Genetic-Fuzzy Programming Based Linkage Rule Miner (GFPLR-Miner) for Entity Linking in Semantic Web

Amit Singh, Jawaharlal Nehru University, New Delhi, India
Aditi Sharan, Jawaharlal Nehru University, New Delhi, India

ABSTRACT

This article describes how semantic web data sources follow linked data principles to facilitate efficient information retrieval and knowledge sharing. These data sources may provide complementary, overlapping or contradicting information. In order to integrate these data sources, the authors perform entity linking. Entity linking is an important task of identifying and linking entities across data sources that refer to the same real-world entities. In this work, they have proposed a genetic fuzzy approach to learn linkage rules for entity linking. This method is domain independent, automatic and scalable. Their approach uses fuzzy logic to adapt mutation and crossover rates of genetic programming to ensure guided convergence. The authors’ experimental evaluation demonstrates that our approach is competitive and make significant improvements over state of the art methods.

KEYWORDS

Entity Linking, Fuzzy Logic, Genetic Programming, Linked Data, Linked Open Data, RDF, Semantic Web

1. INTRODUCTION

Over the past years, the World Wide Web (WWW) is witnessing a rapid change from the document-oriented web to a distributed, proliferated machine readable Semantic Web (SW). The SW contains data sources on various domains such as people, publications, media, government organizations and social web, etc. These data sources are independent and geographically distributed. In order to harness real value from these sources, efforts must be made to integrate these sources and build an interlinked SW. This interlinked SW will facilitate interoperability of data, better utilization, and consumption of data. These links can play a major role in processing federated queries and complex question answering by retrieving information available across data sources.

The SW data sources follow linked data principles. Linked Open Data (LOD) cloud (Bizer, Heath, & Berners-Lee, 2009) is considered as a largest available source of structured information following Linked Data Principles. As per LOD Cloud\(^1\), this LOD currently consists of 1163 interlinked datasets, which consists of billions of facts as resource description framework (RDF) triples. This LOD is

DOI: 10.4018/IJSWIS.2018070107

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
growing rapidly in volume. As per LODStats\(^2\) LOD recorded around 1 billion triples till 2011 which had increased to 149.42 billion by 2017.

According to a study by (Schmachtenberg, Bizer, & Paulheim, 2014), 44% of the LOD datasets are not connected to other datasets at all, and there exist less than 400 million links over 30 billion triples published as LOD. In order to provide accurate links between datasets, the system needs to identify different entities that refer to the same real-world resource. This problem is called entity linking (EL); EL is a fundamental task in semantic web data integration. The main issue behind this lack of links is that manual creation of links is a very tedious process and infeasible in the case of large data sources like DBpedia\(^3\) (4.5 million entities) and LinkedGeoData\(^4\) (1+ million entities). Therefore, there is a need to automatize the process of EL.

In this work, authors have presented a rule based approach to perform EL. In the rule-based technique, a proper selection of similarity/distance measure, an optimal threshold value has to be made to avoid false positive and false negatives. Therefore, the success of rule-based approach highly depends on the proper choice of similarity/distance function and threshold. Datasets belonging to same domain may behave differently so rules designed for one particular dataset may not work on another dataset. Therefore, EL in-data sources is a non-trivial problem. In this work, the authors have developed an automatic domain-independent approach to learn EL rules. These rules will specify the conditions two entities must fulfill in order to be interlinked.

The authors are using Genetic Programming (Koza & Poli, 2005), as a branch of evolutionary computation (Fogel, 2000) to exploit the search space and choose a proper pair of property-similarity measures and threshold. Genetic Programming (GP) allows us freedom from predefining structure and size of the program. GP allows a variable length representation in form of syntax trees. The main contribution of this work lies in improving the performance of GP based LR learning. The performance of GP depends on the proper use of control parameters. These control parameters mutation and crossover rate are adaptively tuned to speed up the process and avoid premature convergence. GP programming also suffers from uncontrolled growth in individual size without any significant improvement in solution quality. This phenomenon is called as bloating. The authors have improved the performance of GP by avoiding the bloating effect. The major contributions of proposed work are summarized as follows:

- The authors propose an adaptive GP based approach to automatically learn LRs. Our approach uses Fuzzy logic (FL) to adopt mutation and crossover rates to ensure fast convergence as compared to other state of art approaches. The authors propose the use of double tournament selection in conjunction with depth limiting as rule pruning strategy to avoid bloating effect in GP.
- The authors propose the use of syntax aware crossover operators instead of sub-tree crossover operators so that a genetic operation always generate a valid LR.
- The authors have compared the proposed approach’s performance in terms of effectiveness and efficiency against the state of the art methods on several datasets.

Rest of the paper is structured as follows: in section 2 the authors present a literature survey of different EL systems based on underlying approach. In section 3 they formally define the problem. The authors have given an overview of the GP technique for LR learning in section 4. In section 5, the authors have presented their proposed approach of generating LRs using adaptive GP. Section 6 deals with the experimental setting and reports the results obtained from the proposed algorithm. The last section covers concluding remarks and future scope.
Applying an Organizational Uncertainty Principle: Semantic Web-Based Metrics

Techniques for Named Entity Recognition: A Survey