International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 08, August-2015

Automatic Extraction of Publicly Available Personal Information from World Wide Web

Prodip Kumar Sarker[1],
[1]Department of Computer Science & Engineering,
Begum Rokeya University,
Rangpur, Bangladesh

Dr. Bimal Kumar Pramanik[2]
[2]Department of Computer Science & Engineering,
University of Rajshahi,
Rajshahi, Bangladesh

Mousumi Saha[3]
[3]Department of Telecommunication and Electronic Engineering,
Hajee Mohammad Danesh Science & Technology University,
Dinajpur, Bangladesh

***Abstract* - Personal information on the World Wide Web is huge in number and widely distributed. Moreover, the information is increasing rapidly with the growth of social network and heterogeneous web. The heterogeneous personal data is mostly available in unstructured or semi- structured format. Automatic extraction of personal heterogeneous data is obviously a challenging task, and getting researchers attention at a rapid pace. In this regard, my research proposal introduce a method of automatic extraction of personal information converting unstructured and semi-structured data into rich semantic Resource Description Framework (RDF), which integrates all personal and their associated information or descriptions by comprising of their relations. Furthermore, introducing an approach of indexing the huge personal information, which is often considered as BIG DATA, to increase the accessibility to the personal information quickly and efficiently.**

Keyword- Information Extraction, RDF, Map reduce, Web of Data, Semantic Web Crawl.

I. INTRODUCTION

Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). The present significance of IE pertains to the growing amount of information available in unstructured form at World Wide Web. Tim Berners-Lee, inventor of the World Wide Web, refers to the existing Internet as the web of documents [4] and advocates that more of the contents be made available as a web of data. Until this transpires, the web largely consists of unstructured documents lacking semantic metadata. Knowledge contained within these documents can be made more accessible for machine processing by means of transformation into relational form, or by marking-up with XML tags. An intelligent agent monitoring continuously growing World Wide Web contents requires IE to transform unstructured data into something that can be reasoned with. A typical application of IE is to scan a set of documents written in a natural language and populate a database with the information extracted [9].

Entity retrieval, that is, returning objects (such as people, organizations, locations, products, etc.) in response to users information needs, has received considerable attention recently from various research communities, as well as from the commercial sector that leads Web of Documents to Web of Data. According to a recent study, more than half of web queries target a particular entity or instances of a given entity type [8]. Supporting the search and discovery of entities, therefore, is essential for ensuring a satisfying user experience. Compared to the Web of Documents, the Web of Data is much more structured. However, since each Web of Data source might have its own defined schema, ranging from loosely to strictly defined, the data structure does not follow strict rules as in a database. Even within a given a data source, the schema might not be fixed and may change as the information grows. The information structure evolves over time, and new records can require new attributes. We therefore consider the Web of Data as being semi-structured [1]. The most important property of the Web of Data is that it is naturally organized around entities, and that each of these entities is uniquely identified by a Uniform Resource Identifier (URI).

II. LITERATURE REVIEW

The field of Information Extraction (IE) and Information Retrieval (IR) is characterized by rigorous attention to evaluation and measurement. International bench-marking campaigns, such as the Text Retrieval Conference (TREC) and the Initiative for the Evaluation of XML Retrieval (INEX) play a key role in fostering IR research by providing a common platform, evaluation methodology, and relevance judgments to assess the quality of information access systems. The introduction of the expert finding task at the TREC Enterprise track in 2005 was a significant milestone on the path to entity-oriented retrieval. The goal of the expert finding task is to create a ranking of people who are experts in a given topical area [5]. Later, in 2007, INEX launched an

Entity Ranking track [6]; here, entities are represented by their Wikipedia page and two tasks are considered: (i) entity ranking, where a query and target categories are given, and (ii) list completion, where a textual query, example entities, and, optionally, target categories are provided as input. In 2009, the Entity track at TREC started with the goal to perform entity-oriented search tasks on the Web, and defined the related entity finding (REF) task [3]. REF requests a ranked list of entities (of a specified type) that engage in a given relationship with a given source entity. The collection used there is a general Web crawl and entities are identified by their homepages. The 2010 edition of the track introduced an entity list completion task [2], similar to that of INEX, but the collection is a Semantic Web crawl, specifically, the Billion Triple Challenge 2009 (BTC-2009) dataset1. Looking at these developments over time, a shift of emphasis can be observed from the document- oriented web to the data-oriented web, or Web of Data.

III. PROPOSED METHOD

In the essence of Web of Data, this paper retrieve person-oriented entity from the World Wide Web documents and recently developed social network like Facebook, twitter, researchGate and so on. A pre-processing step is required in order to extract entities from documents.

To extract information at first develop a crawler that crawls information containing information about persons. A Web crawler is an Internet bot which systematically browses the World Wide Web, typically for the purpose of Web indexing [10]. Web search engines and some other sites use Web crawling or spidering software to update their web content or indexes of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine which indexes the downloaded pages so the users can search much more efficiently.

A Web crawler starts with a list of URLs to visit. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the *crawl frontier*. URLs from the frontier are recursively visited according to a set of policies. If the crawler is performing archiving of websites it copies and saves the information as it goes. The archives are usually stored in such a way they can be viewed, read and navigated as they were on the live web, but are preserved as 'snapshots'.[11]

There are different crawling policy are used. The behavior of a Web crawler is the outcome of a combination of policies: [12]

- a *selection policy* which states the pages to download,
- a *re-visit policy* which states when to check for changes to the pages,
- a *politeness policy* that states how to avoid overloading Web sites, and
- A *parallelization policy* that states how to coordinate distributed web crawlers.

The architecture of a Web crawler is in the following

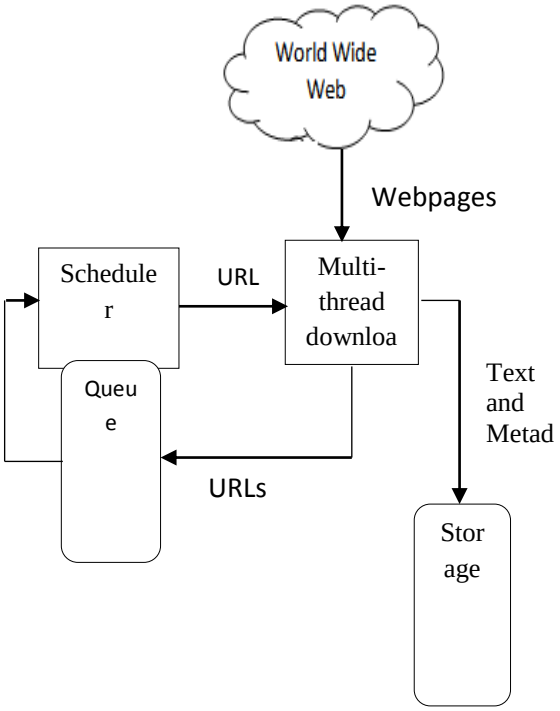

Fig: Architecture of a Web Crawler

Secondly, an analyzer extracts named entity, coreference and relations to define a triple of < subject, predicate, object > format. This relation is so called a "triple" because it has three parts, three parts are described in terms of the grammatical parts of a sentence: subject, predicate, and object. The following figure displays the elements of the tri-part model and the symbology associated with the elements when graphing them [13].

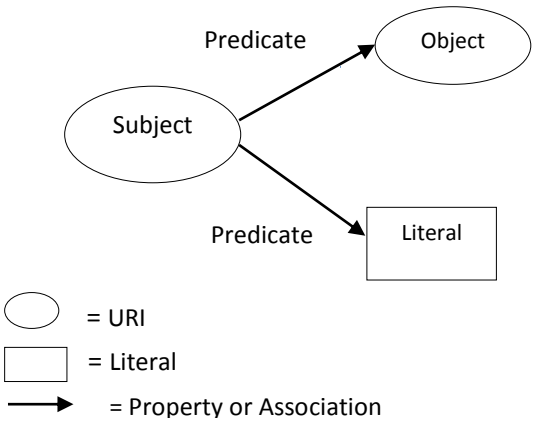

Fig: Relations of triple

The key elements of a relation are as follows:

Subject. In grammar, this is the noun or noun phrase that is the doer of the action. In the sentence "The Company sells batteries," the subject is "the company." The subject of the sentence tells us what the sentence is about. In logic, this is the term about which something is asserted. In RDF, this is the resource that is being described by the ensuing predicate and object.

Predicate. In grammar, this is the part of a sentence that modifies the subject and includes the verb phrase. Returning to our sentence "The Company sells batteries," the predicate is the phrase "sells batteries." In other words, the predicate tells us something about the subject. In logic, a predicate is a function from individuals (a particular type of subject) to truth-values with an arity based on the number of arguments it has. In RDF, a predicate is a relation between the subject and the object.

Object. In grammar this is a noun that is acted upon by the verb. Returning to our sentence "The Company sells batteries," the object is the noun "batteries." In logic, an object is acted upon by the predicate. In RDF, an object is either a resource referred to by the predicate or a literal value.

Statement. In RDF, the combination of the preceding three elements, subject, predicate, and object, as a single unit.

The entity extraction algorithm is implemented using the Map Reduce [7] programming model.

Named Entity Extraction

1. Named entity recognition: recognition of known entity names (for people and organizations), place names, temporal expressions, and certain types of numerical expressions, employing existing knowledge of the domain or information extracted from other sentences. Typically the recognition task involves assigning a unique identifier to the extracted entity. A simpler task is named entity detection, which aims to detect entities without having any existing knowledge about the entity instances. For example, in processing the sentence "P.K. Sarker likes fishing", named entity detection would denote detecting that the phrase "P.K. Sarker" does refer to a person, but without necessarily having (or using) any knowledge about a certain M. Smith who is (or, "might be") the specific person whom that sentence is talking about.
2. Coreference resolution: detection of coreference and anaphoric links between text entities. In IE tasks, this is typically restricted to finding links between previously-extracted named entities. For example,"International Business Machines" and"IBM" refer to the same real- world entity. If we take the two sentences"P.K. Sarker likes fishing. But he doesn't like biking", it would be beneficial to detect that"he" is referring to the previously detected person"P.K. Sarker".
3. Relationship extraction: identification of relations between entities, such as: PERSON works for ORGANIZATION (extracted from the sentence"P.K. Sarker works for BSMRSTU.") PERSON located in LOCATION (extracted from the sentence"P.K. Sarker is in Bangladesh.")

This entity extraction phenomena helps developing a huge network of data, especially a huge network of persons all over the world. An intelligent agent monitoring continuously growing World Wide Web contents acquisite unstructured data to triple that can be reasoned with. Eventually, SPARQL will do the job of reasoning.

Map-Reduce

At the step of identification of named entity, coreference and relationship at a particular document, I plan to apply map-reduce algorithm to develop a network of RDF graph structure. In the map function, the input key < c > is the URL of a document, and the value < s,p,o > is a triple in the document. The map phase simply outputs the same triple for two different join keys, one triple for the entity appearing as a subject and one triple for the entity ap- pearing as an object. The outputs are then partitioned, sorted, and merged by the MapReduce framework. In the reduce phase, all the triples for each join key < c,e > are aggregated together to build a graph containing the incoming and outgoing relations of an entity e. The resulting entity graph is similar to the sub-graphs. Furthermore, we filter all entity graphs that are composed of only one or two triples. In general, such entity graphs do not bear any valuable information, and can be removed in order to reduce the noise as well as the size of the dataset

IV. CONCLUSION

The problem is that natural language processing is a large and challenging topic The Semantic Web structure gives users the power to share and collaboratively generate decentralized linked data. In many cases, though, collaboration requires some form of authentication and authorization to ensure the security and integrity of the data being generated. Personal data are stored in unstructured or semi-structured way for extracting these type of information at first convert these unstructured or semi-structured format data to meaningful structured format data. In this regard, my research proposal introduce an easy method for automatic extraction of personal information converting unstructured and semi-structured data into rich semantic Resource Description Framework (RDF), which integrates all personal and their associated information or descriptions by comprising of their relations.

V. REFERENCES

[1] Serge Abiteboul. Querying semi-structured data. Springer, 1997.
[2] K Balog, P Serdyukov, and AP de Vries. Overview of the trec 2010 entity track. 2011.
[3] Krisztian Balog, Arjen P de Vries, Pavel Serdyukov, P. Thomas, and T. Westerveld. Overview of the trec 2009 entity track. In Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009). NIST, 2010.
[4] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. International journal on semantic web and information systems, 5(3):1–22, 2009.
[5] Nick Craswell, Arjen P de Vries, and Ian Soboroff. Overview of the trec 2005 enterprise track. In Trec, volume 5, pages 199–205, 2005.
[6] Arjen P De Vries, Anne-Marie Vercoustre, James A Thom, Nick Craswell, and Mounia Lalmas. Overview of the inex 2007 entity ranking track. In Focused Access to XML Documents, pages 245–251. Springer, 2008.
[7] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. Communications of the ACM, 51(1):107–113, 2008.
[8] Jeffrey Pound, Peter Mika, and Hugo Zaragoza. Ad-hoc object retrieval in the web of data. In Proceedings of the 19th international conference on World Wide Web, pages 771–780. ACM, 2010.
[9] Rohini K Srihari, Wei Li, Cheng Niu, and Thomas Cornell. Infoxtract: A customizable intermediate level in- formation

extraction engine. In Proceedings of the HLT-NAACL 2003 workshop on Software engineering and architecture of language technology systems-Volume 8, pages 51–58. Association for Computational Linguistics, 2003.

[10] https://en.wikipedia.org/wiki/Web_crawler.

[11] Masanès, Julien (February 15, 2007). Web Archiving

[12] Castillo, Carlos (2004). Effective Web Crawling.

[13] Michael C. Daconta Leo J. Obrst Kevin T. Smith "The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management".