Generating Unified Famous Objects (UFOs) from the Classified Object Tables

Anusha Kola, Harshal More, Sean Soderman, Michael Gubanov
Department of Computer Science
University of Texas at San Antonio

***Abstract*—It is difficult to access data generated by different data sources due to the representation and format differences. ETL, KETL, Jedox, Apatar are some examples of data translation and fusion packages that can be used to resolve representation differences of data coming from different data sources have been favored. However, most tools require significant manual effort to map the data sources. Structural mismatch of data between objects with the same semantics reduces the accessibility of data.**

Here we discuss our initial efforts toward a scalable unsupervised system and algorithms to generate Unified Famous Objects (UFO) - the self-learning "intelligent" data structures that help automate data fusion at scale [Gubanov et al., 2009], [Gubanov et al., 2011]. UFO is a data structure encapsulating different representations of the same data object (e.g. Songs), hence capable of automatically recognizing and mapping such object in different data sources, and significantly reducing manual effort during data integration process.

We evaluate our algorithms on a large-scale Web tables corpus having $\approx$ 64 million of tuples.

***Keywords*-Web-search; Large-scale Data Management; Cloud Computing; Data Fusion and Cleaning; Summarization; Human-Computer Interaction.**

I. Introduction

Big data variety is one of the most challenging problems in Big data research agenda [Stonebraker, 2012], [Gubanov, 2017], [M.Gubanov et al., 2017], [M.Podkorytov et al., 2017]. A variety of data sources, produce valuable data, but also inevitably introduce information representation differences that represent a significant impediment for someone who wants to gain access to *all* relevant data sources. Here we describe our initial efforts to automate data fusion on a large-scale corpus of Web tables containing $\approx$ 64 million tuples.

Solution: To resolve the information representation mismatch between the data sources, there are many solutions including [Bernstein, 2003], [Gubanov et al., 2009], [Gubanov et al., 2011], [Gubanov et al., 2014], [Gubanov and Stonebraker, 2014], [Gubanov and Pyayt, 2013], [Gubanov and Shapiro, 2012], [Gubanov et al., 2008], [Gubanov and Bernstein, 2006]. Among these solutions, the IBM UFO Repository is one one of the attempts to scale data fusion up. It introduces an initial notion of UFO - an object that creates an abstraction over different data representations that have the same semantics. For example, "preis" is German for "price", and both should be treated the same way when considering data stored either of the attributes (except the currency is different). Here, we specifically describe our initial efforts to generate UFOs automatically from tables of classified objects in a large-scale structured dataset.

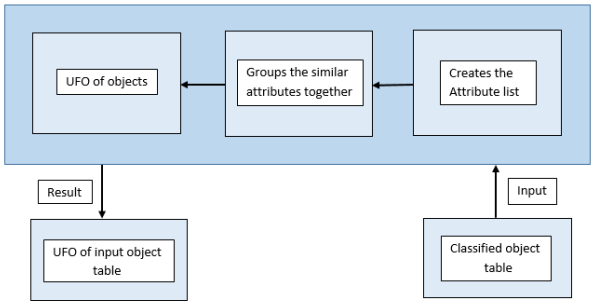

Fig 1 Architecture Diagram for generating UFO from classified tables.

Figure 1. Architecture diagram for generating UFOs from classified tables

II. Architecture

Algorithm: We generate UFOs in three steps. First, we calculate the weights of table attributes within a chosen classified subset of our dataset. Table I shows the attributes sorted by weight for the *Songs* object. *Name*, *price*, *album*, *time*, and *artist* attributes have the most weight.
Next, we group similar attributes together by walking down the set of discovered attributes, generating related words such as synonyms, hyponyms, lemmas and plurals for each attribute. We group attribute a with any attribute b (where $a \neq b$ and is the attribute having a set of words generated) if a word belonging to a's set of words matches b.

Listing 1. Query to Retrieve the Classified Object Metadata from the

```
SELECT file_name,col1,col2,..colN
WHERE isMetadata = true and file_name
IN (SELECT distinct file_name
FROM corpus
WHERE object_classificationID=true)
```

Lastly, in order to recognize the data without a metadata field, or in recognizing the more related attributes that did not have proper metadata fields, we applied our method to

Attributes	Weight
name	32781
price	31123
time	31020
album	21374
artist	11838
länge	1658
preis	1658
nom	729
prix	729
durée	728
interpret	591
	478
格	478
タイトル	478
nome	459
prezzo	415
durata	415
lengte	375
naam	375
prijs	375
アルバム	320
artiste	285
precio	250
duración	248
álbum	224
título	216
artista	208
アティスト	158
description	87
released	81
artiest	71
itunes	47
duração	44
preço	44
nombre	34

Table I
SONGS UFO ATTRIBUTE WEIGHTS

train a classifier identifying *price* attribute. This was done by feeding in the data from different *price* attributes as identified by our generated UFO.

III. EVALUATION

Evaluation: With respect to the above experiment of identifying price attributes, we observe precision of 91.6% and recall of 89.7% on 10-fold cross-validation. In Listing 2, we can see an example of *Songs* UFO, in Listing 3, we can see an example of *Posts* UFO.

Listing 2. UFO Songs in JSON

```
{"Songs": {
      "name": ["name", "nom", "nome",
            "タイトル", "naam", "tÃtulo", "title",
            "lyrics", "nombre", "song"],
      "price": ["price", "preis", "prix",
            "prezzo", "prijs", "precio", "preço",
            "perhour"],
      "artist": ["artist", "artista", "artiest",
            "artiste", "interpret"]
      "time": ["time", "length", "länge",
            "lengte", "durée", "durata",
            "duración", "duração"],
      "album": ["album", "Ã¡lbum", "movie"],
      "description": ["description",
            "descripción"],
      "music": ["music", "type"],
      "date": ["date", "datum"],
      "show": ["show"],
      "type": ["type", "all styles"],
      "download": ["download", "search"]
}}
```

Listing 3. UFO Posts in JSON

```
{
 "Posts": {
  "date": ["date","date &amp; time","date:","date
       added","date/time","dates","-- date --","
      issue / date","date m/d/y","date d'
      inscription"," date ","date posted:","date
      posted","!date","dates","date of findings
      ","date and time","! dates","thedate","
      title/date","{|date}","time"],
  "discussion": ["discussion","discussions","
      general discussion","tech discussion","wrap
      -up discussion","basic discussions","brli
      discussions","band discussion","discussion
      topic","fashion discussion","forum
      discussion","paperdiscussion","discussion
      boards","trading discussion","forum","forum
       topic","forums","antrim forum","forum name
      ","post del forum","recruiting forum","
      forum index","  forum","brave new forum
      ","youth forum","forum","subforums","forum
      description","forum thread","spamfree
      forums"],
  "post": ["latest post","last post","posts","
      post","recent post","last posts","blog post
      ","latest post info","last poster","poster
      ","lastpost"],
  "replies/comment": ["replies","total replies
      :","no replies yet","{replies|views}","no
      replies","replies:","comment","comments","
      last comment","no comment","special comment
      ","regulatory comment","commenti","
      commented:","# comments","comment:","!
      comments","comments:","comments(i)","issuer
       comment","site comments","comment"],
  "topic": ["topics","similar topics","latest
      topics","main topics","general topics","
      agenda topics","photography topics","forum
      topics","active topics","thesis topics","
      topics/messages","title","topic title","!
      title","item title","titles","title","title
      }"],
  "update": ["last updated","updated","date
      created","last update","last updated on","
      update","lastupdated","modified date","
      updates","updated &lt;&lt;","date modified
      "],
  "views": ["views","view","total views","{
      replies|views}","views:","{view}","# views
      "]
 }
}
```

To evaluate object recognition performance of the generated UFOs, we use these UFOs to classify the Web tables:

- Extract the metadata rows from the Web tables corpus.
- Match the metadata field names using the UFO.
- Return the best matching tables.

Stored UFOs can be used in identifying the objects from any data source. This process is illustrated in Figure 2.

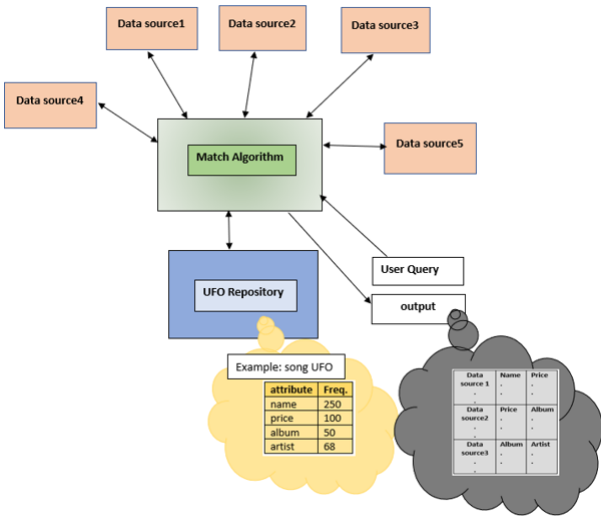

Figure 2. Identify Songs instances using the Songs UFO

IV. Related Work

[Gubanov et al., 2009], [Gubanov and Pyayt, 2012], [Gubanov and Shapiro, 2012] Proposed and evaluated the concept of a Unified Famous Object (UFO) - a unified, standardized interface to heterogeneous data sources. Our constructed UFOs can be used to recognize and query data objects from different data sources.

[Gubanov et al., 2011], [Gubanov and Pyayt, 2013], [Gubanov et al., 2014], [Gubanov and Stonebraker, 2014], [Gubanov et al., 2013] Used UFOs for representing entities in an index of topics extracted from a reference book. Our framework also concentrates on creating the sets of attributes with straight matches then identifying and merging similar attributes into a single UFO attributes.

[Bernstein, 2003] describes Model Management (MM), a framework to manipulate schemas a mappings with many useful common operators. However, MM's operators work with schemas as a whole, hindering its scalability compared to UFOs.

V. Conclusion

Our generated UFOs can be used for:

- Identifying similar data objects in a large-scale structured dataset.
- Fusing/Mapping/Transforming the data instances from different data sources having different schemas.

Our future focus is on training a classifier or algorithm to generate formulas/transformations for the data instances under a particular column for each attribute of an identified UFO. This can be extended further:

1) To form more accurate clusters based on UFOs.
2) To identify similar objects that have similar metadata.
3) To amend objects having none or incorrect metadata with automatically generated metadata.

Finally, the tables can have none or incorrect metadata. Machine learning models can be trained to identify correct metadata if it is present.

References

[Bernstein, 2003] Bernstein, P. A. (2003). Applying model management to classical meta data problems. In *CIDR*.

[Gubanov, 2017] Gubanov, M. (2017). Polyfuse: A large-scale hybrid data fusion system. In *ICDE*.

[Gubanov et al., 2009] Gubanov, M., Popa, L., Ho, H., Pirahesh, H., Chang, J.-Y., and Chen, S.-C. (2009). Ibm ufo repository: Object-oriented data integration. In *VLDB*.

[Gubanov and Pyayt, 2012] Gubanov, M. and Pyayt, A. (2012). Medreadfast: Structural information retrieval engine for big clinical text. In *IRI*.

[Gubanov and Pyayt, 2013] Gubanov, M. and Pyayt, A. (2013). Readfast: High-relevance search-engine for big text. In *ACM CIKM*.

[Gubanov et al., 2011] Gubanov, M., Pyayt, A., and Shapiro, L. (2011). Readfast: Browsing large documents through unified famous objects (ufo). In *IRI*.

[Gubanov and Shapiro, 2012] Gubanov, M. and Shapiro, L. (2012). Using unified famous objects (ufo) to automate alzheimer's disease diagnostics. In *BIBM*.

[Gubanov et al., 2013] Gubanov, M., Shapiro, L., and Pyayt, A. (2013). Readfast: Structural information retrieval from biomedical big text by natural language processing. In *Information Reuse and Integration in Academia and Industry*. Springer.

[Gubanov and Stonebraker, 2014] Gubanov, M. and Stonebraker, M. (2014). Large-scale semantic profile extraction. In *EDBT*.

[Gubanov et al., 2014] Gubanov, M., Stonebraker, M., and Bruckner, D. (2014). Text and structured data fusion in data tamer at scale. In *ICDE*.

[Gubanov and Bernstein, 2006] Gubanov, M. N. and Bernstein, P. A. (2006). Structural text search and comparison using automatically extracted schema. In *WebDB*.

[Gubanov et al., 2008] Gubanov, M. N., Bernstein, P. A., and Moshchuk, A. (2008). Model management engine for data integration with reverse-engineering support. In *ICDE*.

[M.Gubanov et al., 2017] M.Gubanov, M.Priya, and M.Podkorytov (2017). Cognitivedb: An intelligent navigator for large-scale dark structured data. *WWW*.

[M.Podkorytov et al., 2017] M.Podkorytov, D.Soderman, and M.Gubanov (2017). Hybrid.poly: An interactive large-scale in-memory analytical polystore. *ICDM DSBDA*.

[Stonebraker, 2012] Stonebraker, M. (2012). Big data means at least three different things... In *NIST Big Data Workshop*.