Integrating Linked Open Data with Unstructured Text for Intelligence Gathering Tasks

Archit Gupta
CSE Department
Indian Institute of Technology
New Delhi 110016
cs1080164@cse.iitd.ernet.in

Krishnamurthy Koduvayur Viswanathan
CSEE Department
University of Maryland
Baltimore County
Baltimore, MD 21250
krishna3@umbc.edu

Anupam Joshi
CSEE Department
University of Maryland
Baltimore County
Baltimore, MD
joshi@cs.umbc.edu

Timothy Finin
CSEE Department
University of Maryland
Baltimore County
Baltimore, MD
finin@cs.umbc.edu

Ponnurangam Kumaraguru
Indraprastha Institute of
Information Technology
New Delhi, India
pk@iiitd.ac.in

ABSTRACT

We present techniques for uncovering links between terror incidents, organizations, and people involved with these incidents. Our methods involve performing shallow NLP tasks to extract entities of interest from documents and using linguistic pattern matching and filtering techniques to assign specific relations to the entities discovered. We also gather more information about these entities from the Linked Open Data Cloud, and further allow human analysts to add intelligent inference rules appropriate to the domain. All this information is integrated in a knowledge base in the form of a graph that maintains the semantics between different types of nodes involved in the graph. This knowledge base can then be queried by the analysts to create actionable intelligence.

Categories and Subject Descriptors

H.m [**Information Systems**]: Miscellaneous

General Terms

Information Integration

Keywords

Terror networks, information integration, linked open data, intelligence gathering

1. INTRODUCTION

Governments often focus a lot of attention on how intelligence is collected and analyzed in order to protect the country from future terror threats. One of the aspects of this task that is gaining a lot of attention is finding, filtering, and acquiring information from publicly available sources such as news reports, and articles on the Internet. Such artifacts are then analyzed to produce actionable intelligence. This involves integrating the information so acquired with that from other proprietary/classified sources, and with the background knowledge that analysts have from experience. A similar task is performed in the context of business intelligence by large companies. Past work in this area has focused on performing social network analysis on media reports of terrorist violence to uncover links between terrorist organizations. While such methods are likely to be useful for strategic analysis, they have a tendency of generating a lot of false positives.

The emphasis in intelligence gathering tasks generally lies in providing a human analyst with important information gathered and refined from a large corpus of unstructured text (e.g. web pages) which has a significant amount of information, only some of which is relevant to the task at hand. This extracted information then needs to be integrated with information from other sources that an analyst might have, and their background knowledge. The integration is ideally done at a semantic level so that information not just directly present in the text, but that which can be inferred from it, can be used to assist in the integration task.

Elements of this task, especially in the form of populating back-end knowledge bases, have been well studied recently. The Knowledge Base Population (KBP) at the Text Analysis Conference (TAC) and Automated Content Extraction (ACE) tasks coordinated by National Institute of Standards and Technology (NIST) [1] have provided test corpora. However, the target of such exercises have been populating Wikipedia Infobox like structures. In contrast, we extract the knowledge from open sources into an ontology, to better leverage the semantics in the integration

[1] http://www.nist.gov/

process. We focus on the domain of national intelligence, and seek to extract information about terrorist entities and terrorist caused events. We populate a knowledge base with instances of the Semantic Web Technology Evaluation Ontology (SWETO) [1], and store the resulting Resource Description Framework (RDF) [2] triples. In the next step, we get additional data by integrating information from Linked Open Data [3] sources. In particular, for every entity we discover from the text sources, we pull in information from the SPARQL Protocol and RDF Query Language (SPARQL) endpoint of DBpedia. DBpedia, which is a community effort to capture Wikipedia knowledge into RDF format, contains 670 million triples. The result is a graph whose nodes represent persons, organizations, events etc. and whose links have a member of, caused by, works with etc. We also provide the ability for an analyst to write rules, using terms from the SWETO ontology, to capture their background knowledge. These rules, along with the inferences that can be made from the SWETO ontology using Web Ontology Language (OWL)/RDF-S axioms, allow us to infer new relationships in the process of information integration. In particular, it allows the analysts to write rules for dynamically creating a graph with nodes all of the same type (say persons) where all the links mean the same (e.g. associate-of, related-to). These are then suitable for the traditional structural analyses like finding central nodes and connectors etc. In this paper, we demonstrate how intelligence gathering tasks can be achieved using shallow language methods, linguistic pattern matching and semantic web technologies on top of unstructured text.

This paper is structured as follows: Section 2 gives an overview of some of the previous work related to analyzing unstructured text. Section 3 describes the methodology that we adopted for collecting the data, annotating the data. Section 4 presents our main findings and compares the outcomes with another approach. Finally Section 5 presents the conclusions and implications of our study analysis.

2. RELATED WORK

The process of extracting structured information from open and unstructured text has been widely studied in literature. The Knowledge Base Population Track [4] of the Text Analysis Conference is notable in this regard. Participants of this track are given an initial knowledge base (which is derived from Wikipedia infoboxes), and a corpus of documents to learn from. There are two related tasks: Entity linking, where the participants are required to align names to the entities in the KB, and Slot Filling, which requires finding information about entities from the given text.

Yates et. al. [6] describe their TextRunner system, which is a state of the art implementation of the Open Information Extraction idea. It makes a single pass over a huge quantity of data and extract relations from them in the form of triples from sentences. It tags sentences with part-of-speech tags and noun phrase chunks. For each pair of noun-phrases that are not too far and satisfy a certain criteria, a classifier is used to detect whether these should be a part of a triple. Further, parsed sentences are labeled as trustworthy or untrustworthy.

[2]http://www.w3.org/RDF/
[3]http://linkeddata.org/
[4]http://nlp.cs.qc.cuny.edu/kbp/2010/

In an article from the Communications of the ACM, Etzioni et. al. [3] talk about Open Information Extraction as a method that scales to large sizes, and capable of supporting unanticipated questions over arbitrary relations. The TextRunner system makes certain first order Markov assumptions about dependencies, and subsequently using a Conditional Random Field, learns to assign labels to each of the words in a sentence. In the next phase, it extracts the triples from these sentences.

Syed and Finin [5] describe unsupervised methods to discover ontology elements from Wikipedia article links. They state that an article infobox does not contain all entries possible for that particular article; and that more entries can be discovered from the text of the article. However, this paper limits itself to only the (inter-article) links found in the article text. Essentially, it treats the current article as the primary concept, and the linked concept as the secondary concept. For each linked concept, it tries to find an Is-a relationship with nodes in Wordnet. These wordnet sysnets can then serve as infobox labels, and the linked article title can serve as the value for the particular slot.

Kulkarni et. al [4] describe a method for annotating unstructured web text using entity IDs from an entity catalog such as Wikipedia. The main purpose of this work is indexing and search based on annotation, but does not make efforts to discover links between entities that are recognized. Particularly, with respect to terror intelligence, Basu [2] presents a method based on social network analysis to derive a linkage map of terrorist organizations in India. They use concepts such as betweenness, and centrality to mark out key organizations in-spite of having relatively low intensity of links. They report a collaboration or nexus between terrorist organizations is inferred from the data if both the organizations co-occur in a report of a single incident. Based on the frequency of these co-occurrences, an adjacency matrix is created. This approach is easily error prone. For instance, in our own analysis, we found that using such an approach can link famous people quoted in a story about a terrorist organization with it.

We use a better approach to find relationships between entities mentioned in text. Rather than inferring relationships between entities that co-occur in the same document, we limit these to co-occurrences within the same sentence. Further, we use linguistic pattern matching techniques to identify whether there is a known sequence of words between the two entity mentioned. Thus, if we find an expected pattern of words between two entity mentioned in the same sentence, then we report a specific relationship between them. Thus, we can identify not only links between terror entities, but also the kind of relationship between them.

3. METHODOLOGY

The basic idea is centered around creating RDF triple store based KB. This KB is populated by triples generated from text documents such as news reports of terrorist incidents. These triples are generated by first recognizing the named entities in the document and then using linguistic text matching rules to establish certain relationships between the recognized entities. These triples are created in accordance with the SWETO ontology. Additional triples are added to the KB by querying the Linked Open Data cloud for information on the recognized entities. Also, inferred triples are added to the store. These inferred triples

may be generated by OWL rules, or by using additional user defined rules. This knowledge base can then be queried using SPARQL to get information such as affiliations of an individual to certain terror organizations, or individuals involved in a certain terror incident etc. Such information can then be used by intelligence analysts.

3.1 Data Collection

For the purpose of our experiments, we gathered data from news stories available on the Internet. News stories are good for this purpose because of their well formed grammatical structure and reliability. We developed a crawler that collected documents from websites of leading Indian and Pakistani newspapers such as the Hindustan Times, Times of India, The Hindu, The Deccan Herald, Dawn, Jung and NDTV news.

The crawl was done using the IBM Content Analytics [5] technology. The document corpus was further augmented by documents gathered using Google's and Bing's search APIs. The keywords used for these searches were organization names from a government designated list of blacklisted organizations (marked terror organizations). The process of collecting information from these search engines was not regulated. This is because the emphasis of the experiment was to develop a general method for creating a structured datastore from unstructured text; thereby making reliability an issue of user discretion.

3.2 Annotation and Parsing

The documents collected in the previous step were first filtered on the basis of occurrence of blacklisted organizations within them. Documents with no blacklisted organizations were discarded. Next the unstructured text blocks were annotated by an NER (Named-Entity Recognition) system. For this purpose we used the OpenCalais system from Thomson Reuters.[6] We annotated the entities recognized by OpenCalais in the original document, which included names, locations (which would likely correspond to events) and organizations (See Figure 1 for the flow of conversion of unstructured to triple).

The next step after identifying entities within the text document, is to actually identify the relationships between them. We made a simplifying assumption at this juncture that related entities co-occurred in the same sentence atleast once in a document. This assumption was justified by the fact that if a block of text mentions a link between two entities, say A and B, then there would be at least one sentence in the text which would assert this relationship; and moreover would contain both the entities A and B. This assumption meant that we could divide the text into a stream of sentences, and analyze each sentence for possible links. For the actual purpose of link extraction we used simple regular expressions (see Table 1) based matching techniques using the PERL programming language.

The regex based matching is essentially a high precision and low recall technique i.e. there is a high probability that a recognized relationship actually exists in the document; and at the same time, the model was susceptible to errors of omission, i.e. it was prone to missing out on any pattern that wasn't explicitly stated as a regular expression.

[5]http://www-01.ibm.com/software/data/content-management/analytics/

[6]http://www.opencalais.com/

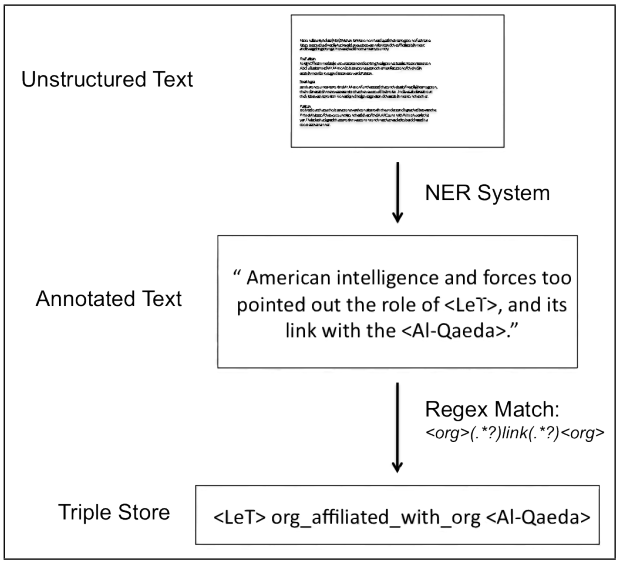

Figure 1: An example demonstrating the conversion of unstructured text to a triple.

The relationships thus extracted were stored as triples. Specifically triples are ways to express links in the form:

$$<Entity\ A>\ Relation{:}R\ <Entity\ B>$$

We enumerated relations which would be identified by the regex parser. These relations linked organizations to organizations and terror events, and people to organizations and terror events. In this way we obtained a collection of triples which we then stored in the KB (as seen in Figure 1). Since we wanted to avoid the name disambiguation issue, we did not extract person-person relationships directly from the text. Such relationships were deduced by the reasoning part of our system described later in Section 3.4.

The use of pattern matching techniques however produced a particular kind of false positives, wherein it identified links that related heads of states and anti-terror bodies with the blacklisted terror organizations. To overcome this, we employed the concept of a white list, which we populated with these commonly occurring names and organizations. The white list was made immune to any kind of assertion during the whole analysis.

3.3 Integrating Information from Linked Open Data

The Linked Data movement is a W3C-backed movement that seeks to connect semantic data across the web. It is a part of the larger semantic web project which is aimed at making the largely unstructured information on the Internet machine-comprehensible. The bulk of the linked open data on the web currently exists in the form of RDF datasets which can be queried through SPARQL endpoints exposed by them. Hence, one of the main motives for representing the relationships between the identified entities as RDF triples was the ability to represent them as instances of relations from an already existing ontology. For this purpose, we specifically chose our relationships as those given in the SWETO Ontology.

Once, the knowledge base was populated with these base facts, we were in a position to query the semantic web for

Predicate	Regular Expression
org_affiliated_with_org	links between(.*)(**org**)(.*)(**org**) (**org**)(.*)(links with|link with|connection|offshoot|wing of)(.*)(**org**) (nexus|connection between)(.*)(**org**)(.*)(**org**) (**org**)(.*)(**org**)(.*)nexus (**org**)(.*)working(.*?)for(.*?)(**org**)
associated_with	(**per**) (of|at the) (**org**) (**per**)(.*)works for(.*)(**org**)
member_of	(**per**)(.*?)(head of|founder[s]? of|leader of|chief|member of|president of)(.*?)(**org**) (**per**)(.*?)(**org**)(.*?)operative (**org**) (operative|activist)(.*?)(**per**)
responsible_for	(**org**)(responsib|behind|implicated|hand|involved|own up)(.*?)(**event**)
directly_involved_in	(**per**)(.*?)(implicated|arrested|alleged|accused|involved)(.*?)(**event**)
based_at	(**org**)(.*?)based(.*?)(**loc**)

Table 1: Regular Expressions and the Predicates/Relations they produce. Here the boldface 'org', 'per','loc' and 'event' represent organizations, persons, locations and events (on locations) respectively which are obtained (directly or indirectly) from the NER.

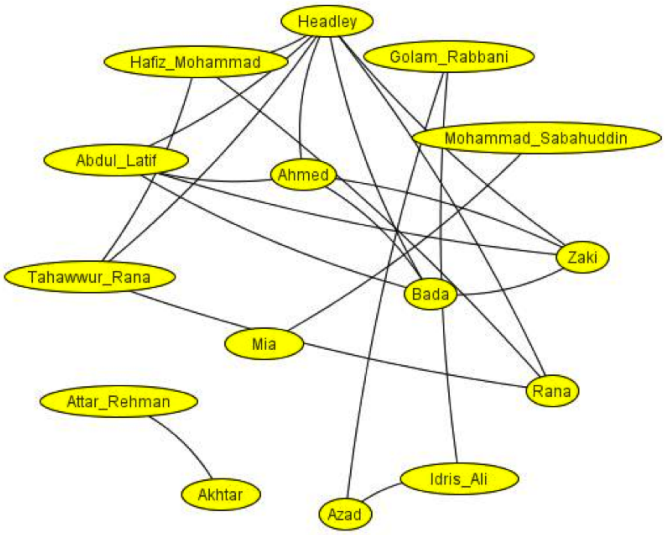

Figure 2: Graph of individuals related to each other on the basis of the query "All people involved in the same event are related to each other."

further *facts* on the identified entities. We issued SPARQL queries about the identified entities to the endpoint exposed by DBPedia.[7] The extra information was sought for groups, and not for individual names we found. While the groups we were interested in, did not need much disambiguation (AlQaeda or Lashkar-E-Tayabba are unlikely to be confused with other similarly named entities in our context), the same cannot be said for individuals. Since we are not explicitly using a name disambiguation step, the information we get back about individuals from DBpedia is likely to be ambiguous. Such triples returned by the SPARQL query results are added to the KB.

3.4 Human Asserted Rules and Inference

The skeleton knowledge base comprises information generated from web articles and news reports, and also information about identified entities from the Linked Open Data cloud. The third and possibly most important component that is integrated into the KB is a set of domain specific heuristics and background knowledge possessed by human analysts. These heuristics go beyond what is explicitly stated in the information gathered so far from web documents and DBPedia. For instance, the analyst might use a rule of the form *"All people involved in the same event are related to each other"* (See Figure 2). Such rules are expressed in the form of RDF statements in accordance with the SWETO ontology and are added to the KB. Now a forward chaining reasoner uses both OWL/RDFS rules and the analyst defined heuristics to generate inferred triples, which are integrated into the knowledge base.

4. RESULTS AND COMPARISON

First, let us give a brief note about the approach where links are assigned between any two entities that appear in the same document, without looking at any other contextual information (refer Section 3.2). The results of such an approach are shown in Figure 4.[8]

[7]http://dbpedia.org/sparql

[8]Thanks to Kanika Narang, Aastha Nigam, Ravi Dhingra,

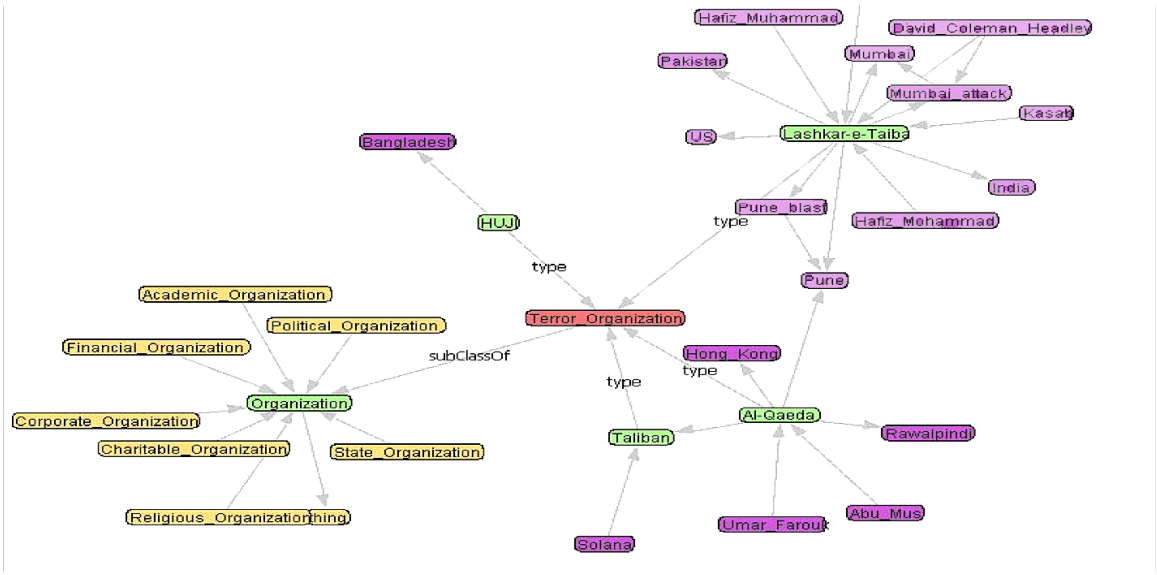

Figure 3: Partially populated SWETO ontology: The representative graph shows a sample KB which stores its relations in the form of triples.

Our approach does better than this, because we restrict the assignment of links/relations between entities based on context. Our input documents were obtained from two sources. Some of them comprised news stories from leading Indian and Pakistani dailies collected using a standard web crawler. The documents obtained at this stage were non- entity-specific. This way of gathering data simulates the situation where human analysts look for small amounts of relevant information in a huge quantity of very general news data. The total number of documents crawled were about 50,000. As expected, the average number of triples extracted per document was very low: about 1 triple/195 documents. More documents were collected from keyword based searches on popular search engines such as Google and Bing. These queries contained the names of black-listed organizations and the top 100 or so results per query were sent through the processing pipeline to generate triples. This was the simulation of targeted entity-specific searches and yielded a much higher triple-document ratio of approx. 1 triple/6.4 documents. The total number of triples obtained before verifying against white lists was 1, 277.

Before extracting triples from the collected documents, we filter out those documents which do not mention any of the blacklisted organizations in our list. We chose to do this because unless a document talks about one of the entities in question, it is unlikely to yield any of the relationships we are looking for. This also explains why we observed a significant improvement in the triple/document ratio when we shifted to an entity specific search of documents where the chances of occurrence of a blacklisted organization name within a particular document were greatly increased.

We next queried the SPARQL endpoint at dbpedia.org using the identified organization names (which were a subset of blacklisted organizations). An extensive query, that handled redirects was able to yield relevant triples. In par-

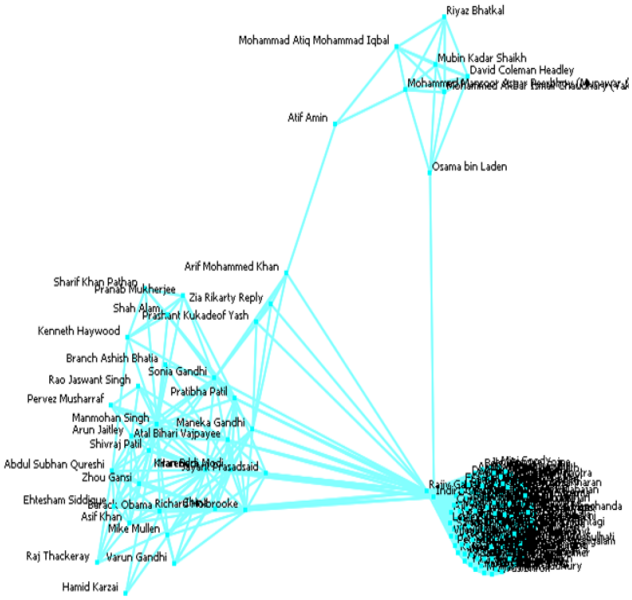

Figure 4: When linkages are assigned between all entities that appear in a document, without further constraints (such as white listing), it leads to a lot of false positives. Heads of state and prominent politicians are often mentioned in articles that describe terror incidents.

Indraprastha Institute of Information Technology, Delhi for helping us with this.

ticular we were able to extract on an average 9 triples for prominent organizations such as the Taliban. This yields a representative graph of the form shown in Figure 3. Notice that this graph contains different kinds and the graph edges have semantics associated with them. In addition to the triples from DBPedia, now we add analyst defined rules such as: If *Org:A Relation:Caused Event:X* and *Org:B Relation:ClaimedResponsibilityFor Event:X* then assert to the KB a new triple *Org:A Relation:RelatedTo Org:B*. This rule essentially means that if Org:A is said to have caused a particular terror event, and if Org:B claims responsibility for it, then these two organizations must be related to each other (either they are the same, or one is a parent organization of the other). Consider another rule: If *Person:A Relation:InvolvedIn Event:X* and *Person:B Relation:InvolvedIn Event:X* then assert Person:A Relation:RelatedTo Person:B. This is simply a user added rule that people who are involved in the same terror incident are related to each other. After accounting for this rule, a graph of related individuals can be extracted as a result of a SPARQL query. This is as shown in Figure 2.

5. CONCLUSIONS AND FUTURE WORK

We presented a framework for integrating data from both unstructured data sources (news reports and web articles), and structured data sources (e.g. Linking Open Data, LOD) with domain specific heuristics and background knowledge possessed by analysts to generate actionable intelligence information. We used shallow methods such as named entity recognition (NER) and linguistic pattern matching to identify entities of interest and discover relations between them. We stored this information in a triple-store based KB. This KB was then augmented with information gathered from DBpedia, and user defined rules expressed in the SWETO ontology. This allowed us to generate a semantically meaningful graph which can then be queried using SPARQL as desired to answer targeted questions. Ongoing work in this research is focused toward dealing with ambiguity in individual names. This will improve the quality of our results further. Also, our initial experiments were performed on a dataset of limited size. In the future, we will be performing more experiments that deal with datasets of large sizes.

6. REFERENCES

[1] B. Aleman-Meza, C. Halaschek, A. Sheth, I. Arpinar, and G. Sannapareddy. SWETO: Large-Scale Semantic Web Test-bed. In *Ontology in Action Workshop, co-located with the 16th International Workshop on Software Engineering and Knowledge Engineering*, pages 490–493, 2004.

[2] A. Basu. Social Network Analysis of Terrorist Organizations in India. In *2006 Conference of the North American Association for Computational Social and Organizational Science*. North American Association for Computational Social and Organizational Science, June 2005.

[3] O. Etzioni, M. Banko, S. Soderland, and D. Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.

[4] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 457–466, New York, NY, USA, 2009. ACM.

[5] Z. Syed and T. Finin. Unsupervised techniques for discovering ontology elements from Wikipedia article links. In *First International Workshop on Formalisms and Methodology for Learning by Reading (FAM-LbR)*, page 78, 2010.

[6] A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland. TextRunner: Open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations on XX*, pages 25–26. Association for Computational Linguistics, 2007.