A Review on Knowledge Base Semantic Integration using Crowdsourcing Through Ontological Model

[1]Mayur A. Pilankar,[2]Prof.Dr.D. S. Bhosale
[1]Student,AMGOI,Vathar tarf Vadgaon,[2]Professor, AMGOI,Vathar tarf Vadgaon
[1]Lecturer,ICRE,Gargoti,

Abstract-The semantic web has enabled the formulation of a growing number of knowledge bases (KBs), which are designed separately using different procedures. Integration of KBs has brought much attention as different KBs normally contain overlapping and interconnected information. Automatic techniques for KB integration have been developed but far from ideal. Therefore, in this paper, the problem of knowledge base semantic integration using crowd knowledge. There are both classes and instances in a KB, in our work, we propose a novel hybrid framework for KB semantic integration considering the semantic heterogeneity of KB class structures. We first perform semantic integration of the class structures via crowdsourcing, then apply the blocking-based instance matching approach according to the integrated class structure. For class structure (taxonomy) semantic integration, the crowd is leveraged to help to identify the semantic relationships between classes to handle the semantic heterogeneity problem. Under the conditions of both large-scale KBs and limited monetary budget for crowdsourcing, here formalize the class structure (taxonomy) semantic integration problem as a Local Tree Based Query Selection. Furthermore, the KBs are usually of large scales and have millions of instances, and direct pairwise-based instance matching is ineffective. Therefore, we adopt the blocking-based strategy for instance matching, taking advantage of the class structure integration result. The experiments on real large-scale KBs verify the effectiveness and efficiency of the proposed strategies.

Index Terms: -knowledge bases, crowdsourcing, semantic web, relevance.

I. INTRODUCTION

The semantic web has enabled the formulation of a growing number of knowledge bases (KBs), which are composed independently using several techniques. Integration of KBs has brought much consideration as different KBs usually contain overlapping and complementary information. Automatic techniques for KB integration have been improved but far from perfect. Therefore, in this work considering the problem of knowledge base semantic integration using crowd intelligence. There are both classes and instances in a KB, here proposed a novel hybrid framework for KB semantic integration considering the semantic heterogeneity of KB class structures. The system can build using semantic integration of the class structures via crowdsourcing, blocking based instance matching approach. Moreover, class structure (taxonomy) semantic integration. The crowd is leveraged to help to identify the semantic relationships between classes to handle the semantic heterogeneity problem. Under the conditions of both large-scale KBs and limited monetary budget for crowdsourcing, Also define the class structure (taxonomy) semantic integration problem as a Local Tree Based Query Selection (LTQS) problem. Also concentrating on rank based and graph-based algorithms for crowdsourced to knowledge refining, which judiciously selects the most beneficial candidate facts to conduct crowdsourcing and prune unnecessary information.

`

II. PROBLEM STATEMENT

Large-scale knowledge bases (KBs) have been constructed and are becoming more plentiful. These knowledge bases are constructed independently from different sources with different techniques, and different KBs usually contain overlapping and complementary information. As knowledge acquisition is an expensive process, reusing existing KBs is strongly desirable to reduce the cost of data management. Therefore, knowledge base integration has a most important approach to consider using crowdsourcing.

`

III. OBJECTIVES

The different objectives of the proposed system are:

- Designing a novel framework for KB integration by conducting taxonomy integration via human intelligence
- Performing blocking-based instance matching based on the integrated taxonomy.
- Designing the crowd assisted taxonomy using Ontology-based model.
- Designing and developing KB integration module for class structure integration and instance matching.

IV. RELEVANCE

There exist some crowdsourcing platforms, such as MTurk and CrowdFlower. In such platforms, asked human "workers" to complete micro-tasks. For example, ask them to answer questions like "Is Italy a country?" Each micro-task is referred to as a human intelligent task (HIT). After having completed a HIT, a worker is rewarded with a certain amount of money based on the difficulty of the HIT. That is, invoking the crowd for knowledge cleaning comes with a monetary cost. In addition, a human worker may not always produce a correct answer for a HIT. To mitigate such human errors, assign each HIT to multiple workers and then take a

V. LITERATURE REVIEW

[1] Rui Meng, Lei Chen, Yongxin Tong, Chen Zhang "Knowledge Base Semantic Integration using Crowdsourcing".

In this paper, Human computation has been used for centuries, which describes computation performed by a human and human computation system organize human efforts to carry out the computation. With the popularity of commercial platforms such as Amazon MTurk (AMT) and Crowd- Flower, the source of human is broadened to the crowd, namely, crowdsourcing.

In this paper, proposed a framework for large-scale knowledge base (KB) integration through human intelligence. The core problem in KB integration, i.e., taxonomy integration, is formulated as the Local Tree Based Query Selection (LTQS) problem, which is proved to be NP-hard. To solve this optimization problem, they propose two greedy-based query selection algorithms, i.e., static query selection algorithm and adaptive query selection algorithm. Based on the taxonomy integration result, align the instance through a shop neighbor based blocking strategy. Finally, we verify our proposed algorithms through extensive experimental studies on real large-scale KBs. In particular, experimental results demonstrate that our approach can operate the KB integration task both efficiently and effectively.

[2] D. Aumueller, H. H. Do, S. Massmann, and E. Rahm, "Schema and ontology matching with COMA++.

In this paper demonstrates an ontology matching tool COMA++, which combines different match metrics, i.e., similarity measure, linearly combining lexical, structural and extensional similarities, to support ontology matching.

[3] O. Udrea, L. Getoor, and R. J. Miller, "Leveraging data and structure in ontology integration.

The author proposes an ILIADS algorithm to integrate the data matching and logical reasoning techniques to conduct ontology integration. Integrates two ontology using lexical and structural features and calculates a similarity measure to derive the alignment, and then verify it to avoid semantic inconsistencies. In this presented ILIADS, a novel ontology integration tool which tightly integrates statistical matching with logical inference. This tight integration means that our algorithm can exploit data and structure effectively to produce high-quality integration results. they have investigated how the ontological structure itself affects the utility of different inference and matching strategies and how our algorithm can adapt to the characteristics of the ontologies to be merged. they have validated our results on an extensive collection of real-world ontologies. Our most important findings are that: (i) the number of inferences steps and the heuristic policies are independent of the particular input ontologies,(ii) ILIADS significantly outperforms COMA++ and FCA- merge, especially so for ontologies with a reasonable amount of instance data and (iii) the parameters of ILIADS correlate to structural properties of the ontologies and are stable for ontology pairs that do not have very different structures.

[4] F. M. Suchanek, S. Abiteboul, and P. Senellart, "Paris: Probabilistic alignment of relations, instances, and schema.

In this paper proposed an automatic approach for ontology alignment, in which not only instances but also classes and relations are aligned.

[5] S. Lacoste-Julien, K. Palla, A. Davies, G. Kasneci, T. Graepel, and Z. Ghahramani, "SIGMa: Simple greedy matching for aligning large knowledge bases.

Julien demonstrates a more efficient algorithm, SiGMa, for large-scale knowledge base alignment. SiGMa adopted an iterative propagation algorithm to align the instances between KBs. We have presented SiGMa, a simple and scalable algorithm for the alignment of large-scale knowledge bases. Despite making greedy decisions and never backtracking to correct decisions, SiGMa obtained a higher F-measure than the previous best-published results on the OAEI benchmark datasets and matched the performance of the more involved algorithm PARIS while being 50x faster on large-scale knowledge bases of millions of entities. Our experiments indicate that SiGMa can obtain good performance over a range of datasets with the same parameter setting. On the other hand, SiGMa is easily extensible to more powerful scoring functions between entities, as long as they can be efficiently computed. Some apparent limitations of SiGMa are a) that it cannot correct previous mistakes and b) cannot handle alignments other than 1-1. Addressing these in a scalable fashion which preserves high accuracy are open questions for future work. We note though that the non-corrective nature of the algorithm did not seem to be an issue in our experiments. Moreover, pre-processing each knowledge base with a de-duplication method can help make the 1-1 assumption, which is a dominant feature to exploit in an alignment algorithm, more reasonable. Another exciting direction for future work would be to use machine learning methods to learn the parameters of the more powerful scoring function. In particular, the 'learning to rank' model seems suitable to learn a score function which would rank the correctly labeled matched pairs above the other ones. The current level of performance of SiGMa already makes it suitable though as a powerful generic alignment tool for knowledge bases and hence takes us closer to the vision of Linked Open Data and the Semantic Web. We could also envision using it to align social networks, as they have rich graph information as well as multiple entity properties

VI. Proposed Work

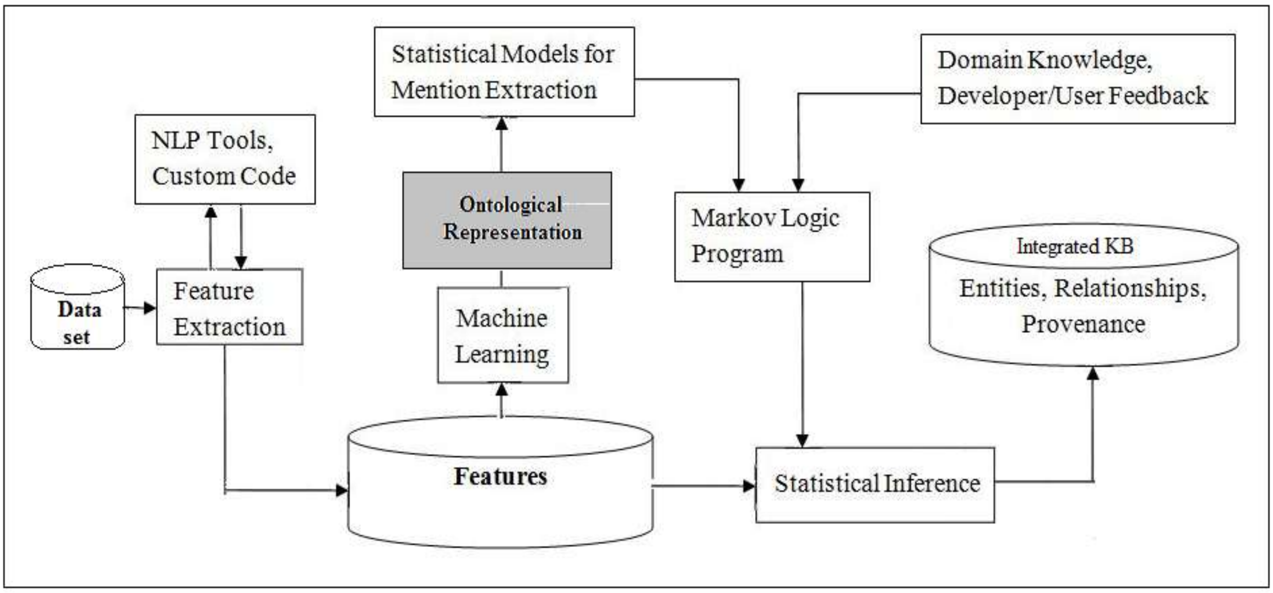
Statistical Models for Mention Extraction
Domain Knowledge, Developer/User Feedback
NLP Tools, Custom Code
Ontological Representation
Markov Logic Program
Integrated KB
Entities, Relationships, Provenance
Data set
Feature Extraction
Machine Learning
Features
Statistical Inference

Figure 1. Block diagram of the proposed system

The System Modules are elaborate as:

1. **Feature Extraction**

To scale up a system to web-scale KBC tasks, employ high-throughput parallel thread computing frameworks, Condor for feature extraction and File System for storage Fortunately, the Condor infrastructure supports a best-effort failure model, i.e., a job may finish even when Condor fails to process a small portion of the input data. Moreover, Condor allows us to simultaneously leverage thousands of machines from across a department, an entire campus, or even the nation-wide Open Science Grid.

2. **NLP**

Natural language processing tool used to process data from crowdsourcing. Data sources are with the different linguistic format so need to process with NLP tool.

3. **Machine Learning**

Traditional KBC systems work on manual annotations or domain-specific rules provided by experts, both of which are scarce resources. To remedy these problems developed the distant supervision approach for relation extraction. The input to distant supervision is a set of seed facts for the target relation together with an (unlabeled) text corpus, and the output is a set of (noisy) annotations that can be used by any machine learning technique to train a statistical relation-extraction model.

4. Ontological Representation

The different KBs are used as input to the Ontological model to measure the similarity between two or more sources. So that system can combine different KBs as per the similarity base.

5. Markov Logic

To scale up statistical inference in Markov logic, the system employs using different systems. To an inference of a grounding step that essentially performs relational operations and a search (or sampling) step that often comprises multiple independent subproblems.

6. Domain Knowledge

Domain Knowledge of the source is trained and used to maintain semantic relations between different KBs.

7. A statistical model for Mention Extraction

In this Model, Entity linking is used of mapping a textual mention to a real-world entity. By running the state-of-the-art to extract textual mentions and corresponding entity types. Then tries to map each mention with string matching.

8. Integrated KB

The relation-extraction models are trained. During feature extraction, perform dependency parsing. Use sparse logistic regression to train statistical relation-extraction models using both lexical and syntactic features.

VII. Operations/Steps Performed on the Proposed Model:

1. Pre-processing the dataset
2. Feature Extraction module
3. Nlp implementation over KB’s

4. Semantic Analysis implementation.
5. Extracting relational data from KB's

VIII. Experimental Setup:

The program is implemented using Java. The experiment is carried out by input various condition database and analyzing output with the trained database in Windows 7 operating system.

IX. Conclusion

In this article, we proposed a framework for large-scale knowledge base (KB) integration through human intelligence. The core problem in KB integration, i.e., taxonomy integration, is formulated as the Local Tree Based Query Selection problem, which is proved to be NP-hard. To solve this optimization problem, we offer two greedy-based query selection algorithms, i.e., a static query selection algorithm and adaptive query selection algorithm. Based on the taxonomy integration result, we align the instance through a shop neighbor based blocking strategy. Finally, we verify our proposed algorithms through extensive experimental studies on real large-scale KBs. In particular, experimental results demonstrate that our approach can operate the KB integration task both efficiently and effectively.

X. References

[1] Rui Meng, Lei Chen, Yongxin Tong, Chen Zhang "Knowledge Base Semantic Integration using Crowdsourcing" IEEE Trans.on Knowledge and data engg.vol 29 .May 2017

[2] D. Aumueller, H. H. Do, S. Massmann, and E. Rahm, "Schema and ontology matching with COMA++," in SIGMOD.

[3] O. Udrea, L. Getoor, and R. J. Miller, "Leveraging data and structure in ontology integration," in SIGMOD, 2007.

[4] F. M. Suchanek, S. Abiteboul, and P. Senellart, "Paris: Probabilistic alignment of relations, instances, and schema," PVLDB, 2011.

[5] S. Lacoste-Julien, K. Palla, A. Davies, G. Kasneci, T. Graepel, and Z. Ghahramani, "SIGMa: Simple greedy matching for aligning large knowledge bases," in KDD, 2013.

[6] J.Hoffart, F.M.Suchanek, K.Berberich,and G.Weikum,"YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," Artif. Intell., 2013. Pages 103–118, 1999.

[7] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in SIGMOD, 2012.

[8] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in SIGMOD, 2008.

[9] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Web-scale information extraction in knowitall:(preliminary results)," in WWW, 2004.

[10] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "DBpedia: A nucleus for a web of open data," in ISWC/ASWC, 2007.

[11] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T.M.Mitchell,"Toward an architecture for never-ending language learning," in AAAI, 2010.

[12] J. Euzenat, P. Shvaiko et al., Ontology matching, 2007.