APOLLO: A General Framework for Populating Ontology with Named Entities via Random Walks on Graphs

Wei Shen[1], Jianyong Wang[1], Ping Luo[2], Min Wang[2]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]HP Labs China, Beijing, China
[1]chen-wei09@mails.tsinghua.edu.cn, jianyong@tsinghua.edu.cn
[2]{ping.luo, min.wang6}@hp.com

ABSTRACT

Automatically populating ontology with named entities extracted from the unstructured text has become a key issue for Semantic Web. This issue naturally consists of two subtasks: (1) for the entity mention whose mapping entity does not exist in the ontology, attach it to the right category in the ontology (i.e., fine-grained named entity classification), and (2) for the entity mention whose mapping entity is contained in the ontology, link it with its mapping real world entity in the ontology (i.e., entity linking). Previous studies only focus on one of the two subtasks. This paper proposes APOLLO, a general weakly supervised frAmework for POpuLating ontoLOgy with named entities. APOLLO leverages the rich semantic knowledge embedded in the Wikipedia to resolve this task via random walks on graphs. An experimental study has been conducted to show the effectiveness of APOLLO.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

General Terms

Algorithms, Experimentation

Keywords

Ontology population, Named entity classification, Entity linking, Label propagation

1. INTRODUCTION

Populating the existing ontology with the newly extracted facts become more and more important. Manually populating ontology requires substantial human efforts and is usually time consuming. The development of the information extraction techniques makes the automatic ontology population techniques possible. Integrating the newly extracted knowledge derived from the information extraction systems with the existing ontology requires to deal with the task of populating ontology with named entities.

Ontology population with named entities is the task to locate the right place of the detected named entity in the ontology. Given a named entity mention detected from the unstructured text, if the mapping entity of the entity mention is not contained in the ontology, we should find the right category node to which the entity mention should be attached in the ontology, which is known as the task of fine-grained named entity classification. Otherwise, if the mapping entity of the entity mention exists in the ontology, the aim of this task is to link this detected entity mention with its corresponding real world entity in the ontology, which is known as the entity linking task. In this paper, we propose APOLLO, a novel graph-based framework to resolve the task of automatic ontology population with named entities integrally. APOLLO is a weakly supervised framework that requires minimal human involvements. Moreover, APOLLO is open-domain as it is independent of the underlying ontology.

Copyright is held by the author/owner(s).
WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. ONTOLOGY POPULATION WITH NAMED ENTITIES

The only input of our framework APOLLO is a collection of documents and an initial ontology. Let $D = \{d_1, d_2, \ldots, d_m\}$ be the collection of the input documents and Ω be the initial ontology. Let ζ be the set of all entity mentions recognized from the document set D, and each entity mention $s \in \zeta$ needs to be populated into the ontology Ω. Let N_Ω denote the set of all named entities contained in the ontology Ω, and C_Ω be the set of all categories in the taxonomy of Ω. An entity mention $s \in \zeta$ is a token sequence which refers to some named entity in the text document. We define the document context η_s of the entity mention $s \in \zeta$ as a window of words around the occurrence of the entity mention s. On the other hand, for each named entity $n \in N_\Omega$, we define the document context η_n of the named entity n as the description context for n in the ontology. As both the entity mention $s \in \zeta$ and the named entity $n \in N_\Omega$ have document contexts, we use η to denote the document context corresponding to an entity mention or a named entity. To capture the semantic information existing in the document context η, we recognize all the Wikipedia concepts γ appearing in η, and consider the set of these detected Wikipedia concepts as the *semantic signature* δ. For the general textual document, we utilize the open source toolkit Wikipedia-Miner[1] to detect the Wikipedia concepts appearing in the context.

Based on the definitions, we propose a framework called APOLLO, to address the task of ontology population with named entities using three modules as follows:

[1]http://wikipedia-miner.cms.waikato.ac.nz/

Graph Creation: This module constructs a graph $G = (V, E, W)$ where the nodes V come from all the entity mentions ζ, the named entities N_Ω and the Wikipedia concepts in their *semantic signatures* δ. Specifically, for each entity mention $s \in \zeta$, we pair it with each Wikipedia concept $\gamma \in \delta_s$ where δ_s denotes the *semantic signature* of s, to create the triple (s, γ, w), and the weight w is set to 1.0 in the experiments. For each triple (s, γ, w), s and γ are added to V and the edge (s, γ) is added to E, with $W(s, \gamma) = w$. And for each named entity $n \in N_\Omega$, we also pair it with each Wikipedia concept $\gamma \in \delta_n$ where δ_n denotes the *semantic signature* of n, to create the triple (n, γ, w), where the weight w is set to 1.0 in the experiments as well. For each triple (n, γ, w), n and γ are added to V and the edge (n, γ) is added to E, with $W(n, \gamma) = w$. To forward the label information over the graph more effectively, the semantically related Wikipedia concept nodes should be connected by some edges to enrich the information propagation paths. In order to measure the strength of the semantic relatedness, we adopt the Wikipedia Link-based Measure (WLM) described in [3] to calculate the semantic relatedness between Wikipedia concepts. For each pair of Wikipedia concept nodes (γ_1, γ_2) in the graph, if the semantic relatedness $SR(\gamma_1, \gamma_2)$ is greater than some threshold τ, we add an edge (γ_1, γ_2) to E, with $W(\gamma_1, \gamma_2) = SR(\gamma_1, \gamma_2)$.

Label Propagation: In this module, we assign each entity mention $s \in \zeta$ to the proper category $c_s \in C_\Omega$ via graph label propagation. Firstly, we annotate each named entity node $n \in N_\Omega$ with its corresponding category label in the graph G. In this paper, the named entity category is used as the label for the node, and we assume that each named entity just belongs to one category for the purpose of simplicity. We then combine the two interpretations (i.e., adsorption via averaging and adsorption via random walks) of the Adsorption label propagation algorithm introduced in [1] and apply it to the graph G. For each entity mention $s \in \zeta$, we obtain the label distribution L_s over the categories C_Ω and consider the category which has the largest distribution in L_s as the predicted category c_s for the entity mention s.

Linking Validation: For each entity mention s, we validate whether its mapping entity $n_s \in N_\Omega$ in this module. Given an entity mention s, we firstly retrieve the set of entities that may be referred by this entity mention s, and we denote this set of entities as the candidate entity set CN_s. We build a dictionary DT by leveraging some useful features of the Wikipedia, such as the entity page, the redirect page, the disambiguation page and the hyperlink in Wikipedia article. The dictionary DT is a <key, value> mapping, where the column of the key K is a list of entity mentions and the column of the mapping value $K.value$ is the set of named entities which are referred by the key K. For each entity mention $s \in \zeta$, we look up the dictionary DT and search for s in the column of the key K. If a hit is found, i.e., $s \in K$, we add the set of the mapping entities $s.value$ to the candidate entity set CN_s. Suppose that the entity does not have the same name entity which belongs to the same category. Thus, if there exists some entity $n \in CN_s$ whose category is also c_s, the same category as the predicted category for the entity mention s, then we can predict that this entity n is the mapping entity n_s of the entity mention s; Otherwise, we can predict that the mapping entity of the entity mention s does not exist in the ontology Ω, that is to say, $n_s \notin N_\Omega$.

Table 1: Experimental results over the data set

	# of mentions	APOLLO		*Ganti-KDD$_{op}$*	
		Accu.	#	Accu.	#
All	1033	**0.7764**	**802**	0.5489	567
Unlinkable	661	**0.7534**	**498**	0.7247	479
Linkable	372	**0.8172**	**304**	0.2366	88

3. EXPERIMENTS

In the experiments, we used the May 2011 version of Wikipedia and YAGO(1)[2] of version 2009-w10-5. We firstly chose 20 categories which are the subclasses of the *person* category from YAGO, and randomly selected at most 200 instances for each selected category by querying the YAGO ontology. The created data set consists of 3304 distinct instances belonging to the 20 categories in total. Then we randomly sampled 80% of these instances (i.e., 2643) as the list of named entities contained in the ontology, and the remaining 661 instances are regarded as the test entity mentions, which are all unlinkable. In addition, we collected 372 test entity mentions which can be linked with the named entities existing in the ontology by querying the names with Google, and added them to the test entity mentions.

We added the Linking Validation module of APOLLO to the method proposed in [2] to create the baseline method *Ganti-KDD$_{op}$*. We used the parameters for the baseline according to the original experimental setting in [2]. For APOLLO, to generate the *semantic signature* for each entity, we used its corresponding entire entity page in Wikipedia as the document context. In the Label Propagation module, the number of iterations for the Adsorption algorithm is set to 10. We adopted the evaluation measure *Accuracy* (Accu.) to evaluate the performance of APOLLO and *Ganti-KDD$_{op}$*.

The experimental results of APOLLO and *Ganti-KDD$_{op}$* over the data set are shown in Table 1. Besides the number of total mentions, we also show the accuracy and the number of correctly assigned entity mentions for both APOLLO and *Ganti-KDD$_{op}$*, according to the different types of the test entity mentions. From the results in Table 1, we can see that APOLLO achieves significantly higher accuracy compared with the baseline method *Ganti-KDD$_{op}$* in all aspects.

4. ACKNOWLEDGMENT

This work was supported in part by National Basic Research Program of China (973 Program) under Grant No. 2011CB302206, National Natural Science Foundation of China under Grant No. 60833003, and an HP Labs Innovation Research Program award.

5. REFERENCES

[1] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravich, and R. M. Aly. Video suggestion and discovery for youtube: Taking random walks through the view graph. In *Proceedings of WWW*, pages 895–904, 2008.

[2] V. Ganti, A. C. König, and R. Vernica. Entity categorization over large document collections. In *Proceeding of SIGKDD*, pages 274–282, 2008.

[3] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of WIKIAI*, 2008.

[2] http://www.mpi-inf.mpg.de/yago-naga/yago/