Fundamental Analysis Powered by Semantic Web

Xian Li
lix15@rpi.edu
Tetherless World Constellation, Rensselaer Polytechnic Institute
Troy, NY 12180, USA

Jie Bao
baojie@cs.rpi.edu

James A. Hendler
hendler@cs.rpi.edu

Abstract—Conducting fundamental analysis within subsets of comparable firms has been demonstrated to provide more reliable inferences and increase the prediction quality in equity research. However, incorporating and representing both firm-specific information and common economic determinants has been widely recognized as the key challenge. This paper investigates how to leverage Semantic Web technologies to assist fundamental analysis by generating flexible and meaningful selections of comparable firms at low costs. We approach the problem by proposing Linked Open Financial Data as the data organization model and ontology modeling for knowledge representation. Results are verified in terms of efficiency with examples of quick mashups, and feasibility by adapting to existing valuation models.

Keywords: fundamental analysis; financial data; financial statement analysis; investment; Semantic Web; linked data

I. INTRODUCTION

Fundamental analysis focuses on valuations of firms and identifying temporarily mispriced securities, which is of obvious interest to shareholders and creditors. Traditional methods rely upon cross-sectional estimations of earnings and returns using information conveyed from financial statements and equity market values [3,18,22]. Considerable observations on the variations have revealed the weakness of cross-sectional method of sacrificing firm-specific characteristics, and have highlighted the importance of analyzing comparable firms within a subset [10]. For example, traditional valuation models have shown to be most effective for high Book-to-Market (BM) firms [24], while they fail to provide meaningful measurements for rapidly growing firms. And the same fundamental signal may have different implications on different groups of firms. For technology firms, “Research & Development Expense” imposes negative impact on current earnings but implies positive growth for future earnings, thus applying the “net earning” across different firms to predict growth rate may decrease the quality of results.

Contextual fundamental analysis based on a set of conditional determinants, e.g. firm size, product type, competitiveness and capital intensity, has demonstrated to be able to provide more reliable results [11,20]. Professional analysts generally perform analysis within a subset of firms which have common primary characteristics. For instance, some analysts focus on “growth” stocks, some specialize in “value” stocks. Analysts’ reports at industry or sector level also show that their analysis is mostly done under a specific context.

The prominent challenge is the choice of conditional variables that further generate groups of comparable firms [4,8]. The selection of an appropriate subset of firms to perform fundamental analysis has been suggested as “an art form” and largely decided subjectively [8]. To allow for simultaneous control over a range of explanatory variables as well as generating appropriate selections of comparable firms by capturing factors that drive cross-sectional variations, it requires a systematic and objective approach. Currently, the major difficulties include:

• Flexible data organization at low cost

Although financial statements are the principal resources used in fundamental analysis, predicting future cash flows and estimating excess stock returns involve a variety of factors that go well beyond accounting data, e.g. market data, exchange data, industry outlook, product demands, and customer groups. How to bring together relevant data from different sources and ensure interoperability? How to facilitate hypotheses by allowing the growth of available data as well as navigating among data without barriers? How to achieve all of these at low cost?

• Meaningful selection to enhance comparability

The sophistication of defining comparability in fundamental analysis comes from intensive domain knowledge and various rationales behind valuation models. Hence, an objective approach needs to not only understand common grounds of concepts but also fit to the specific logic relations behind valuation theories.

To address these challenges, we investigated Linked Open Financial Data (LOFD) as a cost-effective data organization model combined with ontology and rules to represent the domain knowledge in the process of choosing comparable firms. Semantic Web languages, e.g. RDF (the Resource Description Framework) and OWL (the Web Ontology Language), are inherently built with a graph-based open data model and thus naturally support integration from different data sources and applications. In addition, these languages are based on formal knowledge representations so that they enable the automatic processing and inference about data.

The contribution of this paper includes:

Firstly, we designed an incremental data organization model with low cost for fundamental analysis. It enables broader selections of contexts under which to choose comparable firms. It can quickly leverage the value of datasets...
curated at other places while not losing relevancy. In total it produces a comprehensive data environment containing potentially explanatory factors.

Second, we explored Semantic Web knowledge representation to express domain concepts and logic relationships. Flexible aggregations and disaggregations according to semantic meanings are achieved to enhance comparability in terms of profitability, growth and risk.

The rest of the paper is organized as follows. Section 2 describes Linked Open Financial Data as the data organization model for contextual fundamental analysis. In Section 3, we narrow down to financial statements and investigate how to build knowledge representation models to achieve appropriate selections of comparable firms in fundamental analysis. We verify the efficiency and feasibility of proposed solutions in Section 4. Related work is presented in Section 5, and we conclude in Section 6.

II. LINKED OPEN FINANCIAL DATA

There are enormous influential factors in analyzing firms and equity returns. Both firm-specific characteristics and common economic drivers play important roles in understanding causal relationships, co-occurrences and other phenomena. Therefore, having a cohesive picture of facts about the whole business ecosystem which allows controlling over explanatory variables as well as narrowing down for details at low cost is the basic challenge. Meanwhile, the Web is becoming an open environment for data sharing and governments are pioneering publishing agency data for public access. Hence, whether this emerging trend could be an opportunity in addressing this challenge is our first concentration. We start with analyzing the major types of data used in fundamental analysis, current difficulties in integrating them and observed opportunities.

A. Major Data

1) Accounting data

As the principal sources of capital markets research, accounting data communicate the basic fact about a firm’s value through revenues, cash flows, operating costs, etc. In recent years XBRL (eXtensible Business Reporting Language) has been widely adopted for corporate filings. XBRL provides structured data through taxonomies along with the reporting facts and their contexts (e.g. reporting date, currency unit), while not aiming at expressing clear semantics. Firms are identified by the CIK (Central Index Key) in SEC data depository1.

2) Market data, exchange data

Trade-related messages follow the FIX (Financial Information eXchange) Protocols using “Tag=Value” syntax to describe order details. For example, an order about “Preferred Class A of IBM” could be denoted as “55=IBM 65=PRA”. Historical market data are also accessible from the Web in table formats2.

3) Economic data

As governments’ efforts in making data available for public consumption, valuable datasets for fundamental analysis are significantly enlarged. Key economy factors (e.g. labor force, employment, earnings, GDP, CPI) are provided by government agencies, but with broad range of conventions about format, naming, terms etc. The U.S. Data.gov portal3 is an example with both traditional formats and machine-readable formats provided by Semantic Web technologies.

4) Firm profiles

Firm’s website usually contains firm-specific data including products, industry, geography distribution. However, such information is embedded in HTML and lacks both structures and meanings for maching processing. DBpedia4 complements the limitation with basic profiles for public firms in Semantic Web format.

5) News, crowd psychology

Stock prices fluctuate as news comes. Now investors use the Web as a source for latest news and are actively monitoring and participating online social mediums. For instance, discussions on stock message board have been shown to be closely related to contemporaneous market movements [2]. Latest fashion is on Twitter5 where stocks are discussed identified by “dollar tags” followed by stock symbols, such as $GOOG. Applications6 which filter tweets related to capital market investment are gaining popularity among investor communities.

The variety of standards and naming conventions for data contents as well as diversified data formats complicates data integration tasks. Furthermore, structure-oriented nature of these data formats (e.g. XBRL, CSV) makes it difficult to process, browse and query data from a large range of datasets.

Therefore, we adopted a linked data [7] strategy for data integration and built an incremental data organization model called “Linked Open Financial Data”. Data are transformed into machine-readable forms that enable efficient connections of relevant data from independent sources. The value of data could be further realized by generating meaningful and flexible panel data and conditional hypothesis for cross-sectional analysis [16].

LOFD is achieved through two stages. At first, existing financial data are converted into RDF format as the bricks for linked data. Second, RDFized data are interconnected and linked to other linked data based on semantic relevancy.

---

1 http://www.sec.gov/edgar/searchedgar/companysearch.html
2 http://www.finance.yahoo.com
3 http://data.gov
4 http://dbpedia.org/About
5 http://twitter.com
6 http://stocktwits.com/
B. Making Financial Data Linkable

1) Converting Existing Data

We focused our conversion efforts on two major formats of existing financial data: well-structured XBRL files from SEC filings, and simple-structured tabular formats which are widely used for time series data. Considering the amount and complexity of financial data, we make the conversion stage minimal but extensible, just enough to preserve the contents of the existing data.

- Tabular

Data stored in tables with rows and columns of cells can be transformed into Turtle/RDF using TWC LOGD converter [19]. During the conversion, metadata and provenance information critical for discovery are captured and it also transforms into Turtle/RDF using TWC LOGD converter [19].

We focused our conversion efforts on two major formats of existing financial data: well-structured XBRL files from SEC filings, and simple-structured tabular formats which are widely used for time series data. Considering the amount and complexity of financial data, we make the conversion stage minimal but extensible, just enough to preserve the contents of the existing data.

- XBRL

We first converted XBRL-encoded SEC 10-Q reports to RDF format using xbrlimport. Figure 1 is an example showing how a reporting fact is described with details such as accounting concept, value, reporting firm, time and currency:

```
us100:BAC
 dc:identifier "BAC";
 rdf:label "BAC".
us100histdaily-bac:thing_6127
 mdate "2010-10-15"^^xsd:date;
 mclose "11.98"^^xsd:decimal;
 mdv:volume "5.96439E8"^^xsd:decimal;
 mdt:trading_symbol us100:BAC .
```

![Figure 1 SEC 10-Q report after conversion](image)

Furthermore, following the extensible principle, we promoted the CIK identifiers to URIs to describe firms and associated it with the literal value that appeared in XBRL files so as to make them as linkable resources:

```
<http://xbrlimport.sourceforge.net/>
```

1) Semantic Stock Tweets

We collected stock tweets from Twitter by mapping stock trading symbols to a keyword-based search using Twitter’s Streaming API. An instance of TwitLogic [25] was used to collect stock tweets and translate them into the interoperable RDF data model containing details such as author, timestamp, the stock(s) being tweeted, content, retweet relations and links of tiny URLs. Using semantic stock tweets, we captured which firms or stocks were discussed in related to what topics. Here is an example:

```
?s dc:created "2010-11-05T16:10:34.000-04:00"^^dateTime;
 sioc:content "RT @CNBC: BofA Fighting Suits Over $375 Billion http://cnbc.com/id/40031714 $BAC #Mortgages #Foreclosures #SEC #Lawsuits #Banking";
 sioc:links_to <http://cnbc.com/id/40031714>;
```

C. Linking Open Financial Data

We identified several strategies for building up links among RDFized datasets and linking to external resources.

1) Automatic Linking

- Linking within financial data

We found stock trading symbol is the most commonly used identifiers, so the mappings between trading symbols and CIKs are established from predicates dei:TradingSymbol and xbrli:identifier contained in 10-Q data. For firms which do not have trading symbol in their reports, the mappings could be generated from the RDF responses by using Calais API, from which we also get firms’ industry SIC codes. Then we connected to the RDFized CSV datasets by stock trading symbols via the owl:sameAs predicate:

```
<http://logd.tw.rpi.edu/source/sec-edgar/cik/000070858>
dc:identifier "000070858";
 skos:prefLabel "BAC";
 lofd:siccode "6162";
 owl:sameAs us100:BAC .
```

- Linking to economic data

Data.gov has comprehensive datasets about industries, products and consumers. For example, below is a fragment from the RDFized quarterly data about unit labor costs per industry:

```
dgp331:thing1 dgp331:sector_code "3100";
dgp331:sector_name "Manufacturing, Durable Goods";
dgp331:thing2 dgp331:measure_text "Unit Labor Costs";
dgp331:measure_code "11";
dgp331:thing3 dgp331:measure_code "11";
dgp331:sector_code "3100";
dgp331:thing4 dgp331:id "PPUS32006033";
dgp331:value "68.907";
dgp331:period "Q1";
dgp331:year "2009".
```

---

1 Data downloaded from Yahoo! Finance are not for publish purpose.
2 Namespaces are listed in Appendix.
3 http://xbrlimport.sourceforge.net/
We can have these data linked to firms based on the SIC codes at sector level.

c) Linking with stock tweets data

To capture investor psychology as a factor for mispricing [15], we linked to stock tweets data via trading symbols:

dollar:tag:BAC owl:sameAs value_of_trading_symbol:BAC.

d) Linking to other linked data

We adopted a semi-automated approach for mapping to DBpedia, which contains basic profile information of firms e.g. dbpedia-owl:product. First, we generated preliminary mappings using trading symbol, e.g. from “AAPL” to dbpedia:AAPL. Then the remaining ones were heuristically mapped by firm names, e.g from “GOLDMAN SACHS GROUP INC” to dbpedia:Goldman_Sachs_Group.

We further linked to the New York Times Linked Data with DBpedia URIs as bridges, and thus could browse related news through values of the nyt:search_api_query predicate.

2) Leverage Social Intelligence for Linking

With respect to the decentralized and social nature of the Web, we explored using Semantic MediaWiki to facilitate the linking process. We extended collectively by adding contents via friendly user interface as shown in Figure 2.

D. Summary

As a result, we built a linked data model which could incrementally integrate different sources of data to provide a number of contexts for fundamental analysis. An instance of LOFD model is shown in Figure 3.

III. OTOLGY MODELING FOR COMPARABILITY

Despite the broad acceptance of XBRL as the structure model for financial statements, XBRL does not aim to represent clear semantics related to business logics as well as domain knowledge, which are required for fundamental analysis. Our previous work has provided a faithful mapping from XBRL formats to Semantic Web representations [4]. Based on the translation results where reporting concepts are mapped to named classes and relationships (presentationArc, calculationArc and definitionArc) are represented by object properties, in this study we explored how to model the domain semantics with the goal of capturing different meanings of comparability and generating homogeneous groups of firms. We focused on three primary financial statements: Statement of Financial Position, Statement of Comprehensive Income and Statement of Cash Flows, using OWL 2 DL for ontology modeling and SPARQL 1.1 (with SPIN[12] syntax extension) for query.

A. Conceptual Model

The primary information conveyed by financial statements contains the firm’s economic resources and their related changes due to corporate activities. Hence the scope of our modeling includes reporting concepts involved in calculation relationships, and we found frequently used variables in fundamental analysis were covered, such as asset, liability, equity, net income, expense and cash flow. To reflect the conceptual hierarchies reflected through the additive relationships we transformed property xbrl:calculationArc to fa:componentOf property and ensured decidability:

EquivalentClasses(fa:Assets ObjectHasSelf{ ex:pA } )
EquivalentClasses(fa:AssetsCurrentObjectHasSelf{ ex:pC } )
SubObjectPropertyOf( ObjectPropertyChain(ex:pA owl:topObjectProperty ex:pC)
fa:componentOf )

For reporting concepts not connected by calculation relationships, we identified fundamental variables in presentation relationships, e.g.

us-gaap:CommonStockSharesOutstanding, which illustrates the drawback of organizing concepts based on presentation: the type of financial instrument (“CommonStock”) is intermingled with the reporting item (“NumberOfShares”). We mapped such composite concept to independent classes, and added properties to specify the relationships between them:

Declaration (Class (fa:Equity))
Declaration (Class (fa:CommonStock))
SubClassOf(fa:CommonStock fa:Equity)
SubClassOf(fa:Equity ObjectAllValueFrom{fa:hasNumberOfShares xsd:decimal})
EquivalentClasses(us-gaap:CommonStockSharesOutstanding ObjectAllValueFrom(ObjectInverseOf{fa:hasNumberOfShares fa:CommonStock})

B. Aligning Corporate Activities

It is the common practice to analyze firm’s performance independent of its capital structure. Various valuation models

11 e.g. http://data-gov.tw.rpi.edu/wiki/Dow_Jones_Industrial_Average

12 http://www.spinrdf.org/
have identified specific signals to measure profitability, growth and risk \[3,22\] in terms of financing activities (how firms obtain capital resources) and business activities (how they use resources to create value) \[12\]. We modeled such shared knowledge about corporate activities:

\[
\text{DisjointWith(fa:BusinessActivity fa:FinancingActivity)}
\]

\[
\text{DisjointUnion(fa:BusinessActivity fa:OperatingActivity)}
\]

Among business activities, firm has its “core” activities referred as operating activities while other activities are not of the firm’s central business. However, there is no universal classification for all firms. For example, Coca-Cola creates value by converting raw materials into goods for sale, but it may also have some investing activities such as “a portfolio of bonds for trading purposes” which is not of Coca-Cola’s operating activities. In contrast, Goldman Sachs creates value by providing financial services and “a portfolio of trading securities” is of its central business. As operating asset is a commonly used variable in fundamental analysis, whether it is calculated using the same reporting items across all firms or the nature of individual firms is taken into account has different influences on the quality of fundamental analysis results.

To capture such variation, we used property \(fa:hasActivityType\) to connect reporting concepts with activity classes so that operating asset could be generated by asking for reporting items which are of asset category and associated with operating activities, such as:

\[
\text{SELECT ?s WHERE \{ ?s a fa:Asset; fa:hasActivityType fa:OperatingActivity.}\}
\]

Essentially, the most accurate mappings require a management approach \[12\] since the firm’s management is at the best position in judging which assets are part of its ongoing business. Under the current financial reporting environment, such ontology mapping requires extensive domain knowledge in differentiating operating assets (operating liability) from financing assets (financial liability). Here we developed several heuristic rules to automate the process before further manual mapping.

First we designed mapping rules based on the nature of the reporting item relative to the nature of the business as illustrated in the above example. The rules are coded as:

\[
\text{SubClassOf(fa:VendorItem ObjectAllValueFrom(fa:hasActivityType fa:OperatingActivity))}
\]

Third, we developed mapping rules based on the source of incomes and expenses.

\[
\text{SubClassOf(ObjectAllValuesFrom(fa:hasIncomeSource fa:Goods) ObjectAllValueFrom(fa:hasActivityType fa:OperatingActivity))}
\]

Activity alignment based on underlying semantics leads to a common measuring methodology as well as preserves firm-specific characteristics, and thus a more accurate way to calculate key financial ratios. For instance, debt-to-equity ratio is frequently used to assess leverage risk, while it is the “financing liabilities” that should be used rather than “total debt” \[12\]:

\[
\# \text{debt-to-equity ratio} = \text{Financing Liabilities} / \text{Common Stock Equity}
\]

\[
\text{CONSTRUCT \{ ?this ex:flratio ?dfl.?
\text{WHERE \{ ?e a fa:CommonStock; xbrlo:hasContext ?c; xbrlo:hasValue ?equityvalue. }
\text{?d a fa:Liability; xbrlo:hasContext ?c; xbrlo:hasValue ?debtvalue. LET \{ ?dfl := (?debtvalue/\?equityvalue).\}}
\}
\]

C. Expressing Implicit Meanings

1) By Function and By Nature

Although classifying expenses “by function” \[17\] helps to capture the overall business trends (e.g. wholesale revenue), it blurs the impacts driven by different economic characteristics. For example, labor and raw materials might respond to economic events in different ways. Evidence from existing research has proved that aggregating accounting numbers “by nature” facilitates more comparative analysis \[6,8,12\].

To express both perspectives in describing expenses and thus enable flexible comparability, we modeled these two categorizations as two separate sets of class members and asserted the equivalency between the two categories. The disjoint constraints were added to make sure non-overlap within the same category.

\[
\text{DisjointUnion(fa:ExpenseByFunction fa:CostOfGoodsExpense fa:SellingExpense fa:AdministrativeExpense})
\]

\[
\text{DisjointUnion(fa:ExpenseByNature fa:RawMaterialExpense fa:StaffingExpense fa:DepreciationExpense})
\]

\[
\text{EquivalentClasses(fa:ExpenseByFunction fa:ExpenseByNature})
\]

Although the benefits of this approach are not fully realized unless all firms file their incomes with the “by nature” dimension, which FASB (Financial Accounting Standards
Board) is promoting [12], firms that have been preparing their income statements in such way could be analyzed comparably.

2) Level of Uncertainty

Reporting items about incomes carry different prediction signals on future cash flows and stock returns. Reasons could be due to the variations among income components in persisting in future as well as the subjectivity involved in estimating for uncertain factors while filing the reports. For instance, the cash components of income are shown to have higher persistence and thus are better predictions for future cash flows than accrual items are. Reporting recurring remeasurements involves intensive subjectivity to estimate both the timing and amounts of the uncertain items.

Following principles proposed by FASB [12], we expressed the degrees of uncertainty explicitly by using generic classes to express the types of uncertainty as shown in Figure 4. Hence, for a given reporting item, implicit meanings about its uncertainty levels (denoted in yellow) as well as domain-specific knowledge are captured.

![Figure 4 Express the implicit differences among incomes](image)

3) Timing and Risk

Liquidity and financing flexibility are of importance to understand the firm’s risks. However, existing reporting items about assets and liabilities are only classified as “current” and “noncurrent”, which are based on the firm’s operating cycle. Such distinction decreases the comparability because operating cycles vary a lot across firms. It is difficult to assess the timing of future cash flows without a clear benchmark. For instance, reporting item us-gaap:ReceivablesLongTermContractsOrPrograms is defined as “amount to be collected within one year of the balance sheet date (or one operating cycle, if longer)” which carries two meanings for “long term”. To improve comparability, we adopted OWL time ontology to express the timing explicitly. Here is an example where we specified that the expected realization date of an asset is one year after the reporting date. This approach could be applied to other reporting items wherever consistency is absent.

```
_:f :a us-gaap:ReceivablesLongTermContractsOrPrograms;
  xbrlo:hasContext
  [a xbrlo:numericContext;
   xbrlo:period [a time:Instant;
    time:inXSDDateTime "2010-11-05T00:00:00+8:00"^^xsd:dateTime.]];
  time:hasDurationDescription [a time:DurationDescription;
   time:year 1.].
```

D. Aggregating and Totaling

With the ontologies proposed so far, comparability in financial analysis is enhanced by enabling meaningful and flexible classifications of firms to get homogenous groups. Aggregation is required in calculating key financial ratios. The process of aggregating and totaling is accomplished through SPARQL query. Since the aggregated value of monetary reporting items is of central interest for fundamental analysis, we developed mechanisms for querying the net totals according to underlying logic. We reused the calculation rule which compares the balance types [27]: whether a reporting item has positive or negative impact on the totaling of its super class depends on whether their balance types are the same or not. Thus the net total is aggregated by SPARQL query as below:

```
SELECT ?p (?credit-?debit) AS ?netvalue
WHERE {
  select (sum(?vp)) as ?credit
  WHERE {
    ?c1 fa:componentOf ?p;
    xbrlo:balance ?bc1;
    xbrlo:hasValue ?vp.
  }
  FILTER (?bp!=?bc1).
  GROUP BY ?p
}
SELECT (sum(?vn)) as ?debit
WHERE {
  ?c2 fa:componentOf ?p;
  xbrlo:balance ?bc2;
  xbrlo:hasValue ?vn.
  FILTER (?bp!=?bc2).
  GROUP BY ?p
}
```

IV. RESULTS

We verified the results in two aspects. First we demonstrated the efficiency of LOFD in integrating variables from multiple data sources. Then combined with the results from ontology modeling, we showed fundamental variables could be aggregated based on meanings so that the valuation factors from valuation models [1,13,26] are accurately generated.

A. Quick Mashup

Mashups on LOFD lead to quick hypotheses about context under which to generate comparable firms. As an example, a mashup from stock tweets data, SEC 10-Q reports, NYSE US 100 index data and short interest data reveals that whether stocks have short interests or not may have different effects on to what extent social dialogs on Twitter are related to the firm’s earning performances as shown in Figure 5.
Furthermore, we can easily use data curated at other places. For instance, the following query on corporate management published by rdf.about.com would explore whether board sizes provide context for comparable performances:

```
SELECT ?firm ?epsd ?boardsize {
  {SELECT ?firm count(?mgmt) as ?boardsize
   WHERE {
     ?o a sec:OfficerRelation;
     sec:corporation ?firm.
   }
  SERVICE <http://plato.cs.rpi.edu:8890/sparql>
  {SELECT ?firm ?epsd WHERE {
    GRAPH <http://tw.rpi.edu/lofd/resources> {
      ?r rs:name ?firm;
      dcterms:identifier ?cik.
    }
    GRAPH <http://tw.rpi.edu/lofd/marketdata> {
      ?s owl:sameAs ?r;
      md:closing ?price;
      md:date ?day.
    }
    GRAPH <http://tw.rpi.edu/lofd/2010/sec-edgar> {
      ?f0 a fa:NetIncome;
      fa:hasActivityType fa:OperatingActivity;
      xbrlo:value ?earning;
      xbrlo:hasContext ?c1.
      ?f1 a us-gaap:EarningsPerShareDiluted;
      xbrlo:value ?epsd;
      xbrlo:hasContext [xbrlo:entity ?cik].
    }
  }
  LET (?pe := ?price/?epsd).
  LET (?fsize := ?price*?shr).
  LET (?bm := ?bv/(?price*?shr)).
  FILTER (?earning >0).
}
```

2) Earning multiple in related with profit margin, firm size, book-to-market ratio

Fundamental analysis is inseparable from market-based valuation. We mashed up accounting data and market data with flexible selection of subset firms. We compared the earning multiples, profit margin, firm size and book-to-market among firms which have positive operating profits [13].

```
CONSTRUCT { ?this ex:firmname ?firm;
  ex:priceearningratio ?pe;
  ex:profitmargin ?pm;
  ex:firmsize ?fsize;
  ex:bmratio ?bm.
}
WHERE {
  GRAPH <http://tw.rpi.edu/lofd/resources> {
    ?r rs:name ?firm;
    dcterms:identifier ?cik.
  }
  GRAPH <http://tw.rpi.edu/lofd/marketdata> {
    ?s owl:sameAs ?r;
    md:closing ?price;
    md:date ?day.
  }
  GRAPH <http://tw.rpi.edu/lofd/2010/sec-edgar> {
    ?f0 a fa:NetIncome;
    fa:hasActivityType fa:OperatingActivity;
    xbrlo:value ?earning;
    xbrlo:hasContext ?c1.
    ?f1 a us-gaap:EarningsPerShareDiluted;
    xbrlo:value ?epsd;
    xbrlo:hasContext [xbrlo:entity ?cik].
    ?f2 a us-gaap:CommonStockValueOutstanding;
    xbrlo:value ?bv;
    xbrlo:hasContext ?c2.
    ?f3 a us-gaap:CommonStockSharesOutstanding;
    xbrlo:value ?shr;
    xbrlo:hasContext ?c2.
  }
  LET (?pe := ?price/?epsd).
  LET (?fsize := ?price*?shr).
  LET (?bm := ?bv/(?price*?shr)).
  FILTER (?earning >0).
}
```

V. RELATED WORK

The importance of choosing comparable sets of firms has been well studied in areas of accounting and finance. Contextual analysis was proposed since it was found that the cross-sectional variations are related to firm-specific attributes [1]. Choosing firms from the same industry improved predictions [26]. Benefits of contextual fundamental analysis were also demonstrated in investigating the subset of firms with extreme returns [6]. The importance of the contextual approach was enhanced by developing growth-oriented valuation measures separately for low BM stocks and applying traditional valuation factors for high BM stocks [21].
Although these works have demonstrated the advantages of identifying appropriate selections, few of them led to a systematic and scientific approach. The difficulties were pointed out to be the relative costs associated with collecting data [23]. Efforts include seeking reasonable industry classification mechanisms [9], and performing regressions on valuation models’ parameters to improve the objectivity in selecting comparable peers [8].

From a technical solution perspective, compared with the direct mapping from XML Schema to OWL [14] and a faithful translation of XBRL taxonomies into OWL DL [4], we went further to model the domain knowledge and enhance the diversity of LOFD.

VI. CONCLUSION

We explored leveraging Semantic Web technologies as systematic and objective approaches for choosing comparable firms in fundamental analysis. We investigated Linked Open Financial Data as an incremental data organization model at low cost. Relevant data from various sources are connected enabling simultaneous control over a range of explanatory variables. By applying ontology modeling, we represented both generic and domain-specific meanings along with the logic relations underlying accounting numbers used by fundamental analysis, making implicit information explicit and machine-readable. Collectively we explored how comparability is improved according to valuation rationales with flexible aggregation and disaggregation among data.

For future work, we will continue to study a fuzzy representation for uncertain factors, and improve the reasoning ability of our knowledge model to assist fundamental analysis.

ACKNOWLEDGMENT

We particularly thank Timothy Lebo, Li Ding and Ankesh Khandelwal for their insightful comments.

REFERENCES


APPENDIX

owl http://www.w3.org/2002/07/owl/
rdf http://www.w3.org/1999/02/22-rdf-syntax-ns/
dcterms http://purl.org/dc/elements/1.1/
audioscropper http://www.w3.org/2004/02/audioscropper/core/
foaf http://xmlns.com/foaf/0.1/
time http://www.w3.org/TR/owl-time/
ex http://www.example.com/
un-gaap http://xbrl.us/un-gaap/2009-01-31/
xs http://www.w3.org/2001/XMLSchema/
xbrl http://www.xbrl.org/2003/instance/
dollartag http://w3.org/2001/XMLSchema-instance/
ex http://www.w3.org/2001/XMLSchema-instance/
dollartag http://w3.org/2001/XMLSchema-instance/
ex http://www.w3.org/2001/XMLSchema-instance/
un100 http://data.gov.tw.rpi.edu/source/YahooFinance/dataset/us100hisndaily
xml http://data.gov.tw.rpi.edu/source/YahooFinance/dataset/us100hisndaily/vocab/enhancements