Ontology Learning from Text

Mangala Gowri Krishnamoorthy(260511307)

ABSTRACT

An ontology formally represents knowledge as a set of concepts within a

domain, and the relationships among those concepts. The process of defining and

instantiating a knowledge base is referred to as knowledge markup or ontology

population. Machine Learning techniques are extensively used to infer meaningful relationships among the data and build a logically coherent set of formal vocabulary. Of particular interest and the focus of this paper is to study the efforts made towards developing an ontology for web based content gathering ie knowledge acquisition from text. A popular method of inducing taxonomies from textual data is to use hierarchical clustering methods for synonym extraction and term clustering.

This project aims to perform a comparative study of methods that are being used for ontology engineering especially in the context of text based inferences. We also try to find available implementations of these algorithms and try to quantify them in terms of their varying degrees of success.

KEYWORDS

Ontology, Knowledge representation, Knowledge base, Hierarchical clustering, Semantic tagging, Semantic web

1. INTRODUCTION

Knowledge representation, inspite of being a central concept of AI (Artificial Intelligence) has no well proven or universally accepted definition. The concept can best be understood in terms of the roles that it plays under different circumstances. KR is at best a surrogate - ie while designing a system that senses its environment and reacts to it intelligently, we come across an important, inescapable fact- reasoning is a process that goes on internally, while most things it wishes to reason about exist only externally. Thus KR can be visualised as a surrogate for the outside world that an agent uses to reason and draw conclusions from. Modelling the external world is no trivial task. We have to abstract out the unnecessary details and figure out how and what to see in the world. KR consists in making a set of ontological commitments ie, an answer to the question: In what terms should I think about the world? An ontology is thus required for a compact and refined view of a specific subset of the world. In other words, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents.

A common ontology defines the vocabulary with which queries and beliefs are exchanged among agents. Ontological commitments are agreements to use the shared vocabulary in a coherent and consistent manner. The agents sharing a vocabulary need not share a knowledge base; each knows things the other does not, and an agent that commits to an ontology is not required to answer all queries that can be formulated in the shared vocabulary. Ontologies can be written in a variety of ways (logical rules, programming code etc). The important thing here is not the method used but the content of these ontologies and their application in real world problem solving scenarios.

Recently, a lot of research is focused on developing an ontology for web based content gathering ie knowledge acquisition from text. This presents a problem of enormous complexity due to the heterogeneity of data available and the sheer amount of content present on the web. A popular method of inducing taxonomies from textual data is to use hierarchical clustering methods for synonym extraction and term clustering.

The rest of this paper is organized as follows. Section 2, describes the basic concepts and approaches used. Section 3 describes the various tools that have been used for the experiment. Section 4 describes the experimental setup and the data that it has been tested on. Section 6 describes the findings and results. Section 7 contains the conclusion and possible future improvements. Section 8 presents the conclusion and future enhancements.

2. BASIC CONCEPTS AND APPROACHES

These are some of the fundamental concepts used in this experiment.

2.1. HIERARCHICAL CLUSTER LABELING

Clustering, in Machine Learning, is a useful technique for grouping data points such that points in a single cluster have similar characteristics (or are close to each other) as compared to data points in other clusters. Traditional clustering methods include K-means and K-medoids. K-means algorithm represents each cluster by a single mean vector that is then used to classify the data. This method is very sensitive to noise and outliers as they could distort the entire classification. K-medoids, on the other hand, chooses the most representative point in a cluster and uses it to further classify the data. In both these methods, one has to specify the number of clusters beforehand.

*Hierarchical clustering* (or *hierarchic clustering* ) outputs a hierarchy, a structure that is more informative than the unstructured set of clusters returned by flat clustering. Hierarchical clustering does not require us to pre specify the number of clusters and most hierarchical algorithms that have been used in IR are deterministic. These advantages of hierarchical clustering come at the cost of lower efficiency. Hierarchical clustering algorithms are mainly classified as agglomerative methods and divisive methods. Agglomerative clustering starts with N clusters and each of them includes exactly one object. A series of merge operations are then followed out that finally lead all objects to the same group. Divisive clustering proceeds in an opposite way. In the beginning, the entire data set belongs to a cluster and a procedure successively divides it until all clusters are singleton clusters. There has been many traditional Hierarchical clustering algorithms, such as BIRCH,CURE, ROCK, Chameleon and so on.

2.2.TEXT AND SEMANTIC EXTRACTION

Text documents are one of the means to store information. These documents can be found on personal desktop computers, intranets and on the Web. Thus, knowledge is often present in an unstructured and cluttered form. Having an automated system that can extract these information from the texts is very desirable. However, the major challenging issue in developing such an automated system is that no natural language is free from ambiguity and uncertainty problems. Thus till this day, semantic extraction remains a challenging task to researchers in this area.

Typical text mining tasks include text categorization, text clustering, concept/entity and fact extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling. Entity extraction is used to identify people, companies, organizations, cities, geographic features, and other typed entities within HTML pages and text documents/content. Keyword / Term Extraction refers to the task of extracting important terms and "topic" keywords from text documents/content. Advanced statistical and linguistic algorithms analyze the content, "tagging" it with the most important words and phrases. Sentiment Analysis is another interesting task that can be performed . It basically aims to identify positive, negative and neutral sentiment within HTML pages and text documents. Sentiments can be of various types such as document-level sentiment, user-targeted sentiment, entity-level sentiment, and keyword-level sentiment is provided. One of the most important tasks of performing semantic extraction is Relation Extraction that identifies facts and Subject-Action-Object relations within HTML pages and text documents/content. The main aim of semantic data extraction is the creation of a Semantic Web that enables machines to interpret, combine, and use data on the Web. Whereas the current “eyeball” Web is only understandable for humans, the Semantic Web can be used by computers as well. The basis for the Semantic Web are computer-understandable descriptions of resources. We can create such

descriptions by annotating resources with metadata, resulting in “annotations”

about that resource.

4. TOOLS USED

To perform the task of synonym extraction from textual data, I used 4 popular tools that use the technique of clustering as described above. These tools take as input a text document or a html page and produce an output that is in machine readable format. All the tools first extract relevant keywords using clustering methods. They then adopt various techniques to identify relations and other meaningful information. This section provides a brief description of the algorithms and their specific issues.

4.1. Alchemy API

AlchemyAPI is a cloud-based text mining platform providing semantic tagging. AlchemyAPI utilizes natural language processing technology and machine learning algorithms to analyze content, extracting semantic metadata: information about people, places, companies, topics, facts & relationships, authors, languages etc. The API endpoints are provided for performing content analysis on Internet-accessible web pages, posted HTML or text content. The API adopts the approach of SA (Semantic Tagging) to perform its task. Semantic tagging is a process of annotating the data with other metadata like names, attributes, comments, descriptions etc. Compared to tagging, Semantic tagging goes a step further as it enriches the unstructured or semi-structured data with a context that is further linked to the structured knowledge of a domain. It allows results that are not explicitly related to the original search. SA helps to bridge the ambiguity between natural language and a formal machine representation by telling the computer how data items are related and the context in which they appear.

4.2. OpenCalais

OpenCalais is a web service that automatically creates rich semantic metadata for the content that is submitted. Using natural language processing (NLP), machine learning and other methods, Calais analyzes the document and finds the entities within it. But, Calais goes well beyond classic entity identification and returns the facts and events hidden within the text as well. The web service is free for commercial and non-commercial use. It considers the entire document as a dump of data and again uses the technique of semantic tagging (SA) to extract meanings from text.

4.3. Thewikimachine

Thewikimachine is an application that uses SA to add semantic tags to all relevant terms in the text and also other named entities such as persons, locations and organisations. Using DBpedia as knowledge base, it captures cross references to the same concepts even when the documents are in different languages. This application does not need long texts to find relevant tags, just one or two sentences are typically sufficient. It is ideal for analysing comments and texts uploaded in Social Web sites. This becomes particularly useful while performing sentiment analysis for a large group of people.

4.4. Zemanta

Zemanta is a content suggestion engine that analyzes user-generated content (e.g. a blog post) usingnatural language processing andmachine learning techniques to suggest pictures, tags and links to related articles. It performs assisted online content production for any web user based on a blog, an article or a web resource. It was written mainly for bloggers who write about similar topics. Zemanta also acts as a writers social networking tool by discovering similar blogs and cross recommending them to each other, helping them to express affection and to build credibility in their peer group. Zemanta claims to be the only service that connects many well known databases in a single-point solution for discovering more context of the content. It helps to leverage the power of the databases by collecting more structured metadata about the content to build semantics-powered smart apps.

5. EXPERIMENTAL SETUP

To test the efficiency of these 4 tools and to evaluate their correctness, I used various kinds of data (text documents and html pages) and attempted to observe the accuracy of the results. The first data was a simple and short text file that talked about a scientific topic. I then attempted to use another similar but larger piece of text about the same subject matter to see if the results are similar. Next, I used random web pages (html pages) as input to see whether the algorithms were scalable or not. Finally, I tested the programs by using an abstract piece of text from a classical literature book as input.

It is interesting to note that there is no specific data set that must be used to test these programs. They must work well for any data as long as it falls within the domain of the knowledge base being used. This is important as the labels of the clusters are already defined in the ontology. The algorithm’s job is just to parse through the text infer its topic, important phrases and relations connecting these phrases. While sounding like a trivial problem, it is often a humongous task to figure out the various possible combinations of meanings and contexts that could occur within a single short piece of text. Therefore one has to be very careful while selecting the ontology one wishes to work with as the efficiency of the algorithm to a large extent depends on how rich the underlying knowledge base is and how well the ontology is designed. This however results in a trade-off between a rich and expressive ontology that captures various nuances of the underlying knowledge base and a simple to use and understand ontology that doesn’t have lots of complicated relations and hierarchies to deal with. The choice between the two sets of ontologies is application dependent and is left for the user to decide.

6. RESULTS AND FINDINGS AND DISCUSSION

When the above data were run on the 4 different programs, the following observations were made.

- All algorithms extracted topics and clusters accurately ie they performed Domain terminology extraction with a high degree of success. Looking at the algorithms individually,

AlchemyAPI - No confidence level provided about the extracted tags

Calais - Confidence measure not completely relevant

Thewikimachine - Random tags presented(hyperlinks)

Zemanta - Most accurate. Extracted relevant tags

For a sample text, the results looked as follows:

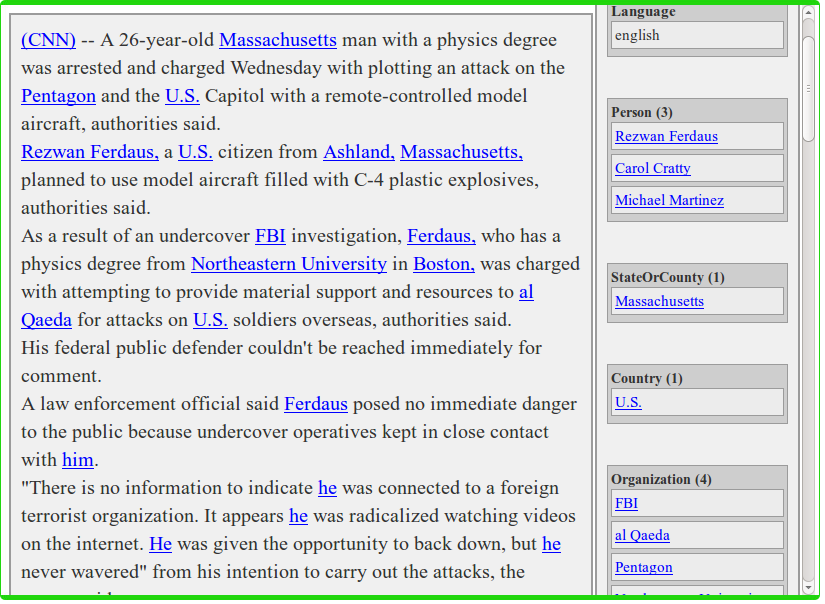


Figure 6.1: Alchemy API



Figure 6.2 : Calais API

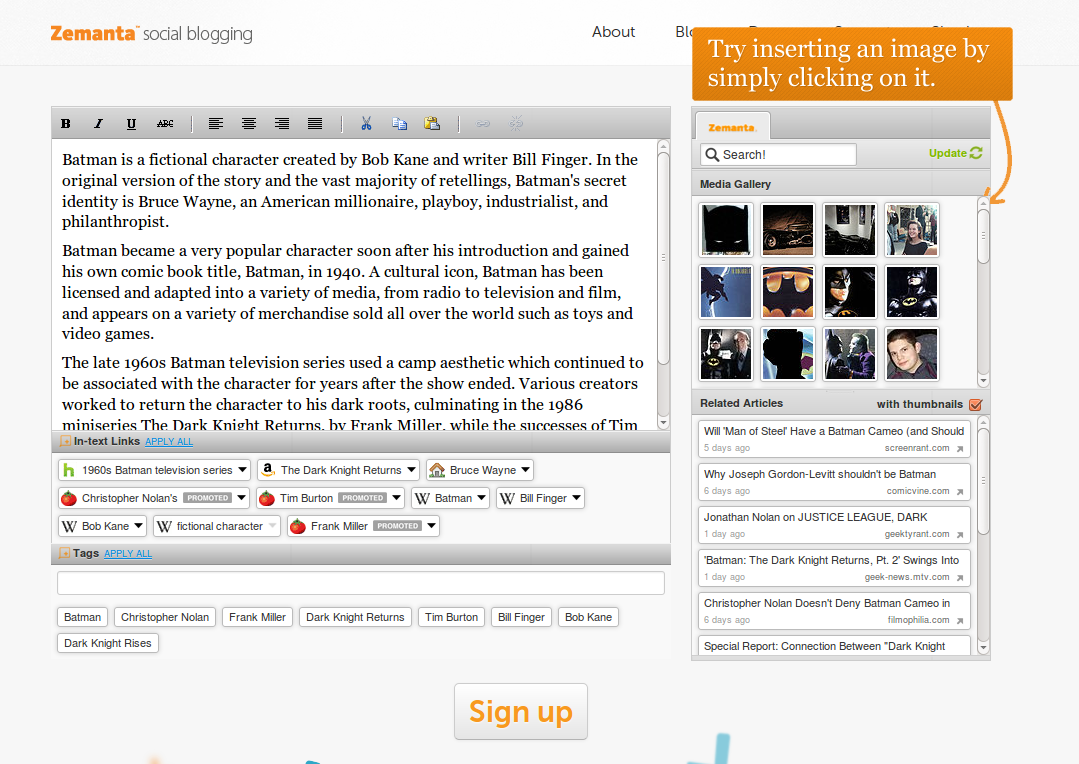


Figure 6.3 : Zemanta

\* While all the applications extracted the relevant keywords in a piece of text and managed to link them appropriately to other sources on the web, they could not distinguish between the most relevant tags and the secondary ones ie they could not identify the main topic accurately.

\* The programs work better for scientific data that have a specific topic, are well structured and are mainly about facts and other verifiable information.

\*When presented with an abstract piece of text, the algorithms falter and are unable to identify the key topic, the context and the relations existing among them.

\* When it comes to the second part of finding the context of the text and inferencing relationships, the algorithms fail to perform convincingly and produce output that may not be relevant.

One reason for this is that the Concept Discovery and Rule discovery algorithms use previously defined rules about the sentence structure (subject, verb and object) to derive relations. In this case, the program will output that subject performs verb on object or subject is related to object through verb. The algorithm tries to find subject, object pairs in the text and the map them. But this simple rule is not sufficient to infer the complicated relationships that may exist in the data as the mapping of words to concepts is ambiguous. Humans use context to disambiguate various meanings of a given piece of text. Machine translation systems cannot easily infer context. And thus, the extracted relations need to be evaluated by an ontologist for relevance and correctness.

7. CONCLUSION

Despite the large number of research works devoted to the ontology learning task, there are still many issues that are not well-covered. Among these issues are the problem of knowledge integration and the lack of deep semantic analysis approaches. Also, many extracted structures are just relational triples that are more lexically-based than conceptual in nature. Another concern is related to the accuracy of the syntactic parses, especially when parsers are confronted with the diversity of Web texts. Despite major advances in the field, one cannot count on entirely correct parses. Semantic analysis systems should then incorporate mechanisms for filtering and checking the accuracy of the syntactic aspects and the semantic aspects. One way to do this is to rely on redundancy across multiple texts and the Web. Other approach might be to repair the inconsistencies in the learned ontologies or provide a mechanism for their easy update and modification. The ultimate aim of ontology learning is the creation of a Semantic Web. By the inclusion of semantic content in web pages, the Semantic Web aims to convert the cluttered and completely random data present on the web into a meaningful and richly connected ‘web of data’ that will not only improve user search but also facilitate for unsupervised learning for intelligent agents.

REFERENCES

[1] Alchemyapi [www.alchemyapi.com/api/entity/](http://www.alchemyapi.com/api/entity/)

[2] Opencalais. [www.opencalais.com/](http://www.opencalais.com/)

[3] Zemanta api. <http://developer.zemanta.com/>

[4] Thewikimachine [www.thewikimachine.fbk.eu/](http://www.thewikimachine.fbk.eu/)

[5] Ontology Learning and Reasoning —Dealing with Uncertainty and Inconsistency

Peter Haase, Johanna Volker

[6] A Multi-Relational Hierarchical clustering algorithm based on shared nearest neighbour similarity : Jing-Feng Guo, Yu-Yan Zhao, Jing Li

[7] Automatically Labeling Hierarchical Clusters : Pucktada Treeratpituk, Jamie Callan

[8] Learning Taxonomies from Multi-Relational Data via Hierarchical Link-Based Clustering : Maximilian Nickel**,** Volker Tresp

[9] Topic Extraction from Messages in Social Computing Services Determining the number of Topic Clusters : Basabi Chakraborty, Takako Hashimoto