper we present a research effort for suggesting the right way to search for potential Open Innovation problem solvers in Linked Data sources, by looking at the structure of available data sources. In addition, we seek to develop ways of suggesting domains of expertise that are in some way relevant to the domain of the Open Innovation problem, in order to enable a cross-domain solution transfer.

Keywords: Linked Data, User Profiling, Open Innovation, Solver Search.

1 Introduction

Open Innovation [1] websites like hypios.com, NineSigma.com and Innocentive.com serve as a showcase for R&D problems of big companies, where researchers and practitioners from diverse domains can propose solutions and earn rewards. In addition to broadcasting problems, an absolute requirement of clients that use such websites is the possibility to identify potentially good solvers for a particular problem and to invite them to submit solutions. This challenge is similar to expert finding, which has been treated for decades in Computer Science, but differs from standard expert finding scenarios in the following two significant points (challenges of solver finding):

- **Challenge 1.** Higher level of expertise does not necessarily make a solver more desirable. In fact studies like [2] show that people who are marginally related to the problem domain are likely to bring winning solutions.
- **Challenge 2.** The domain of the problem is not necessarily the only domain where the solvers could come from. Open Innovation approaches aim to enable solution (knowledge) transfer from one domain to another. The challenge is to target other potentially promising domains.

In addition, users tend to leave more and more traces about their interests, activities, results of their work, achievements, etc. – information that is vary valuable for detecting their fields of knowledge and interest. Many of those traces get published as data sets using the

Linked Data (LD) principles[1] – in the so-called LD Cloud. In this work we explore how the challenge of finding potential solvers might be effectively realized on LD.

2 Related Work in Expert Finding

Earlier expert finding approaches mostly took a legacy data set (research paper pdf files [3], e-mails, documents, etc.) and tried to extract structured data containing some traces about user's knowledge and interest. These data are later used for ranking them by the level of expertise. In each of those approaches there is an implicit assumption about what does it take to be an expert, and what makes one a better expert then another. The dataset that is used partially determines this assumption. For instance, an approach that uses research papers as a corpus, implicitly assumes that persons who wrote a research paper on a certain topic are experts on the topic. We call such assumptions *expertise hypotheses*[2]. Different expert search approaches use different expertise hypotheses. The first kind of expertise hypothesis relies on content owned and generated by users (e-mail, blogs, papers, etc.). Another branch of approaches relies on user activities (such as bookmarking, question answering and participation in projects). The third branch of approaches uses user's reputation, for expert finding and ranking. An exhaustive overview of expert finding approaches and their respective expertise hypotheses is given in our paper [4] and is omitted here due to the paper size. Many of the types of traces used by the abovementioned approaches are starting to appear in the LD cloud[3].

2 Proposed Approach and Methodology

In order to address the challenges faced by problem-solving websites, we plan to construct a system that would enable the discovery of additional topics related to the main topics of the problem in question (Challenge 2), and that would be able to use traces available in the LD cloud to identify potential solvers (Challenge 1). In our basic scenario a user from a company (called Seeker) provides the description of a problem for Open Innovation solving. He later enriches it with topic concepts and verifies those coming from concept extraction tool such as Zemanta. In the next step the user explores the space of related topics that are somewhat relevant for the problem in question, and includes them in the description (Challenge 2). We outline our approach for using Semantic Web to help the user explore the space of relevancy in the Section 2.1. The system then suggests the best expertise hypotheses to use for the given problem and the data set available to the user. The suggestion is made by looking at the structure of LD (see Section 2.2.) In the final step the user should invoke the search for potential solvers that correspond to given topics. The list of potential solvers could then be used for inviting them to solve the problem, or in any other Open Innovation scenario (e.g., giving the list of solvers to the Seeker, asking solvers to identify existing solutions in literature).

[1] http://www.w3.org/DesignIssues/LinkedData.html

[2] An example expertise hypothesis would be: "authors of two or more research papers on a particular topic can be considered as experts on that topic".

[3] We have created a map of existing LD sources that contain relevant user traces. The map can be accessed at http://milstan.net/hypios/competence-map/mec.html

2.1 Exploring the Space of Relevant Topics

Companies that involve in Open Innovation usually expect from unexpected domains that they did not target themselves. Existing research in recommender systems and semantic proximity [5] allows spotting different dimensions of relevancy. Apart from structural proximity of topics in a graph, topics might also be related by co-occurrence, or by being associated to structurally similar or analogous problems. We intend to rely on LD cloud in order to find potentially interesting topics related to the problem in question. For this purpose we will model the different dimensions of relevancy. In order to address the similarity of problems we will construct an ontology of problem structure, relying on existing early models [6].

2.2 Suggesting the Expertise Hypothesis based on Linked Data Structure

As a difference from the time before LD when expert search approaches had to focus on a particular hypothesis based on the data that was available to them, LD-based approaches can benefit from plenty of different kinds of traces and can choose among many expertise hypotheses. Therefore the choice of an appropriate expertise hypothesis becomes a challenge for LD-based expert finding. We propose a way to suggest appropriate expertise hypotheses for a given problem topic, based on the type of user trace that is used in those hypotheses. We argue that experts from different domains would use different communication channels (e.g., one domain mostly tweets, the other mostly blogs) and leave different user traces. Detection of such patterns would allow us to choose the expertise hypotheses that rely on traces significant for the given domain. We propose to explore the structure of LD and establish LD metrics that would help to identify good evidence types for particular topics. These metrics might also be beneficial in choosing the right data set in a scenario of running distributed queries over several data sets.

2.2.1 Metrics Based on Data Quantity

As the simplest metric we define Q_t to be the number of available instances of type *t*. Further on, it would be interesting to know the number of instances of a certain type, having some particular concept as value of dcterms:subject property. We thus define $Q_{t, C}$ where *C* is a set of concepts that identify topics that are associated with the instances to be counted. A similar metric is in fact already used by systems that use data summaries to accelerate query execution like [7]. Those graph summaries could directly serve as a source of $Q_{t, C}$ metric. If the data taken into account is a representative subset of world's data, then higher values of $Q_{t, C}$ should indicate that most of interactions around a particular topic are happening on a particular type of medium, and thus the use of such sources might result in higher precision. Experimentation should reveal more substantial correlations.

2.2.2 Metrics Based on Topic Distribution

We assume that prevailing use of particular topics with particular type of trace instances could positively influence the effectiveness of expert search. We define subject homogeneity $SH_{t,s}$ as number of instances of type *t* that are associated with topic *s*, divided by the total number of instances of type *t*. Subject homogeneity shows the degree of use of the subject *s* within the type *t*. We also define type homogeneity $TH_{t,s}$ as number of instances of type *t* that are associated with topic s, divided by the total number of trace instances associated with topic s. This metric shows the ratio of use of particular trace type with instances relevant for a particular topic. At the same time $\mathrm{TH}_{t,\,s}$ represents the upper bound of recall for expertise hypothesis using trace type t and searching for expert on topic s. In our experimentation we will certainly explore the possibility of including close matches of the topic in question, as well as its broader topics into calculation of homogeneity.

$$SH_{t,s} = \frac{Q_{t,s}}{Q_t} \quad TH_{t,s} = \frac{Q_{t,s}}{Q_{topTraceClass,s}} \tag{1}$$

2.2.4 Metrics Based on Data Quality

Although a number of approaches propose to Asses data quality based on provenance or based on data set popularity determined through link analysis [8], there is not much work on data completeness and usefulness. We are interested to assess the completeness of data in a particular source with regards to a particular expertise hypothesis.

3 Ongoing Initial Experimentation

After examining the current state of LD cloud and its ability to answer expert search queries (see [4]), we have discovered various issues including incomplete data, or missing links to topics and to trace authors. Having this in mind we have constructed a sample, rich enough data set with different types of traces and many different topics. Such a data set will enable us to launch experiments for evaluating the correlation of particular LD metrics with the precision and recall of expert identification. We first imported existing public LD data about user traces (mostly publications). In addition Sindice.com gave us a number of public data sources containing blog and publications data. To enrich the set we constructed an extractor for conference event tweets. For any type of trace, we processed the textual content of the trace using Zemanta to enrich them with DBPedia concepts. We have a similar approach for inferring the missing author data from the trace URI wherever possible (blog an tweet URIs identify their authors).

Table 1. Values of Linked Data Metrics from Preliminary Experimentation. [4]

Trace Type (*t*)	$Q_{t,LD}$	True Positives	Precision	Recall	Q_t	$TH_{t,LD}$	$SH_{t,LD}$
Tweets	77	57	.7402	.3608	6631	.3990	.0116
BlogPosts	1	1	1	.0063	837	.0052	.0012
ResearchPapers	87	86	.9885	.5506	1013	.4508	.0859

At the moment our data set contains 6631 tweets (including tweets from 5 latest Semantic Web research conferences), 837 latest blog posts, and 1013 research papers (including last 3 years' Semantic Web conferences). As the data set is mostly representative for the Semantic Web domain we have decided to run initial calculation on a topic related to Semantic Web, in particular we have calculated the values for proposed metrics as if we searched for solvers knowledgeable in Linked Data. The values of metrics are provided in the Table 1. Although it is too early to speak of any correlation, the correspondences of high recall values with the high values of $\mathrm{TH}_{t,\,LD}$ and $\mathrm{SH}_{t,\,LD}$ give hope to continue the research in this direction. We hope

[4] LD is the set of topics that contains only one concept - dbpedia:Linked_Data. True positives are counted by manual identification of real Linked Data experts available in the data store.

to include more domains in the future and conduct further evaluation in order to find eventual correlations between the values of different metrics and precision and recall.

4 Conclusions and Future Work

We have identified main challenges for the Semantic Web technologies to help perform Open Innovation on the Web. We have designed an approach for finding potential solvers on LD, by picking the right expertise hypothesis. We have proposed several LD metrics that might be useful for choosing the right expertise hypothesis and directing the solver search. In the future we plan to load greater quantities of data into our local LD store and test if there is a correlation between the values of proposed LD metrics and the precision and recall of expert/solver search. We also intend to construct an ontology to define different types of user traces and their correspondence to various existing concepts in ontologies used in LD cloud. This ontology should facilitate the passage from expertise hypotheses to SPARQL queries.

We are also considering other aspects that might be significant for the choice of hypothesis, like the temporal aspects of traces, expert candidates availability for problem solving, etc. We also plan to conduct experiments to test the performance of different approaches for discovery of relevant topics for cross-domain solution transfer. Our intention is to further develop several relevancy models (based on existing work from artificial intelligence, cognitive science and LD) that would enable the users to discover new relevant topics. We will test those models in a user study.

Acknowledgments. The work of Milan Stankovic has been partially funded by ANRT (CIFRE 789/2009). The Author would like to thank to his supervisor Philippe Laublet for his guidance.

References

1. Chesbrough, H. W. (2003). Open Innovation: The New Imperative for Creating and Profiting from Technology. Harvard Business Press.
2. Jeppesen, L., & Lakhani, K. (2009). Marginality and Problem Solving Effectiveness in Broadcast Research. Organization Science, 20.
3. Buitelaar, P., & Eigner, T. (2008). Topic Extraction from Scientific Literature for Competency Management. *The 7th International Semantic Web Conference*. Karlsruhe, Germany.
4. Stankovic, M., Wagner, C., Jovanovic, J., & Laublet, P. (2010). Looking for Experts ? What can Linked Data do for You? In Proceedings of Linked Data on the Web 2010, Raleigh, NC.
5. Passant, A., & Decker, S. (2010). Hey! Ho! Let's Go! Explanatory Music Recommendations with dbrec. (L. Aroyo, G. Antoniou, E. Hyvönen, A. Teije, H. Stuckenschmidt, L. Cabral, et al.)The Semantic Web: Research and Applications, Lecture Notes in Computer Science, 6089, 411-415. Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-13489-0.
6. Motta, E. 1999 Reusable Components for Knowledge Modelling: Case Studies in Parametric Design Problem Solving. 1st. IOS Press.
7. Harth, A., Hose, K., Karnstedt, M., Polleres, A., & Sattler, K. (2010). Data Summaries for On-Demand Queries over Linked Data. In *Proceedings of the 17th international conference on World Wide Web, WWW2010* (pp. 411-420). Raleigh, NC, USA: ACM Press.
8. Delbru, R., Toupikov, N., Catasta, N., Tummarello, G., & Decker, S. (n.d.). Hierarchical Link Analysis for Ranking Web Data. In *Proceedings of ESWC2010*. Heraklion, Grece.