This essay discusses how to determine the validity of financial information. Financial information is essential for investors. In the capital market, companies are required to provide their financial statements to market regulators such as U.S. Securities and Exchange Commission (SEC). Market regulators collect and disseminate companies’ financial information to investors. Professional financial analysts compile and interpret financial data and other relevant information and analyze the current situation and future trends of the markets in various financial analysis reports, in particular, investment newsletters. These reports help investors make their investment decisions. An interesting question is whether those financial data can be believed.

In a speech on 3 March 2006, Christopher Cox, Chairman of U.S. Securities and Exchange Commission, reveals a surprising fact about the validity of financial data. Cox addressed that the error rate of the data used by financial analysts in their valuation models is up to 28% or higher. The direct reason for this problem is that the financial data used in financial analysis come from neither the companies nor SEC, instead the data is bought from intermediate companies that manually re-key financial statements into the data of suitable formats. Another reason from a technical view is that the financial data is mixed with text in financial statements, so that the data cannot be extracted automatically by machines.

The above situation yields the needs of “interactive data” (refer to Cox’s speech), machine-processable data that can be easily extracted from financial statements and customized into the reports that meet the different needs of financial information users. To this end, XBRL (eXtensible Business Reporting Language) provides a solution to this problem by allowing financial data to be marked up with semantic information, making it machine-readable and allowing for the creation of interactive reports.
Reporting Language has been developed to facilitate electronic communication of business and financial data. Many countries have initiated or suggested the use of XBRL. For examples, US SEC set up an “interactive-data test group” in which companies are using XBRL to make their SEC fillings (refer to Cox’s speech); UK Government proposes to make the use of XBRL mandatory for company tax filing from March 2010. The use of XBRL will make the data in companies’ financial statements be able to be easily extracted and make the data used in financial analysis reports be able to be validated by machines.

Another approach to extracting financial data is to use web data extraction technologies such as Lixto, by which targeted data can be extracted by machines from text in general web documents such as html even pdf files.

However, XBRL and web data extraction technologies alone cannot completely solve the validity problem in financial information. This is because these technologies only help to transform financial data automatically, but not to tell who create these data and whether the sources of data can be trusted. In addition, financial analysis reports usually contain not only companies’ financial data but also many other forms of information or knowledge such as news stories, personal interpretations about data, personal opinions, personal assumptions, assertions and derived results. Web data extraction technology can help to extract financial data from text by machines. The combination of knowledge provenance (to be introduced in the next section) and XBRL as well as web data extraction technologies can provide solutions for investors to determine the validity of financial information. In the following, we will discuss a solution that combines XBRL and knowledge provenance.

What is Knowledge Provenance?

Knowledge Provenance (KP) is an approach to determining the origin and validity of knowledge/information on the web by means of modeling and maintaining information sources and dependencies, as well as trust relationships. The major questions KP attempts to answer include: Can this information be believed to be true? Who created it? Can its creator be trusted? What does it depend on? Can the information it depends on be believed to be true?

As addressed by Wilson, “we can trust a text if it is the work of an individual or group of individuals whom we can trust”. The basic approach of knowledge provenance is to find the information source, and to judge whether

---

3 http://www.xbrl.org/Announcements/UK-XBRL22March2006.htm (visited on 1 Nov. 2006)
this information source can be trusted. In addition to trust, knowledge provenance also considers dependency relations among each piece of information.

We identify four modules in KP: Static KP \[4\] focuses on provenance of static and deterministic information; dynamic KP \[8\] considers how the validity of information may change over time; uncertain KP \[7\] (ch.5) addresses provenance problem where truth values and trust relationships are uncertain; trust module \[7\] (ch.6,7,8) focuses on judging the trustworthiness of information creators.

To use KP, information creators need to annotate web documents with KP metadata \(^{4}\) to describe the provenance-related attributes, such as who is proposition creator and what is the premise proposition on which this proposition depends; information users (provenance requesters) need to define their personalized trust relationships to tell whom they trust; an online KP software agent (a KP reasoner) \(^{5}\) will trace KP tags in web documents across web pages, combining information sources and dependencies, as well as trust relationships, to deduce the origin and validity of tagged information.

The basic unit of web information to be considered in KP is a “proposition”. A proposition, as defined in Propositional Logic, is a declarative sentence that is either true or false. A proposition is the smallest piece of information to which provenance-related attributes may be ascribed. An information creator may define a phrase, a sentence, a paragraph, even a whole document as a proposition. Not only text but also an xml element could be defined as a proposition.

The taxonomy of the propositions in KP is illustrated in figure 1. KP_prop is the most general class of propositions; an Original_prop is a proposition which content is created by the information creator(s); an Asserted_prop is an assert-

---

\(^{4}\) see KP ontology at http://www.eil.utoronto.ca/kp/2006/06/kp.owl

\(^{5}\) An online demo can be found at http://www.eil.utoronto.ca/kp/
tion that is not dependent on any other propositions; a Dependent\_prop is a proposition whose truth is dependent on other propositions; (correspondingly, the propositions a dependent proposition depends on are called the support propositions of this dependent proposition); a Derived\_prop is a derived conclusion based on some premises; an Equivalent\_prop is a copy and its truth value is the same as the proposition it depends on; a Compound\_prop could be the logical “and” / “or” / “negation” of other proposition(s).

Every Asserted\_prop or Derived\_prop has an “assigned truth value” that is the truth value given by the proposition creator, and its default value is “true”. Every KP\_prop has a “believed truth value” that is evaluated and trusted by a specific provenance requester. To handle uncertainty, subjective probability is introduced to represent uncertain truth value. Correspondingly, assigned/derived certainty degree is used for uncertain assigned/derived truth value.

The major rules of knowledge provenance are presented as follows.

**Rule 1 (authentication):** For any KP\_prop, it has an authenticated information creator, if and only if: either (1) the signature of this proposition signed by this information creator is validated; or (2) the proposition is a valid web publication of this information creator. For example, the url of the proposition shows that it is in the official web pages of the information creator.

**Rule 2 (trust and belief):** the belief degree of a proposition is the maximal degree of trust placed in an authenticated information creator of the proposition in a field covering that proposition.

**Rule 3 (truth value of asserted proposition):** the believed certainty degree ($bcd_x$) of an asserted proposition $(x)$ is dependent on: (1) the belief degree ($bd_x$) place on this proposition; (2) the assigned certainty degree ($acd_x$) to the proposition.

$$bcd_x = bd_x \cdot acd_x + (1 - bd_x) \cdot (1 - acd_x).$$  \hspace{1cm} (1)

**Rule 4 (truth value of derived proposition):** the believed certainty degree ($bcd_x$) of an derived proposition $(x)$ is dependent on: (1) the belief degree ($bd_x$) of this proposition; (2) the believed certainty degree ($bcd_y$) of the support proposition $(y)$; (3) the assigned certainty degree ($acd_x$) to the proposition.

$$bcd_x = bcd_y \cdot (bd_x \cdot acd_x + (1 - bd_x) \cdot (1 - acd_x)) + (1 - bcd_y) \cdot 0.5.$$  \hspace{1cm} (2)

**Rule 5 (truth value of equivalent proposition):** the believed certainty degree of an equivalent proposition is the same as the believed certainty degree of the proposition it depends on.
**Rule 6 (truth value of compound proposition):** the believed certainty degree \(bcd_x\) of an compound proposition is the probability for the believed truth values of this compound proposition to be true, e.g. \((x = y \land z)\)

\[
bcd_x = pr(BTV_y \land BTV_z).
\]  

(3)

where \(BTV_y\) and \(BTV_z\) denote the believed truth values of \(y\) and \(z\).

The derivation of these rules as well as the complete and formal description of KP can be found in [7]. The discussion about the semantics of trust can be found in [9].

**Example of Investment Newsletter**

Come back to the theme of how to determine the validity of financial information. In order to demonstrate how to use knowledge provenance and XBRL to solve financial information validity problem, we give a simplified example of investment newsletter as follows.

**Issued by:** W-finance.com

**Issued Date:** 15 November 2006

**Company:** Lockheed Martin Corporation (LMT-NYSE)

**Recommendation:** Buy

**Target Price:** $94.13

---

6 This sample is adapted from:
(1) Zacks Digest on Lockheed Martin Corp., October 26, 2006.
(2) [http://www.fool.com/News/mft/2006/mft06102549.htm](http://www.fool.com/News/mft/2006/mft06102549.htm)
(7) [http://www.sec.gov/Archives/edgar/data/936468/000119312506219311/lmt-20060930.xml](http://www.sec.gov/Archives/edgar/data/936468/000119312506219311/lmt-20060930.xml)
http://www.sec.gov/Archives/edgar/data/936468/000119312506219311/lmt-20060930.xsd

52wk High (2006-10-24): 89.89

52wk Low (2005-11-15): 59.55

Key Positive Arguments:
- Expansion of margins;
- Strong free cash flow and improving balance sheet;
- Big increase of profits and reduced tax rate;
- Won major contracts, for example, Orion project;

Key Negative Arguments:
- Uncertainty in defense policy.

Recent Events:
- On November 8, 2006, Defense Secretary Donald H. Rumsfeld stepped down, which clouds defense policy.
- Democrats wins 2006 Elections.
- On October 24, 2006, LMT released Q3 earnings. Third-quarter net income rose to $629 million, with a growth rate of 47%.
- On August 31, 2006, NASA selected Thursday Lockheed Martin Corp. as the prime contractor to design, develop, and build Orion, America’s spacecraft for a new generation of explorers. The estimated value for the project is $3.9 billion.

Other sections including Revenue Highlights, Margins Highlights, Cash Flows Highlight, and Balance Sheet Highlights are omitted for length limitation. Detailed example can be found in [7] (ch.9).

In this example, some information comes from LMT’s financial statements in the format of XBRL, and some others are news stories and derived conclusions.

In the following, we demonstrate how to annotate this simplified financial analysis report with KP metadata, how to validate the authentication of XBRL data, and how to make financial knowledge provenance reasoning.

Financial Information Annotation

Assume the above sample investment newsletter is published in HTML. In the following, we demonstrate how to annotate financial reports with KP
metadata. In order to focus on KP, we neglect most xhtml tags and only give KP metadata.

For general propositions (non-XBRL data) in financial reports are directly annotated with KP tags. The web ontology of KP[^7] defines a KP annotation language.

In the sample newsletter, **Recommendation**: Buy” is a derived proposition, which is dependent on the truth of the propositions given in arguments. This derived proposition is annotated as follows.

```
<kp:DerivedProp rdf:id="#RecommToBuy"
kp:isDependentOn = “#Argument-pos-1”
kp:isDependentOn = “#Argument-pos-2”
kp:isDependentOn = “#Argument-pos-3”
kp:isDependentOn = “#Argument-pos-4”
kp:isDependentOn = “#Argument-neg-1”
kp:inField = “investment advices”
>
  Recommendation: Buy
</kp:DerivedProp>
```

**Target Price**: $94.13” is an assertion about the target price of the recommended company. This asserted proposition can be annotated as follows.

```
<kp:AssertedProp rdf:id="#TargetPrice"
kp:inField = “investment advices”
>
  Target Price: $94.13
</kp:AssertedProp>
```

Other general propositions can be annotated in a similar way.

Now, we look at a special type of information in financial reports, XBRL data items. In a news report in “Recent events” section, “On October 24, 2006, LMT released Q3 earnings. Third-quarter net income rose to $629 million, with a growth rate of 47%”, “$629 million” is an XBRL data item coming from LMT’s third quarter financial statement in 2006. KP defines XBRL data items as a subclass of asserted propositions. So that, the above XBRL data item can be annotated as follows.

[^7]: http://www.eil.utoronto.ca/kp/2006/06/kp.owl
In KP, an XBRL data item has “believed truth value” of “true” if it is contained in a valid XBRL instance file.

Next, we discuss how to validate the authenticity of XBRL data items.

**Authentication of XBRL Data**

KP reasoner validates XBRL data items appeared in financial reports by making a XQuery. We present this process by continuing our the previous example as follows:

1. get the XBRL instance file containing the data in question. The url of the XBRL file is specified in the `inDoc` attribute of `XBRL_Item`. In this example, the url is: http://www.sec.gov/Archives/edgar/data/936468/000119312506219311/lmt-20060930.xml.

2. make a XQuery from the XBRL instance file to get item elements with tag “usfr-pte:NetIncome”. The query is:

   `doc(lmt-20060930.xml)/xbrl/usfr-pte:NetIncome`.

   `doc()` is the function that is used by XQuery to extract elements from an XML document. XQuery will return all elements with the tag.

3. check whether the questioned XBRL data item matches any XBRL element extracted by XQuery. The attributes need to be validated include: value, the unit and the context (entity and the period of the item). Validating context requires to make another XQuery to extract the context element by “con-

---

8 http://www.w3.org/TR/xquery/
textRef" attribute. Validating entity needs to check the entity’s identification number, typically, CIK issued by SEC.

Note that it is the information creator’s responsibility to make the annotated text have the same semantics as the metadata of the XBRL DataItem has; it is the information users’ responsibility to check the text and the metadata have the same context. KP reasoner only validate whether the metadata match the item in the specified XBRL instance file.

After the authenticity of this XBRL data item in question is successfully validated, if an information user trusts LMT about its financial statement, then this XBRL data item has truth value of true.

Provenance Reasoning

Now, we demonstrate how to make provenance reasoning for the propositions in the sample investment newsletter.

When an investor, “John”, reads the newsletter, he is interested in the message about Lockheed Martin, so he makes provenance request regarding the proposition “RecommToBuy” by using a KP reasoner; then KP reasoner makes provenance reasoning and answers John to what an extent this proposition can be believed to be true.

First, we assume John has the following trust relationships, which is inferred from John and his trusted friends’ inter-individual trust relationships 9.

\[
\begin{align*}
\text{trusted}_\text{in}(John, W, \text{“investment advice”}, 0.92). \\
\text{trusted}_\text{in}(John, W, \text{“finance analysis”}, 0.95). \\
\text{trusted}_\text{in}(John, W, \text{“news & comments”}, 0.88). \\
\text{trusted}_\text{in}(John, SEC, \text{“finance statement”}, 1.0). \\
\text{trusted}_\text{in}(John, NASA, \text{“news in NASA”}, 1.0). \\
\text{trusted}_\text{in}(John, CNN, \text{“news in US Politics”}, 0.99). \\
\end{align*}
\]

The information dependency among the propositions in the newsletter is illustrated in figure 2.

In provenance reasoning, KP reasoner handles a derived proposition with several support propositions as the derived proposition has one implicit support

\footnote{In order to focus on knowledge provenance, we omit how these trust relationships are derived from a social network.}
Fig. 2. Dependency relations in the sample investment newsletter

proposition and this implicit proposition is the conjunction of those support propositions.
The provenance reasoning process is given as the following steps.

(1) Provenance reasoning on derived proposition “RecommToBuy”. This proposition is authored by W whom provenance requester (John) trusts in the field “investment advices” in the degree of 0.92. Since this proposition is dependent on several other propositions, to determine the believed truth value of this proposition, KP reasoner needs to infer the believed truth values of these support propositions first.

(2) Provenance reasoning on derived proposition “Argument-Posi-1”. This proposition is dependent on four items of data in LMT’s financial statement in an XBRL instance file. To calculate the believed certainty degree of this proposition, the believed certainty degree of six XBRL items need to be calculated first.

As the first step, each XBRL item is validated (as discussed in the previous section) that the item is an item element in the specified XBRL instance file (lmt-20060930.xml); then, since this XBRL file is published in SEC’s official website, the XBRL file is believed in degree of 1.0, so each XBRL item is also believed in degree of 1.0; each item in the XBRL is claimed to be true by the information creator, i.e. assigned certainty degree \((acd)\) is 1.0; an XBRL item is an asserted proposition, so by rule 3, the believed certainty degree \((bcd)\) of each XBRL item is 1.0.

Given that the believed certainty degree of each support proposition is 1.0, and the degree of trust in W in field of “financial analysis” is 0.95, (so that the belief degree to proposition “Argument-Posi-1” is also 0.95 by rule 2), by rule 4, the believed certainty degree of the derived proposition “Argument-Posi-1” is calculated as 0.95.

(3) Provenance reasoning on derived proposition “Argument-Posi-2”. Similarly, the believed certainty degree of this proposition is calculated as 0.95.

(4) Provenance reasoning on derived proposition “Argument-Posi-3”. Similarly, the believed certainty degree of this proposition is calculated as 0.95.

(5) Provenance reasoning on derived proposition “Argument-Posi-4”. This derived proposition depends on another derived proposition “Event-LMT-Orion” given in recent events, which further depends on a news released by NASA. The believed certainty degree of “Argument-Posi-4” is calculated as 0.896 by a provenance reasoning process as shown in figure 3.

(6) Provenance reasoning on derived proposition “Argument-Neg-1”. This proposition is supported by two events and comments, which are
Fig. 3. Provenance reasoning on proposition “Argument-posi-4”

Fig. 4. Provenance reasoning on proposition “RecommToBuy” (continue)

Further supported by four pieces of news reports provided by CNN. By a provenance reasoning process, the believed certainty degree of this derived proposition is 0.844.

(7) Provenance reasoning on derived proposition “RecommToBuy” (Continue).

After the believed certainty degrees of five support propositions are calculated, as shown in figure 4, the believed certainty degree of this proposition can be calculated as 0.754, which shows the subjective probability to believe this investment recommendation.
Discussion

The Web has fast become an open decentralized global information/knowledge repository, where anyone is able to produce and disseminate information, so that the information on the Web may be true or false, current or outdated; however, few tools exist to discern the difference. Information validity has become a serious problem on the Web. For example, in 1999, two men posted fraudulent corporate information on electronic bulletin boards, which caused the stock price of a company (NEI) to soar from $0.13 to $15, resulting in their making a profit of more than $350,000. In order to solve this problem, methods and tools need to be developed to determine the validity of web information. From this background, knowledge provenance is proposed to determine the origin and validity of knowledge/information on the web by means of modeling and maintaining information sources and dependencies, as well as trust relationships.

Finance is a typical area where people are seriously concerned about the origin and validity of the information they get. In this essay, we introduced the basic concepts and methods of KP, and presented how to use KP and XBRL to determine the validity of financial information by a simple example. KP can be applied not only to financial reports but also to any other information service in finance such as news, financial digests, messages (electronic bulletin board) and blogs. By using KP in financial information, investors can learn where the information comes from and determine to what extent this information can be believed. Reconsider the story of fraudulent corporation information in BBS discussed earlier. If KP had been applied, people would have found that message comes from an untrusted information source, and that cheating event would not happen. KP can help to make financial information more trustworthy and also help investors to judge the validity of these information.

Several research projects relevant to knowledge provenance have emerged. Most recently, EU provenance project [6] proposed an open provenance architecture to enable documentation of provenance information in grid computing. This system mainly focuses on the documentation of data processing process, and provide tools for creating, recording and querying provenance information.

Buneman et al [3] proposed “Data Provenance” to answer two types of provenance of a piece of data generated by a database query: (1) why-provenance – why the returned data is there; (2) where-provenance – where the data come from. Data provenance actually focuses on the “provenance” of information retrieval process rather than the provenance of the information.

Coming from an automated reasoning perspective, KSL at Stanford [11] developed “Inference Web (IW)” which enables information creators to register proofs with provenance information in IW, and then IW is able to explain the provenance of a piece of requested knowledge. IW only provides provenance information (registered by creators) to information users, and the users make decisions whether to trust or not trust the requested knowledge. In addition, IW may be suitable for only formalized information rather than various forms of web data.

TRELLIS [5] is a web-based tool where users are able to judge the quality of a piece of information used in argumentation, by examining other users’ evaluations of the information in other argumentation, and by tracing back to the provenance of the information. TRELLIS considers the provenance of information, that is, the use and the original context of the yield of this information; but there are no formal models used for information validity judgments, and such judgments need to be made by human users.

As addressed in [2], “provenance information is extremely important for determining the value and integrity of a resource”. Provenance tells not only what is the information source but also how this information is derived, to what extent this information can be believed to be true, what is the context of this information, and how this information is used. However, processing rich provenance information needs human beings’ participation, so that research have to trade off the range of provenance information, what human beings do, and what machines do.

References


