Exploring Freebase Potentials for Big Data Challenges

Mahmoud Elbattah
College of Engineering & Informatics,
National University of Ireland
Galway, Ireland
m.elbattah1 {at} nuigalway.ie

Mohamed Roshdy
Faculty of Computer and Information Sciences
Ain shams University,
Cairo, Egypt

Mostafa Aref
Faculty of Computer and Information Sciences
Ain shams University,
Cairo, Egypt

Abdel-Badeh M Salem
Faculty of Computer and Information Sciences
Ain shams University,
Cairo, Egypt

Abstract-- **As Big Data investments are persistently growing worldwide by businesses and governments as well, enterprises currently recognize their Big Data as a priceless source for business process improvement and sustainability. Consequently, Big Data practitioners might face difficulties in accessing real-world big datasets. However, open-data portals, such as Freebase, present new insights for researchers to facilitate accessibility to massive datasets. The paper provides an exploratory study of Freebase in an attempt to demystify its actual potentials for building applications able to process Big Data. For that, Freebase is deconstructed by explaining its technical features, data model and querying capabilities. Furthermore, the differences between the two approaches of Freebase and Wikipedia are highlighted in a comparative perspective. . (Abstract)**

Keywords– *Freebase, Big Data, Open Data*

I. INTRODUCTION

Earlier in 2014, UK government introduced £73 million of new funding to help the public and academics unlock the capabilities of Big Data [1]. However, the massive volume of Big Data is just a single challenge according to a Gartner's report [2]. More complex challenges lie in the variety of data sources from rigidly structured data such as business transactions to loosely unstructured data such as social networks. Moreover, "Velocity" of data where analyzing data-in-motion as fast flows of data streaming into data repositories.

On the other hand, Big Data practitioners could probably find difficulty in accessing big datasets which are a part of a big organization due to data privacy and protection issues. Collaboratively created databases, such as Freebase, DBpedia and Linked Data, provide unprecedented opportunities for diverse researchers to conduct studies on real big datasets. Though, the "Know-How" is a must to harness the capabilities of such immense data sources. In this paper, a road-map exploratory study is presented in order to provide the must-know knowledge and technical characteristics of Freebase.

II. DEFINING FREEBASE

Described by Tim O'Reilly upon Freebase launch, "Freebase is the bridge between the bottom up vision of Web 2.0 collective intelligence and the more structured world of the semantic web" [3] . Freebase is an open, writable, semantic database with information on millions of topics ranging from genes to jeans [4]. Freebase stores data from international and government agencies, private foundations, university research groups and individual users.

Freebase contains dozens of millions of topics, thousands of types and tens of thousands of properties. Each topic is linked to other related topics and annotated with important properties like movie genres and people's dates of birth. Over two billion facts or relations that make Freebase one of the greatest ever sources of knowledge.

A brief historic overview of Freebase [5], it was developed by the American software company Metaweb and has been running publicly since March 2007. Metaweb was acquired by Google in July 16, 2010. Google's Knowledge Graph is powered in part by Freebase.

A. Main Features of Freebase

- **An Identity Database [3]:** Freebase ensures that each topic is a single reconciled identity, that there should be only one GUID representing each real world entity, topic, or concept. For example, Arnold Schwarzenegger appears in Freebase as an actor, a politician, a governor and a champion. In Freebase, however, there is only one topic for Arnold Schwarzenegger that brings all those facets together.
- **Graph-Shaped Data Store [6]:** Freebase data structure is based on networked graphs where nodes representing entities are connected by edges. By storing the data as a graph, Freebase can quickly traverse arbitrary connections between topics and easily add new schema without having to change structure of the data.

- **A Large Data Object Store (LOB) [7]:** Comprising a store of large data objects such as text documents, images and sound files. LOB objects are indexed and annotated in the graph store.
- **Integrated Versioning Mechanism:** Freebase has built-in reversion support of all database edits, thus allowing "undo" of large, complex operations to any degree.
- **User-Friendly Web UI:** Casual and non-technical users can use Freebase's Web UI to search, browse, create, and edit the data stored in Freebase.

B. Freebase Data Model

Data models of Freebase are called "Schemas" which are broken down into a set of a few components as follows:

- **Topic:** Currently, Freebase has over 40 million topics about real-world entities like people, places and things. A topic object may be specific and concrete (e.g. Albert Einstein, London) or an abstract concept (e.g. Euler's number {e}, Globalization). Freebase adopts "Entity Reconciliation": a topic may be associated with many names or abbreviations, but each topic should represent one and only one entity or concept in the world. And each topic is given exactly one globally unique identifier (GUID). Figure (1) demonstrates an example of entity reconciliation in Freebase.
- **Type:** A type is an object that is used as a conceptual container of properties that are most commonly needed for describing a particular aspect of information. A topic associated with a type is considered to be an instance of that type. Examples of types include "Film Actor", "Book Author", "Location", and "Programming Language". Topics in Freebase can have any number of assigned types which may be added or removed over time. Unlike object-oriented models, or some of the RDF models, Freebase types do not have inheritance [3]. Figure (2) demonstrates the usage of types in Freebase with "Winston Churchill" as an entity example.

Figure 1. Example of how Freebase handles the ambiguity and multiplicity of representing entities: How people can refer to "University of California Los Angeles" with many different styles. Freebase maps all those different representations into a single entity with unique ID.

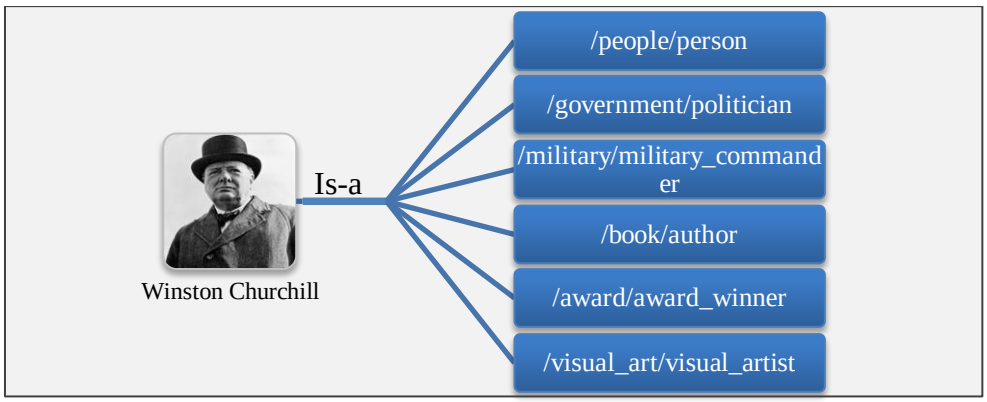

Figure 2. Example of how Freebase uses the concept of "Types" to address the multi-faceted nature of topics: "Winston Churchill" is of type "Politician", which resides in the "Government" domain (government/politician), and he is also typed as a person, which resides in the "People" domain (/people/person), and he can be found in other domains like "Military Commander", "Author", "Nobel Laureate" or "Visual Artist".

- **Property [8]:** Properties of a topic define a "Has-a" relationship between the topic and the value of the property (e.g. Paris {topic} has a population {property} of 2,153,600 {value}). It's very common that the property name is a verb, or verbal phrase (e.g. "directed by").
- **Domains**: Types are grouped into domains which are similar to ordinary sections in a newspaper such as Business, Arts and Entertainment, Politics, Economics, etc. "Commons" are special types of domains which have met certain standards to be considered as well-known topics. The commons are listed at "Category: Commons".

C. Knowledge Graphs

Data structure in Freebase is defined in the form of a "Knowledge Graph" as a set of nodes and a set of links that establish relationships between the entities. Subsequently, Freebase data is non-hierarchical and can model more complex relationships rather than conventional databases. Moreover, relationships can be simply extended or narrowed by adding or removing links to other nodes. Figure (3) illustrates an example of the usage of knowledge graphs in Freebase

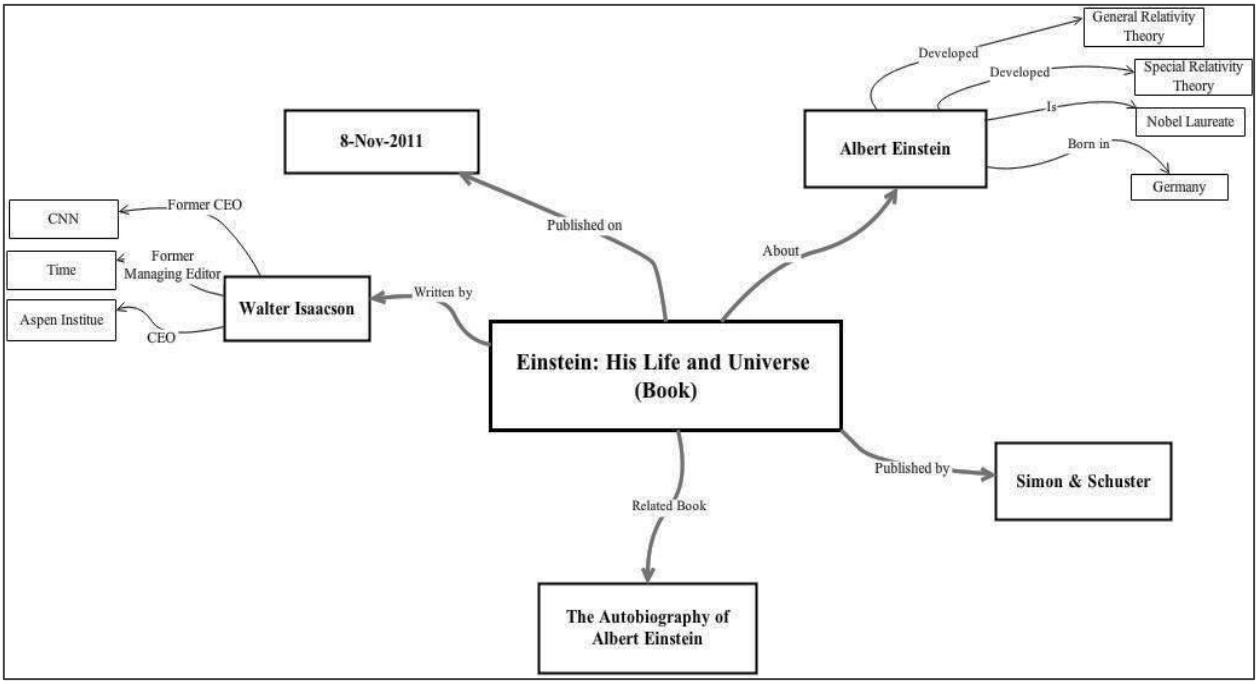

Figure 3. Example of a knowledge graph representing a Freebase topic, a biography book, "Einstein: His Life and Universe". The building block of Freebase knowledge graph is the "Triplet" that connects two entities (topics) such as "Albert Einstein" and "Germany" with a "Born-in" relationship (type).

III. QUERYING FREEBASE

Freebase offers a powerful query language, MQL (Metaweb Query Language), for performing programmatic complex queries. This allows incorporating knowledge from the Freebase database into external applications or websites [9]. MQL syntax is JSON-based and can be submitted via HTTP with responses returned also in JSON.

MQL is an easy-to-use, object-oriented query language with a tree-based result structure of objects. It includes dynamic schema support without the need for a DDL, path-based node naming and idempotent transaction-less write support. Important MQL features include mixing structural data matching with approximate string matching of literals, cursors, and at semantics of all data, which makes mixing of data and metadata easy. [10]

An example of MQL query, retrieving all diseases that have the "Anorexia", "Nausea" and "Dysphagia" symptoms in common:

```
[{
  "name": null,
  "type": "/medicine/disease",
  "sym1:symptoms": [{ "name": "Anorexia"  }],
  "sym2:symptoms": [{ "name": "Nausea" }],
  "sym3:symptoms": [{ "name": "Dysphagia"  }]
}]
```

IV. DATA SOURCES OF FREEBASE [11]

Wikipedia is one of the major data sources for Freebase and provides the core set of topics. However, other sources are used including the following:

- Wikimedia Commons
- EDGAR
- Open Library Project
- Stanford University Library
- TVRage
- ISFDB
- MusicBrainz
- National Register of Historic Places
- OurAirports
- NFDC FAA
- ITIS - Taxonomy of plants and animals
- World of Spectrum
- WordNet

V. FREEBASE USERS AND USAGE POLICY

Freebase is licensed under the "Creative Commons" which greatly facilitates the usage and redistribution of Freebase data. Freebase mainly supports the following classes of users:

- **Researchers:** Wide varieties of researches can benefit from Freebase in areas such as data mining, knowledge discovery, semantic web, ontology creation and analysis, and graph analysis.
- **Data Contributors:** The data holders who upload their datasets into Freebase so that they can be open-access from a structured and graph-based database.
- **Application Builders:** Developers who are interested in building public data services that access the data in Freebase are supported through the Freebase API.

VI. DISCUSSION: FREEBASE VS. WIKIPEDIA

Discussing the differences between Freebase and Wikipedia is a point of controversy as they may seem similar for laymen. Nevertheless, Freebase has outstanding characteristics apart from Wikipedia as follows:

Firstly, Freebase can be considered as well-structured database for providing rich types and pre-defined schemas for the entities while Wikipedia depends primarily on categories. In Freebase, the entity can rarely be assigned meaningless types as the types of an entity determine the appropriate attributes schema. Accordingly, Freebase enjoys a better type taxonomy and more complex schemas that make it more convenient to be used by both human and machine as well. Figure (4) illustrates the different approaches of structuring data in Freebase and Wikipedia.

Secondly, Freebase contains much more topics than Wikipedia. According to the current statistics on each site, Freebase is about 10 times larger than English Wikipedia, (43 million Freebase topics vs. 4.5 million Wikipedia articles) [13].

Finally, the powerful query language of Freebase is a competitive advantage for Freebase. MQL can support complex and nested queries in an automated fashion that harnesses the large datasets of Freebase.

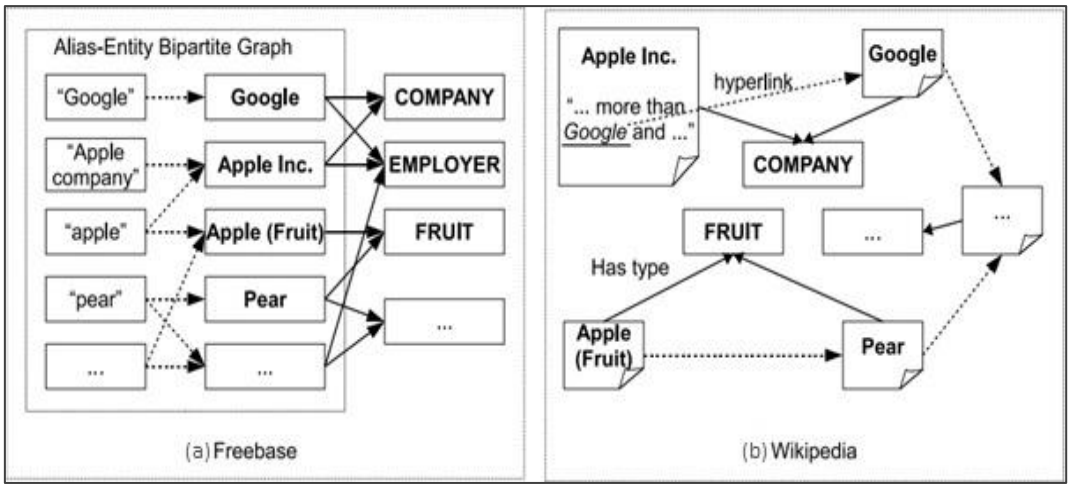

Figure 4. [12] An Example of Data Structure in Freebase compared to Wikipedia: Data model of (a) Freebase depends on well-structured schemas whereas that of (b) Wikipedia depends on rich textual content and hyperlinks.

VI. CONCLUSION

Freebase can be considered as the transition from the soft-semantic knowledge structure of Wikipedia into a hard-semantic knowledge structure. Freebase explicitly encodes entity-entity network via knowledge graphs in order to build a huge well-structured data source to the world's knowledge.

Freebase holds promising potentials for Big Data practitioners concerning different perspectives: First, it provides open-access huge datasets spanning diverse domains. Secondly, Freebase can be considered to handle one of the key challenges of Big Data, "Data Variety", since Freebase models data with highly structured schemas. Over and above, the Freebase API and the strong query language (MQL) can help build complex applications that are capable of processing massive datasets for purposes of knowledge discovery or data mining.

Nevertheless, comparing Freebase to other open-data sources such as DBpedia, Linked Data or Wikiepdia is still controversial. More applications need to be built around Freebase to investigate the efficiency and completeness of Freebase approach.

REFERENCES

[1] "https://www.gov.uk/government/news/73-million-to-improve-access-to-data-and-drive-innovation", retrieved on 12 March 2014.

[2] Mark Beyer, Anne Lapkin, Nicholas Gall, Donald Feinberg, Valentin T. Sribar ,"'Big Data' is Only the Beginning of Extreme Information Management", Gartner Report, April 2011.

[3] Tim O'Reilly ,"Freebase Will Prove Addictive", O'Reilly Radar, retrieved on 12-03-2014.

[4] Toby Segaran, Colin Evans, and Jamie Taylor, "Programming the Semantic Web", O'Reilly Media, P.116, 117,119, 2009.

[5] "http://en.wikipedia.org/wiki/Freebase", retrieved on 12-03-2014.

[6] "http://developers.google.com/freebase/guide/basic_concepts#graph". Retrieved on 13-03-2014.

[7] Kurt Bollacker, Robert Cook, Patrick Tufts, "Freebase: A Shared Database of Structured General Human Knowledge", AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence - Volume 2, 2007.

[8] "http://wiki.freebase.com/wiki/Property", retrieved on 17-03-2014.

[9] David Flanagan,"MQL Reference Guide", Metaweb Technologies, Inc., P. 2, 27, 2009.

[10] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, Jamie Taylor "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge", Proceedings of the ACM SIGMOD international conference on Management of data, 2008.

[11] "http://wiki.freebase.com/wiki/Data_sources", retrieved on 21-03-2014.

[12] Zhicheng Zheng, Xiance Si, Fangtao Li, Edward Y. Chang, Xiaoyan Zhu, "Entity Disambiguation with Freebase", Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology- Vol. 1, 2012.

[13] "http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia", retrieved on 21-03-2014.