Advances in Intelligent Systems and Computing 373

Sigeru Omatu · Qutaibah M. Malluhi
Sara Rodríguez González · Grzegorz Bocewicz
Edgardo Bucciarelli · Gianfranco Giulioni
Farkhund Iqba *Editors*

Distributed Computing and Artificial Intelligence, 12th International Conference

Springer

Advances in Intelligent Systems and Computing

Volume 373

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within “Advances in Intelligent Systems and Computing” are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

More information about this series at http://www.springer.com/series/11156

Sigeru Omatu · Qutaibah M. Malluhi
Sara Rodríguez González · Grzegorz Bocewicz
Edgardo Bucciarelli · Gianfranco Giulioni
Farkhund Iqba
Editors

Distributed Computing and Artificial Intelligence, 12th International Conference

Springer

Editors
Sigeru Omatu
Faculty of Engineering
Department of Electronics, Information and Communication Engineering
Osaka Institute of Technology
Osaka
Japan

Qutaibah M. Malluhi
College of Engineering
Qatar University
Doha
Qatar

Sara Rodríguez González
Department of Computing Science
Faculty of Science
University of Salamanca
Salamanca
Spain

Grzegorz Bocewicz
Department of Electronics and Computers
Koszalin University of Technology
Koszalin
Poland

Edgardo Bucciarelli
Dep. of Philosophical, Pedagogical and Economic-Quantitative Sciences
University of Chieti-Pescara
Pescara
Italy

Gianfranco Giulioni
Dep. of Philosophical, Pedagogical and Economic-Quantitative Sciences
University of Chieti-Pescara
Pescara
Italy

Farkhund Iqba
Abu Dhabi Campus
Zayed University
Abu Dhabi
Utd.Arab.Emir.

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-19637-4 ISBN 978-3-319-19638-1 (eBook)
DOI 10.1007/978-3-319-19638-1

Library of Congress Control Number: 2015940025

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Preface

The 12^{th} International Symposium on Distributed Computing and Artificial Intelligence 2015 (DCAI 2015) is a forum to present applications of innovative techniques for solving complex problems in these areas. The exchange of ideas between scientists and technicians from both the academic and industrial sector is essential to facilitate the development of systems that can meet the ever-increasing demands of today's society. The present edition brings together past experience, current work and promising future trends associated with distributed computing, artificial intelligence and their application in order to provide efficient solutions to real problems. This conference is a stimulating and productive forum where the scientific community can work towards future cooperation in Distributed Computing and Artificial Intelligence areas.

Nowadays it is continuing to grow and prosper in its role as one of the premier conferences devoted to the quickly changing landscape of distributed computing, artificial intelligence and the application of AI to distributed systems. This year's technical program will present both high quality and diversity, with contributions in well-established and evolving areas of research. Specifically, 62 papers were submit-ted to the main track from over 18 different, representing a truly "wide area network" of research activity. The DCAI'15 technical program has selected 46 papers, the best papers will be selected for their publication on the the following special issues in journals such as Neurocomputing, Frontiers of Information Technology & Electronic Engineering, Journal of Artificial Intelligence (IJAI), the International Journal of Imaging and Robotics (IJIR) and the International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI). These special issues will cover extended versions of the most highly regarded works.

Moreover, DCAI'15 Special Sessions have been a very useful tool in order to complement the regular program with new or emerging topics of particular interest to the participating community. Special Sessions that emphasize on multi-disciplinary and transversal aspects, such as *AI-driven methods for Multimodal Networks and Processes Modeling* and *Multi-Agents Macroeconomics* have been especially encouraged and welcome.

We thank the sponsors (Indra, INSA - Ingeniería de Software Avanzado S.A., IBM, IEEE Systems Man and Cybernetics Society Spain, AEPIA *Asociación Española para la Inteligencia Artificial,* APPIA *Associação Portuguesa Para a Inteligência Artificial,* CNRS *Centre national de la recherche scientifique*), and finally, the Local Organization members and the Program Committee members for their hard work, which was essential for the success of DCAI'15.

Sigeru Omatu
Sara Rodríguez González
Gianfranco Giulioni
Qutaibah M. Malluhi
Grzegorz Bocewicz
Edgardo Bucciarelli
Farkhund Iqbal
(Eds.)

Organization

Honorary Chairman

Masataka Inoue	President of Osaka Institute of Technology, Japan

Scientific Committee

Sigeru Omatu (Chairman)	Osaka Institute of Technology, Japan
Qutaibah M. Malluhi (Co-Chairman)	Qatar University, Qatar
Adriana Giret	Politechnich University of Valencia, Spain
Alberto Fernández	University Rey Juan Carlos, Spain
Ana Carolina Lorena	Federal University of ABC, Brazil
Ângelo Costa	University of Minho, Portugal
Antonio Moreno	University Rovira y Virgili, Spain
Antonio Manuel de Jesús Pereira	Leiria Polytechnic Institute, Portugal
Araceli Sanchís	University Carlos III of Madrid, Spain
B. Cristina Pelayo García-Bustelo	University of Oviedo, Spain
Bianca Innocenti	University of Girona, Spain
Bogdan Gabrys	Bournemouth University, UK
Bruno Baruque	University of Burgos, Spain
Carina González	University of La Laguna, Spain
Carlos Carrascosa	Politechnich University of Valencia, Spain
Carmen Benavides	University of Leon, Spain
Daniel Glez-Peña	University of Vigo, Spain
David Griol Barres	University Carlos III of Madrid, Spain
Davide Carneiro	University of Minho, Portugal
Dídac Busquets	University of Girona, Spain
Dongshik KANG	Ryukyu University, Japan

Eladio Sanz	Unversity of Salamanca, Spain
Eleni Mangina	University College Dublin, Ireland
Emilio Corchado	University of Burgos, Spain
Eugenio Aguirre	University of Granada, Spain
Eugénio Oliveira	University of Porto, Portugal
Evelio J. González	University of La Laguna, Spain
Faraón Llorens Largo	University of Alicante, Spain
Fernando Díaz	Univesity of Valladolid, Spain
Fidel Aznar Gregori	University of Alicante, Spain
Florentino Fdez-Riverola	University of de Vigo, Spain
Francisco Pujol López	Polytechnic University of Alicante, Spain
Fumiaki Takeda	Kochi Institute of Technology, Japan
Germán Gutiérrez	University Carlos III, Spain
Grzegorz Bocewicz	Koszalin University of technology, Poland
Gustavo Santos	University of Salamanca, Spain
Ivan López Arévalo	Lab. of Information Technology Cinvestav-Tamaulipas, Mexico
Jamal Dargham	University of Malaysia, Saba, Malaysia
Javier Carbó	University Carlos III of Madrid, Spain
Javier Martínez Elicegui	Telefónica I+D, Spain
Jesús García Herrero	University Carlos III of Madrid, Spain
Joao Gama	University of Porto, Portugal
Johan Lilius	Åbo Akademi University, Finland
José R. Villar	University of Oviedo, Spain
Juan A. Botia	University of Murcia, Spain
Juan Pavón	Complutense University of Madrid, Spain
José M. Molina	University Carlos III of Madrid, Spain
José R. Méndez	University of Vigo, Spain
José V. Álvarez-Bravo	University of Valladolid, Spain
Joseph Giampapa	Carnegie Mellon, USA
Juan Manuel Cueva Lovelle	University of Oviedo, Spain
Juan Gómez Romero	University Carlos III of Madrid, Spain
Kazutoshi Fujikawa	Nara Institute of Science and Technology, Japan
Lourdes Borrajo	University of Vigo, Spain
Luis Alonso	University of Salamanca, Spain
Luis Correia	University of Libon, Portugal
Luis F. Castillo	Autonomous University of Manizales, Colombia
Luís Lima	Polytechnic of Porto, Portugal
Manuel González-Bedia	University of Zaragoza, Spain
Manuel Pegalajar Cuéllar	University of Granada, Spain
Manuel Resinas	University of Sevilla, Spain
Marcilio Souto	Federal University of Rio Grande do Norte, Brazil
Margarida Cardoso	ISCTE, Portugal
Maria del Mar Pujol López	University of Alicante, Spain
Masanori Akiyoshi	Osaka University, Japan
Masaru Teranishi	Hiroshima Institute of Technology, Japan

Michifumi Yoshioka	Osaka Prefecture University, Japan
Miguel Delgado	University of Granada, Spain
Miguel Ángel Patricio	University Carlos III of Madrid, Spain
Miguel Rebollo	University of Vigo, Spain
Miguel A. Vega-Rodriguez	University of Extremadura, Spain
Mohd Saberi Mohamad	University of Technology Malaysia, Malaysia
Michael Zaki	University Rostok, Germany
Noki Mori	Osaka Prefecture University, Japan
Norihisa Komoda	Osaka University, Japan
Paulo Novais	Polytechnic University of Minho, Portugal
Pawel Pawlewski	Poznan University of Technology, Poland
Peter Fobrig	University Rostok, Germany
Ichiro Satoh	National Institute of Informatics, Japan
Rafael Corchuelo	Catholic University of Sevilla, Spain
Ramón Rizo	University of Alicante, Spain
René Leistikow	University Rostok, Germany
Ricardo Campello	University of Sao Paulo, Brazil
Ricardo Costa	Polytechnic of Porto, Portugal
Rubén Fuentes	Complutense University of Madrid, Spain
Rui Camacho	University of Porto, Portugal
Sebastian Ahrndt	Technische Universität Berlin, Germany
Sebastian Bader	University of Rostock, Germany
Seiichi Ozawa	Kobe University, Japan
Seiji Ito	Hiroshima Institute of Technology, Japan
Seyedsaeid Mirkamali	University of Mysore, India
Shanmugasundaram Hariharan	B.S.Abdur Rahman University Vandalur, India
Silvana Aciar	University of Girona, Spain
Xinyu Que	IBM Research, USA
Shiro Tamaki	University of the Ryukyus, Japan
Sofya V. Zhukova	St. Petersburg State University, Russia
Toru Yamamoto	Hiroshima University, Japan
Vicente Botti	Politechnich University of Valencia, Spain
Vicente Julián	Politechnich University of Valencia, Spain
Victor J. Sosa-Sosa	Laboratory of Information Technology (LTI) CINVESTAV, México
Viet-Hoang VU	Hanoi University of Technology, Vietnam
Waldo Fajardo	University of Granada, Spain
Worawan Diaz Carballo	Thammasat University, Thailand
Yi Fang	Purdue University, USA
Yusuke NOJIMA	Osaka Prefecture University, Japan
Yutaka MAEDA	Kansai University, Japan
Zbigniew Pasek	IMSE/University of Windsor, Canada

AI-Driven Methods for Multimodal Networks and Processes Modeling Special Session Committee

Chairs

Grzegorz Bocewicz	Koszalin University of Technology, Poland
Izabela E. Nielsen	Aalborg University, Denmark

Co-chairs

Pawełł Pawlewski	Poznan University of Technology, Poland
Zbigniew Banaszak	Warsaw University of Technology, Poland
Peter Nielsen	Aalborg University, Denmark
Sitek Pawełł	Kielce University of Technology, Poland
Robert Wójcik	Wrocłlaw University of Technology, Poland
Marcin Relich	University of Zielona Gora, Poland
Justyna Patalas-Maliszewska	University of Zielona Gora, Poland
Krzysztof Bzdyra	Koszalin University of Technology, Poland

Multi-agents Macroeconomics Special Session Committee

Chairs

Edgardo Bucciarelli (Economist)	University of Chieti-Pescara, Italy
Gianfranco Giulioni (Economist)	University of Chieti-Pescara, Italy

Scientific Committee

José Carlos R. Alcantud (Mathematical economist)	University of Salamanca, Spain
Claudia Casadio (Philosopher of science)	University of Chieti-Pescara, Italy

Shu-Heng Chen (Economist)	National Chengchi University, Taipei, Taiwan
S. Barry Cooper (Mathematical logician)	University of Leeds, United Kingdom
Felix Freitag (Computer scientist)	Universitat Politécnica de Catalunya, Spain
Jakob Kapeller (Economist)	Johannes Kepler University of Linz, Austria
Amin Khan (Computer scientist)	IST, University of Lisbon, Portugal
Alan Kirman (Economist)	Aix-Marseille Université, France
Nicola Mattoscio (Economist)	University of Chieti-Pescara, Italy
Giulio Occhini (Computer scientist)	Italian Association for Informatics and Automatic Calculation, Milan, Italy
Enrico Rubaltelli (Cognitive psychologist)	University of Padua, Italy
Anwar Shaikh (Economist)	The New School for Social Research, New York, United States
Antonio Teti (Computer scientist)	Italian Association for Informatics and Automatic Calculation, Milan, Italy
Katsunori Yamada (Economist)	Osaka University, Japan
Stefano Zambelli (Economist)	University of Trento, Italy

Organizing Committee

Sara Rodríguez (Chairman)	University of Salamanca, Spain
Juan F. De Paz (Chairman)	University of Salamanca, Spain
Javier Bajo	Technical University of Madrid, Spain
Fernando De la Prieta	University of Salamanca, Spain
Gabriel Villarrubia González	University of Salamanca, Spain
Javier Prieto Tejedor	University of Salamanca, Spain
Alejandro Hernández Iglesias	University of Salamanca, Spain
Emilio S. Corchado	University of Salamanca, Spain
Belén Pérez Lancho	University of Salamanca, Spain
Angélica González Arrieta	University of Salamanca, Spain
Vivian F. López	University of Salamanca, Spain
Ana de Luís	University of Salamanca, Spain
Ana B. Gil	University of Salamanca, Spain

Review Process

DCAI welcomed the submission of application papers with preference to the topics listed in the call for papers.

All submitted papers followed a thorough review process. For each paper was be refereed by at least two international experts in the field from the scientific committee (and most of them by three experts) based on relevance, originality, significance, quality and clarity.

The review process took place from January 20, 2015 to February 18, 2015 and was carried out using the Easychair conference tool.

The review process granted that the papers consisted of original, relevant and previously unpublished sound research results related to any of the topics of the conference.

All the authors of those papers requiring modifications were required to upload a document stating the changes included in the paper according to the reviewers' recommendations. The documents, together with the final versions of the papers, were revised in detail by the scientific committee chairs.

Contents

Part I: Main Track

Part I
Main Track

A Natural Language Interface to Ontology-Based Knowledge Bases

Mario Andrés Paredes-Valverde[1], José Ángel Noguera-Arnaldos[2], Cristian Aarón Rodríguez-Enríquez[3], Rafael Valencia-García[1], and Giner Alor-Hernández[3]

[1] Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo, 30100, Murcia, Spain
[2] PROASISTECH, S.L. Edificio CEEIM, Campus Universitario Espinardo s/n, 30100, Murcia
[3] Division of Research and Postgraduate Studies, Instituto Tecnológico de Orizaba, Mexico
{marioandres.paredes,valencia}@um.es, jnoguera@proasistech.com, crodriguezen@gmail.com, galor@itorizaba.edu.mx

Abstract. The aim of the Semantic Web is to improve the access, management, and retrieval of information on the Web-based. On this understanding, ontologies are considered a technology that supports all aforementioned tasks. However, current approaches for information retrieval on ontology-based knowledge bases are intended to be used by experienced users. To address this gap, Natural Language Processing (NLP) is deemed a very intuitive approach from a non-experienced user`s perspective, because the formality of a knowledge base is hidden, as well as the executable query language. In this work, we present ONLI, a natural language interface for DBpedia, a community effort to structure Wikipedia's content based on an ontological approach. ONLI combines NLP techniques in order to analyze user's question and populate an ontological model, which is responsible for describing question's context. From this model, ONLI requests the answer through a set of heuristic SPARQL-based query patterns. Finally, we describe the current version of the ONLI system, as well as an evaluation to assess its effectiveness in finding the correct answer.

Keywords: natural language processing, ontology, semantic web.

1 Introduction

The aim of the Semantic Web is to provide well-defined and understandable information not only by humans, but also by computers, allowing these last ones to automate, integrate and reuse high-quality information across different applications. Ontologies are considered a technology which supports all aforementioned tasks. In the context of computer and information sciences, an ontology defines a set of representational primitives allowing to model a domain of knowledge or discourse [1]. Nowadays, data stored in ontology-based knowledge bases has significantly grown, becoming in an important component in enhancing the Web intelligence and in supporting data

S. Omatu et al. (eds.), *Distributed Computing & Artificial Intelligence, 12th Int. Conference*, Advances in Intelligent Systems and Computing 373, DOI: 10.1007/978-3-319-19638-1_1

integration. Indeed, ontologies are being applied to different domains such as biomedicine [2], finance [3], innovation management [4], cloud computing [5] [6], medicine [7], and human perception [8], among others.

At the same time, Wikipedia, a free-access and free content Internet encyclopedia, has become one of the most popular Web-based knowledge sources, maintained by thousands of collaborators. DBpedia [9] has emerged as a community effort to structure Wikipedia's content and make it available on the Web. Currently, the English version of DBpedia describes 4.22 million of things in a consistent ontology, including persons, places, creative works, organizations, species and diseases.

The most known formal query language for information retrieval from semantic knowledge bases is SPARQL (*Simple Protocol and RDF Query Language*). The use of SPARQL demands a high level of knowledge and expertise about technologies such as RDF (*Resource Description Framework*) and query language expressions. Because of this, several research efforts have been carried out in order to make accessible the Semantic Web information to all kind of users. From this perspective, NLI (*Natural Language Interface*) is deemed to be a very intuitive approach from a user point of view [10] in order to address the gap between knowledge bases systems and end-users. A NLI allows users to access information stored in any repository by formulating requests in natural language [11]. In this work, we present ONLI, a NLI for DBpedia. ONLI processes user's question by means of Natural Language Techniques in order to obtain semantically meaningful words. These elements are queried against knowledge base in order to establish a context. All information obtained is organized in an ontological model, from which the possible ambiguities are managed. Finally, the answer search is carried out, this search is based on a set of heuristic SPARQL-based query patterns. Once possible answers are obtained, these ones are organized and shown back to the user.

This paper is structured as follows: Section 2 presents a set of related works about NLI for knowledge bases. The architecture design, modules and interrelationships of the proposed approach are described in Section 3. Section 4 presents the evaluation of the ONLI system. Finally, conclusions and future work are presented.

2 Related Works

In recent years, several research works concerning to NLI for knowledge bases (NLIKB) have been carried out. These works provide to end-users a mechanism to access knowledge in ontologies, hiding the formality of ontologies and query languages. In this context, there are works focused to query Wikipedia content such as [16] where authors presented a search interface for Wikipedia, which enable users to ask complex queries based on a faceted approach. On the other hand, there are works focused on a specific context such as QACID [12], an ontology-based Question Answering applied to the Cinema Domain. QACID groups user queries into clusters which are associated to a SPARQL-based query. Then, QACID's infer engine deduces between a new query and the clusters in order to associate the query with its corresponding SPARQL-based expression. Other works are focused on portable NLI development. NLP-Reduce [11] identifies triple structures contained in user's question and match them to the synonym-enhanced triple store previously generated from an OWL knowledge base. The identified triples are translated to SPARQL-based statements. Then, NLP-Reduce infers implicit triple statements by using the *Pellet Reasoner*. Finally the results are displayed to the user. NLP-Reduce issue is the limited set of natural language processing techniques used to process question, which limits the identification of triples. ORAKEL [10] is an ontology-based natural language system which works with input queries that contain *wh*-pronouns. The query is parsed to a query logical form, which in turn, is translated to a knowledge representation language. Then, ORAKEL evaluates the query with respect to the knowledge base and presents the results to the user. The main issue of ORAKEL is the need for a lexicon engineer to carry out the adaptation process for a specific domain. Aqualog [13] takes natural language queries and an ontology as input. Then, Aqualog translates queries into ontology-compatible triples. Aqualog takes into account features such as voice and tense in order to facilitate the reasoning about the answer. Aqualog identifies 23 different linguistic categories of intermediate representations of queries. This classification is based on the kind of triple needed. Querix [14] is a domain-independent NLI for the Semantic Web. It processes query and obtains a syntax tree and a query skeleton of word categories such as *Noun*, *Verb*, among others. Then, a synonym search is performed for each element contained in the syntax tree, Querix matches the query skeleton with a small set of heuristic patterns. Next, Querix searches for all matches between synonym-enhanced query, with the resources and their synonyms in the ontology. Finally, Querix composes SPARQL-based queries from the joined triples. SWSNL [15] uses ontologies for storing the domain structure and the description of user queries. SWSNL analyses query by using NLP techniques and then built a semantic and KB-independent representation of query which is translated to a query language.

Most of above presented works were used and tested on different knowledge bases which vary in their size and complexity. Because of this, it is difficult to establish which interface provides the best result. Also, each work employs a different set of natural language processing techniques, in our approach a wider set of these techniques is used in order to obtain a more detailed analysis of the user's question, among these techniques are: POS (Part-of-speech) tagging, lemmatizing, synonym expansion, and NER.

3 ONLI System

In this work, we present a NLI called ONLI, which allows users to obtain answers to questions concerning to DBpedia's content. ONLI system is composed of three main modules: (1) NLP, (2) Question classification, and (3) Answer searching and building. In brief, ONLI system works as follows: User provides a question expressed in natural language. NLP module processes the question with the aim of detecting words potentially belonging to a semantic category. Question classification module queries the obtained elements against DBpedia in order to obtain all semantic information about them, and then, a question's context is established. All this information is organized in an ontological model called *Question model,* whose main purpose is to describe the user's question so that subsequent modules can use this information to search the answer. Next, question is classified based on the information contained in the *Question model*. The Answer Searching and Building module looks for an answer though a set of heuristic SPARQL-based query patterns. Finally, this module filters the results obtained and provides an answer to the user. The ONLI's functional architecture, which is based on the generic architecture for NLI proposed by [17], is composed of three different modules that are described in detail in next sections.

3.1 NLP Module

This module carries out the natural language processing of the question in order to obtain all semantically meaningful of the words for the system. To this end, POS tagging, Lemmatization, and Named Entity Recognition (NER) techniques are executed. POS-tagging technique allows annotating all question elements with their lexical category such as verb, proper noun, and adjective, to mention but a few. Sometimes the word contained in the question is not the same that defines a knowledge entity, e.g. a word could be declared in a different verb time, or in plural. Lemmatization technique provides the base form for a word, known as *lemma*. The *lemma* can be used instead of the word as mentioned in the original question for linking against the knowledge base. The aim of the NER technique is to detect Named Entities [18], i.e. entities belonging to a predefined semantic category such as person names, organizations, locations, among others. This task is performed by using gazetteers which identifies entity names in the text based on lists, and annotation patterns i.e., rules which act on annotations assigned in earlier phases, in order to produce outputs of annotated entities.

Furthermore, NLP module searches for interrogative particles such as *what*, *where*, w*hen*, and w*ho*, among others. These particles will help to determine the answer type expected, thus reducing the search space. Also, NLP module performs a synonym search for each entity found. This task is supported by the lexical database WordNet [19]. The synonym search aims to extend the knowledge entities' search through use of both the entity's lemma and the entity's synonyms.

At the end of NLP phase, there is a set of entities belonging to a specific entity type (*questionType*, *individual*, *objectProperty*, *datatypeProperty*, and *class* or *concept*), which in turn have two main properties: *originalContent*, which refers to the word as it appears in the question, and *lemma*. All these information is mapped to the *Question model*.

3.2 Question Classification Module

Once question has been processed, each entity contained in the *Question model* is queried against DBpedia in order to establish the question's context. According to the entity type, a SPARQL-based query is executed. In this query, entity's lemma and entity's synonyms are compared against *rdfs:label* property of each knowledge entity. The *rdfs:label* property can be used to provide a human-readable version of a resource name. Hence, each class, property and instance, contained in the knowledge base must have the *rdfs:label* property defined, in order to ensure the correct functioning of the ONLI system. It must be noted that the final result, i.e. the answer, depends of quality of these labels. The knowledge entities' search needs to be flexible, i.e., it does not require to retrieve only entities with a high level of similarity because of this, the present module assigns a score to each knowledge entity using the *Levenshtein* distance metric. This score determines the similarity level between the entity's *lemma* and the knowledge entity's label. The knowledge entity with highest *score* probably is the most crucial to find the correct answer. The question's context establishing does not require all information about a knowledge entity, because these entities are only the base for future queries. Based on this understanding, the properties recovered of each knowledge entity are: *URI*, *rdfs:label*, and *rdf:type*.

On the other hand, most questions contain interrogative particles which provide a guide about the desired response, thus reducing the search space. For instance, when the question contains the interrogative particle *who*, user is generally searching for a person or an organization. In the same context, the interrogative particle w*here* makes reference to a location, therefore, the search space can be reduced to knowledge entities belonging to semantic categories such as country, city, and buildings, among others. Taking this into account, this module classifies the question according to the classification of question proposed in [20], furthermore, it determines the answer type expected, which in turn is mapped to the *Question model* as an *answerType* entity. The answer type can be an instance of any class contained in the knowledge base, a set of instances, a date, or a number, among others. All information recovered in this phase is organized in the *Question model*.

3.3 Answer Searching and Building Module

Once question has been processed, and all the entities found have been organized into the *Question model*, the search for an answer is carried out. This module extracts the knowledge entities contained in the *Question model* and it includes each of them in a set of heuristic SPARQL-based query patterns. According to the knowledge entity´s type and question's type, a query pattern is selected. For instance, if the following query is processed: "*How many films has Christian Bale done?*", a total of two knowledge entities are found, *Film* and *Christian Bale* corresponding to the entity types *Concept* and *Individual* respectively. Also, the "*How many*" question type has been identified, therefore, the user is looking for an amount, i.e. the answer expected corresponds to a *number*. Taking into account the aforementioned information, the Answer Searching and Building module include the entities found in the next heuristic SPARQL-based query pattern:

```
SELECT count (distinct ?uri) as ?count
WHERE {
 ?uri rdf:type <http://dbpedia.org/ontology/Film>.
 ?uri ?prop <http://dbpedia.org/resource/Christian_Bale>.
 ?uri rdfs:label ?label.
 FILTER langMatches( lang(?label), 'EN' )
}
```

Finally, the result obtained from the SPARQL-based query above presented, is shown to the user, i.e. according to the DBpedia's content, Christian Bale has been in 34 movies. It should be noted that each pattern used by this module is focused in different contexts in order to retrieve results consistent with the answer type expected.

4 Evaluation and Results

In order to assess the effectiveness of the present work, we conducted an evaluation focused on the capability of ONLI to provide the correct answer to user queries. For this purpose, we performed the study with a set of five students of Computer Science. The students were introduced to the DBpedia knowledge content, and then they were asked to express 20 queries in natural language. The questions used in this study had to meet the following constraints:

1. The query must be able to be answered through DBpedia's content.
2. The questions must begin with *Wh*-pronouns such as *who*, *what*, *where* and *which*, or contain the expressions "*How many*" and "*How*" followed by an adjective.
3. The question must not have orthographic errors.

Examples of the questions generated are: "*What is the height of Lebron James?*", "*Where were Rafael Nadal and Roger Federer born?*", "*What are the movies of Michael Caine?*", and "*How many astronauts born in Spain?*"

Once natural language queries were defined, these queries were executed by using the ONLI system. The answer provided by ONLI was compared to the correct answer provided by the SPARQL-based query previously built by an experienced user on SPARQL and DBpedia. In this evaluation, we employed the *precision* and *recall* metrics, as well as their harmonic mean, known as *F-measure*. These metrics are commonly applied to information retrieval systems and researches in the context of NLI [10].

The results obtained from the evaluation indicated that ONLI system obtained an F-measure score of 0.82. The global experiments results are reported in Table 1. Despite the fact that the results obtained seem promising, there are yet some issues to be solved. On one hand, the DBpedia has full narrative labels, however we cannot guarantee that all knowledge entities contained in this knowledge base are correctly annotated. On the other hand, once both set of queries (natural language questions used on ONLI, and their corresponding SPARQL-based queries manually-built by an expert user) were analyzed, we ascribe the unsuccessfully answered questions to the complexity of the questions generated i.e., some natural language questions contained more than two knowledge entities of different types, which in turn required more than one triple pattern for their representation on SPARQL-based language. Therefore, the set of heuristic SPARQL-based patterns was a bit limited to deal with this kind of questions.

Table 1. Evaluation results

Precision	Recall	F-measure
0.81	0.84	0.82

5 Conclusions and Future Work

The main contribution of this research is a methodology for the development of a question answering system for ontology-based knowledge bases. Specifically, we present ONLI, an ontology-based NLI developed for answering queries related to the DBpedia's content. Although the results obtained seem promising, ONLI needs to be improved in order to provide a system able to return precise answers to questions, since ONLI system provides a set of possible answers instead of a unique answer.

As subsequent work we plan to develop a method that allows carrying out the lexicalization of the DBpedia ontology, i.e. to obtain a set of words that can be found in the user's question, thus improving the identification of potential ontology concepts, which in turn, can be used to build the SPARQL-based query. Also, we plan to enlarge and enrich the set of heuristic SPARQL-patterns by taking into account a wider set of question and answer types. Other ONLI issue is the lack of robustness, because of all input questions must contain *Wh*-pronouns or expressions "*How many*" and "*How*" followed by and adjective. For solving this, we are considering to extend the question's classification and include a wide set of question patterns that allows establishing the question context and the type of answer expected, avoiding the need to use aforementioned words to express a question.

Acknowledgments. Mario Andrés Paredes-Valverde is supported by the National Council of Science and Technology (CONACYT) and the Mexican government.

References

1. Gruber, T.: Ontology. In: Encyclopedia of Database Systems, vol. 5, p. 3748. Springer
2. Ruiz-Martínez, J.M., Valencia-García, R., Martínez-Béjar, R., Hoffmann, A.: BioOntoVerb: A top level ontology based framework to populate biomedical ontologies from texts. Knowl.-Based Syst. 36, 68–80 (2012)
3. Lupiani-Ruiz, E., García-Manotas, I., Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J.T., Camón-Herrero, J.B.: Financial news semantic search engine. Expert Syst. Appl. 38(12), 15565–15572 (2011)
4. Hernández-González, Y., García-Moreno, C., Rodríguez-García, M.Á., Valencia-García, R., García-Sánchez, F.: A semantic-based platform for R&D project funding management. Comput. Ind. 65(5), 850–861 (2014)
5. Rodríguez-García, M.Á., Valencia-García, R., García-Sánchez, F., Samper-Zapater, J.J.: Ontology-based annotation and retrieval of services in the cloud. Knowl.-Based Syst. 56, 15–25 (2014)

6. Rodríguez-García, M.Á., Valencia-García, R., García-Sánchez, F., Samper-Zapater, J.J.: Creating a semantically-enhanced cloud services environment through ontology evolution. Future Gener. Comput. Syst. 32, 295–306 (2014)
7. Rodríguez-González, A., Alor-Hernández, G.: An approach for solving multi-level diagnosis in high sensitivity medical diagnosis systems through the application of semantic technologies. Comput. Biol. Med. 43(1), 51–62 (2013)
8. Prieto-González, L., Stantchev, V., Colomo-Palacios, R.: Applications of Ontologies in Knowledge Representation of Human Perception. Int. J. Metadata Semant. Ontol. 9(1), 74–80 (2014)
9. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P., Hellmann, S., Morsey, M., van Kleef, P., Bizer, C.: DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal (2014)
10. Cimiano, P., Haase, P., Heizmann, J., Mantel, M., Studer, R.: Towards portable natural language interfaces to knowledge bases – The case of the ORAKEL system. Data Knowl. Eng. 65(2), 325–354 (2008)
11. Kaufmann, E., Bernstein, A.: Evaluating the Usability of Natural Language Query Languages and Interfaces to Semantic Web Knowledge Bases. Web Semant. 8(4), 377–393 (2010)
12. Ferrández, Ó., Izquierdo, R., Ferrández, S., Vicedo, J.L.: Addressing ontology-based question answering with collections of user queries. Inf. Process. Manag. 45(2), 175–188 (2009)
13. Lopez, V., Pasin, M., Motta, E.: AquaLog: An Ontology-Portable Question Answering System for the Semantic Web. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 546–562. Springer, Heidelberg (2005)
14. Kaufmann, E., Bernstein, A., Zumstein, R.: Querix: A Natural Language Interface to Query Ontologies Based on Clarification Dialogs. In: 5th ISWC, pp. 980–981 (2006)
15. Habernal, I., Konopík, M.: SWSNL: Semantic Web Search Using Natural Language. Expert Syst. Appl. 40(9), 3649–3664 (2013)
16. Hahn, R., Bizer, C., Sahnwaldt, C., Herta, C., Robinson, S., Bürgle, M., Düwiger, H., Scheel, U.: Faceted Wikipedia Search. In: Abramowicz, W., Tolksdorf, R. (eds.) BIS 2010. LNBIP, vol. 47, pp. 1–11. Springer, Heidelberg (2010)
17. Smith, R.W.: Natural Language Interfaces. In: Brown, K. (ed.) Encyclopedia of Language & Linguistics, 2nd edn., pp. 496–503. Elsevier, Oxford (2006)
18. Grishman, R., Sundheim, B.: Message Understanding Conference-6: A Brief History, Stroudsburg, PA, USA, pp. 466–471 (1996)
19. Miller, G.A.: WordNet: A Lexical Database for English. Commun. ACM 38(11), 39–41 (1995)
20. Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Girju, R., Goodrum, R., Rus, V.: The Structure and Performance of an Open-domain Question Answering System. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 563–570 (2000)

Retweeting Prediction Using Meta-Paths Aggregated with Follow Model in Online Social Networks

Li Weigang and Jianya Zheng

TransLab, Department of Computer Science, University of Brasilia, Brasilia – DF, Brazil
{weigang,zhengdarcy}@cic.unb.br

Abstract. Studying the mechanism of retweeting is useful for understanding the information diffusion in Online Social Networks (OSNs). In this paper, we examine a number of topological features that may affect the retweeting behavior. We apply the *Follow Model* to formulate the user relations and then propose *Relationship Commitment Adjacency Matrix* (*RCAM*) to present the connectivity between users in OSNs. We define three meta-paths to identify the people who may retweet. With these meta-paths, various instance-paths are revealed in the retweeting prediction problem. A framework based on Conditional Random Field model is developed and implemented with the data from Sina Weibo. The case study obtains the results of retweeting prediction with the indices of precision larger than 61% and recall larger than 58%.

Keywords: Follow model, Online Social Network, retweeting prediction, Weibo.

1 Introduction

Online Social Networks (OSNs) have become important social activities for human on the Internet [1-5]. They provide great potential utility for both business marketing and academic research. While business focuses on promoting brand and products, the academic community is interested in user interactions and evolution of network technique. Although the OSN data has great value, how to extract valuable knowledge from the massive data, within a certain time limitation, is still a great challenge.

Many functions have been developed to meet user's demand and facilitate the communication, such as activities *retweet*, *mention*, *like* and *comment* in social networks. The information produced by these activities could open up a new space in the study of human behaviors. Retweeting is one of the most popular functions in OSNs, by which the user could forward the message they concerned about to their followers, relatives and friends. In this way, the information carried by the message can spread widely and quickly. The prediction of retweeting action becomes an intense research area because it can help scientists understand better how the information will propagate in the online social networks.

In this paper, based on the concept of meta-path from the study of heterogeneous network [3], we use *mention* information to help predicting the user's retweeting

S. Omatu et al. (eds.), *Distributed Computing & Artificial Intelligence, 12th Int. Conference*,
Advances in Intelligent Systems and Computing 373, DOI: 10.1007/978-3-319-19638-1_2

action. Also the composite relationship between users which are represented by the *Follow Model* [4,6] are included to make our approach more effective.

From the conventional graph theory, the relationship between the nodes is represented by *Adjacency Matrix* (AM). In this paper, we apply the *Follow Model* to formulate the user relations and then propose Relationship Commitment Adjacency Matrix (RCAM) to present the connectivity between users in OSNs. We define three meta-paths to identify the people who may retweet. The paper extends the basic idea of [7] and aggregates the followship relations in this matrix presentation to define *Relationship Committed Adjacency Matrix* (RCAM): $\boldsymbol{A_{in}}$ as *Follower Adjacency Matrix* and $\boldsymbol{A_{out}}$ as *Followee Adjacency Matrix*.

In OSNs, the multi-constraint queries such as "who, when and what" are important for mining analysis and retweeting prediction. This paper proposes and implements RCAM associated querying and retweeting prediction framework to a real online social network, Sina Weibo in China (weibo.com). The data involves 58.66 million users with 265.11 million followship relations and 369.80 million messages, of which 51.62% were retweets.

2 Related Work

Follow Model was proposed by Sandes et al. [4, 6] in 2012 to describe the following relationship in OSNs, where the terms of followee, follower and r-friends (the users are their follower and followee with each other) can be represented by three functions: $f_{out}(.)$, $f_{in}(.)$ and $f_r(.)$. Given that objects u and v are associated in the vertices of network, the directed edge set E: $V \times V$ represents the edges:

$f_{out}(u)=\{v|(u,v)\in E\}$, is the followee function to present the subset, V^*, of all followees of user u, $V \to V^*$, $V^* \subset V$;

$f_{in}(u)=\{v|(v,u)\in E\}$, is the follower function to present the subset, V^*, of all followers of user u, $V \to V^*$, $V^* \subset V$;

$f_r(u)=f_{out}(u) \cap f_{in}(u)$, is the r-friend function to present the subset, V^*, of all r-friends of user u, $V \to V^*$, $V^* \subset V$.

These functions collectively form the *Follow Model* [4, 6], which has the following three properties: reverse relationship, compositionality and extensibility.

For retweeting prediction, Peng et al. [8] studied the retweet patterns using Conditional Random Fields (CRFs) method [9] with three types of user-tweet features: content influence, network influence and temporal decay factor. They also investigated the approaches to partition the social graphs and construct the network relations for retweeting prediction. Hong et al. [10] analyzed the problem of predicting the popularity of messages as measured by the number of future retweets and shed some light on what kinds of factors influence the information propagation in Twitter.

Sun et al. [3] proposed meta-path to describe the relationship in a heterogeneous network which consists of multi-typed objects and multi-typed relations. The meta-path can be used to define topological features with different semantic meanings.

For example, the connection path between user *u* and *v* in online social network is unique, but the relationship can be various, including following, mentioning, retweeting etc. In order to define the meta-path in this research, we have to identify the target relation in the retweeting prediction. The target relation is the one we wanted in the problem space, and here it is the retweeting relationship. We use $R_{rt} < u, v >$ to represent the action that the user *u* retweets *v*.

To establish the target relationship, we also need the similarity relationship R_{sim} as a bridge to connect the originate and the terminate. Sun et al. [3] listed three general forms of meta-path as topological features.

$u\ R_{sim}\ v\ R_{rt}\ w$, means *u* will R_{rt} *w* if *v* retweets *w* and *u* is similar to *v*.

$u\ R_{rt}\ v\ R_{sim}\ w$, means *u* will retweet *w* if *u* retweet *v* and *v* is similar to *w*.

$u\ R_1\ v\ R_2\ w$, in this case, R_1 and R_2 are some relations undefined and could be defined by the researchers in the practical needs, for example R_1 may be mention etc.

Using meta-path we extended it into OSN with the combination of *Follow Model* to define the relationship between users [7], the detail is described in Section 3.

3 Who May Retweet?

In this section, we establish three meta-paths considering the target and similarity relations to represent who is likely to retweet based on the methods of [7] with necessary extension.

3.1 The User Re-tweets Others He/She is Concerned with

In the situation of “The user re-tweets others he/she is concerned with”, if two users belong to the same category, one may retweet another with a high probability. The formulation of this relation can be represented as follow:

$$v \in f(w) R_{rt}\, w$$

where ***f(.)*** is a function to describe the relationship between *v* and *w* defined in *Follow Model* [4,6], also see section 2. In this case, user *v* is likely to retweet *w*’s tweet. We extend the ***f(.)*** in more detailed cases:

1) $v \in f_{in}(w)\ R_{rt}\ w$. In *w*´s follower subset $f_{in}(w)$, *v* is likely to retweet *w* ´s message.

2) $v \in f_r(w)\ R_{rt}\ w$. In *w*´s r-friends subset $f_r(w)$, *v* is more likely to retweet *w*´s message.

3) $v \in m_{out}(w)\ R_{rt}\ w$. Among the people mentioned by *w*, $m_{out}(w)$, *v* is likely to retweet *w* ´s message after mentioning.

3.2 The User Re-tweets Others He/She is Similar with

The situation of “The user re-tweets others he/she is similar with” can also be found in OSNs. If the similar person of a user retweets a message, then this user may also retweet this message with a high probability. This can be represented as the following:

$$v\, R_{sim}\, f(v) R_{rt}\, w$$