International Journal of Web Portals
Volume 10 • Issue 1 • January-June 2018

Concept Identification Using Co-Occurrence Graph

Anoop Kumar Pandey, Centre for Development of Advanced Computing, Bangalore, India

ABSTRACT

In a community setting, Utilitarian Knowledge or "Knowledge that works" are routinely diffused through social media interactions. The aggregation of this knowledge is a divergent process, where common knowledge gets segregated into several local worlds of utilitarian knowledge. To capture and represent this knowledge, several data models have been proposed. One of the model organizes concepts (atomic elements) in a hierarchy namely concept hierarchy ("is-a") in which concepts are added manually at the most appropriate level inside the hierarchy. To minimize manual intervention in entity resolution, this article proposes entity resolution based on co-occurrence graph and continuous learning, thereby eliminating the bottleneck of manual concept entry. While traditional Supervised Learning methods require sufficient training data beforehand which is not available in a community setting at start, Continuous Learning method could be useful which can acquire new behaviours and can evolve as the community data evolves.

KEYWORDS

Co-Occurrence, Concept Hierarchy, Containment Hierarchy, Continuous Learning, Entity Resolution

INTRODUCTION

In on-line communities, several netizens interact and exchange or share large amount of knowledge among themselves. There are several types of knowledge that they share. In this paper we focus on two types of knowledge elements: Encyclopaedic or Informational Knowledge (knowledge that informs) & Utilitarian Knowledge that can be put to use on a daily basis. Encyclopaedic knowledge is primarily meant for informational purposes. Its objective is to have members of the community agree upon a common world-view, that is aggregated from individual world-views each of which, may be incomplete or have conflicting views of underlying issues. They can be common beliefs or world models. For example, in a university context, a knowledge fragment like "The Spring Semester begins on the first Monday of January" may refer to a knowledge fragment that is commonly held by all members of the university. While in the same university setting, knowledge about the next faculty meeting, the decisions taken in the last meeting, the different points raised by members on an issue, etc. form utilitarian knowledge. For maintaining Encyclopaedic knowledge many ready-made ontologies like DBpedia (http://wiki.dbpedia.org) are available, however there is no such knowledge base to capture Utilitarian Knowledge, since it is very context specific and there are too many contexts. Some generic ontologies are available but they need to be adapted to a particular context. The aggregation of utilitarian knowledge is a divergent process, where common knowledge gets segregated into several local worlds of utilitarian knowledge. If the community as a whole is coherent, these different worlds end up denoting different aspects of the community's dynamics. For example, the community "Cancer Patients in Bangalore" will start with generic information about

DOI: 10.4018/IJWP.2018010103

"Cancer" and then move on to different utilitarian world like "Hospitals in Bangalore for Cancer", "Current advancements in Cancer diagnostics", "Latest services and medical facilities available in Bangalore for Cancer Patients" and so on. Utilitarian Knowledge is subjective in nature. A piece of knowledge may be utilitarian or not, based on who is consuming it and what kind of use they have in mind. They also require much finer privilege management than encyclopaedic knowledge.

To capture and represent this knowledge, several data models have been proposed. One of the model called Many Worlds on a Frame (MWF), as discussed in (Srinivasa, 2012), contains several concepts organized in hierarchies. Concept in general, could refer to all the terminologies and vocabulary of a particular domain which is used to describe it. The definition of concepts and relationship between the concepts is typically captured using a simple structure called concept tree that captures two kinds of relationships: 'is-a' and 'is-in'. For instance, the concepts 'Java' is-a 'Programming Language' depicts 'is-a' relationship while "Bangalore" is-in "Karnataka" depicts 'is-in' relationship. These hierarchies contain many concepts beyond the encyclopaedic concepts which are relevant to that domain. Therefore, in such kind of knowledge base, concepts need to be added manually and hence poses a bottleneck in evolution of the knowledge base. A need for an entity resolution system, that could, given a concept, label it with proper 'is-a' parent thereby placing it appropriately in the concept hierarchy, could help the knowledge base to grow with minimal manual intervention. There are various methods for entity resolution based on supervised learning. Supervised Learning is the process of learning a mapping from examples to labels, by inferring labelled examples provided as input. The model tries to identify the correlation between various label/classes. It then tries to classify an untagged document called as the test dataset. While conventional Supervised Learning methods require sufficient training data beforehand which is not available in a community setting at start, Continuous Learning method (explained later) could be useful which can acquire new behaviours and can evolve as the community data evolves. To minimize manual intervention in entity resolution, this paper proposes entity resolution based on co-occurrence graph and continuous learning, thereby eliminating the bottleneck of manual concept entry.

MOTIVATION AND RELATED WORKS

There are several techniques to achieve named entity recognition (NER) in free text. Among them, dictionary based techniques (Chandel, Nagesh, & Sarawagi, 2006) are one of the most obvious and implementation friendly. In this approach, real world entities from different contexts are manually added into a list segregated based on the type. For example, terms like "Bangalore" and "Delhi" are placed under 'City' type and "India" is placed under the 'Country' type and 'City' and 'Country' are both sub-types of 'Place'.

Theoretically we could construct a dictionary of all the entities and NER can be reduced to a dictionary match, but practically it is infeasible in terms of constructing such a dictionary. The biggest problem with dictionary-based approaches is the unknown entity recognition problem. These algorithms can't recognize entities which are not present in the dictionary and assume the dictionary to be sacrosanct. A dictionary can also not be distinguish between the various contextual roles an entity plays; for example, "Sachin Tendulkar" can either be a 'cricketer' or a 'politician' or both but can't be types as a 'cricketer' in one document and 'politician' in another by a dictionary based algorithm.

Probabilistic models like Hidden Markov Models (HMMs) (Rabiner, 1989) and Conditional Random Fields (CRFs) (Lafferty, McCallum, & Pereira, 2001) address entity recognition by modeling the distribution of entities belonging to a particular type. Significant amount of manually tagged data is observed to construct a suitably trained model that represents the patterns of various entity types. When they encounter an entity in the text, they compare the distribution of other entities in its proximity with that of the distributions of entity types present in the trained model.

This solves the problem of unknown entity recognition to a certain extent as there is no set of predefined entities. The location of the extracted entity becomes key in guessing which class the