2014

An Ontology-Based Information Extraction (OBIE) Framework for Analyzing Initial Public Offering (IPO) Prospectus

Jie Tao
Dakota State University

Amit Deokar
Dakota State University

Omar F. El-Gayar
Dakota State University

Recommended Citation

Tao, J., Deokar, A. V., & El-Gayar, O. F. (2014, January). An ontology-based information extraction (OBIE) framework for analyzing initial public offering (IPO) prospectus. In 2014 47th Hawaii International Conference on System Sciences (pp. 769-778). IEEE.

An Ontology-based Information Extraction (OBIE) Framework for Analyzing Initial Public Offering (IPO) Prospectus

Jie Tao
Dakota State University,
Madison, SD, USA, 57042
jtao16065@pluto.dsu.edu

Amit V. Deokar
Dakota State University,
Madison, SD, USA, 57042
Amit.Deokar@dsu.edu

Omar F. El-Gayar
Dakota State University,
Madison, SD, USA, 57042
Omar.El-Gayar@dsu.edu

Abstract

With the large amounts of information associated with the Initial Public Offering (IPO) process, an intelligent tool is needed for assisting the decision-making activities for both the investors and the underwriters. Even though a large body of related studies exists in extant literature, minimum attention has been devoted to the aspect of understanding hidden semantics within the informative contents of IPO prospectus. In this paper, we present a framework for processing the textual content of IPO prospectus based on an emerging technique named Ontology Based Information Extraction (OBIE). Preliminary results indicates that the framework is capable of meeting the design requirements identified. Moreover, lessons learned during the design and implementation span technical and organizational considerations and can serve as guidance for future research and development in related areas.

1. Introduction

Vast amounts of textual information is available in financial documents such as Initial Public Offering (IPO) prospectus. IPO refers to the process in which organizations raise capital for supporting business decisions through issuing stocks in the market. Under the regulations from the US Security and Exchange Commission (SEC), particular information needs to be disclosed through certain documents – these documents are termed as the IPO prospectus [1]. Rich text-based information is provided in the IPO prospectus. These documents are composed according to a similar structure; thus, identical sections are included in different documents. Among the different sections in the IPO prospectus, four of them have been recognized of more significance than the others with respect to their informativeness, namely: *Prospectus Summary, Use of Proceeds, Risk Factors, and Management's Discussion and Analysis* [2]. Number of studies have been conducted in order to suggest explanations to various phenomena related to the IPO process (e.g., changes in the price offerings) through understanding the relationships between these phenomena and the information disclosed in the prospectus [3], [4], [5]. Even though the prior studies have been able to provide insights into these IPO related phenomena, researchers have increasingly realized that the rich implicit knowledge snippets hidden in the textual parts of the IPO prospectus have been largely overlooked [2]. Several recent studies have focused on analyzing information content [6], [7], yet such studies have typically concentrated only on identifying relevant keywords/concepts (e.g. *risk, finance, loss*, etc.) from particular sections (e.g. *risk factors*, *Management's Discussion and Analysis*, etc.) in the IPO prospectus. As a result, the rich semantic information, such as the semantic relationships between such concepts, remains unrevealed. Further, multiple versions of the prospectus are released in a single IPO process, and each draft involves some crucial changes in the content and the changes in (estimated) price offers are potentially associated with these changes. Comprehending such associations would be helpful in understanding and assessing the validity of price changes in the offering process.

The overarching research question of this project can be stated as follows – "*How to extract meaningful information from IPO prospectus and utilize the hidden links between the informative contents to support various tasks associated with the IPO process, e.g., understanding and estimating changes in price offerings?*" In this paper, we present the design of an analytical framework for processing textual contents within IPO prospectus, in order to identify and extract hidden semantics based on a particular type of Information Extraction (IE) methodology, namely Ontology-based Information Extraction (OBIE). With ontology as the formalized conceptualization of the domain knowledge, not only the relevant concepts, but also the relations between them and the properties defining them, can be identified and added to the

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.103

domain knowledge base. Contextual and delta-analysis inferencing can be supported through the proposed design artifact. An analytical and reasoning dashboard will enable such analysis leading to knowledge discovery for user desired purposes. The knowledge nuggets extracted from the informative contents of IPO prospectus can assist average investors to better analyze IPO process (including tasks such as estimating the "true" price of a stock) as well as underwriters when composing the prospectus.

The paper is structured following the Peffers et al.'s [8] design science research presentation guidelines. The rest of this paper is organized as follows: In Section 2, two motivational scenarios are described, one from an investors' perspective and another from an underwriters' perspective. Following the motivation, the design requirements are elaborated for the research problem under consideration. A brief review of extant work on IE and IPO prospectus studies is provided in Section 4. Next, The structure and functionalities of the proposed framework, as well as evaluation of the same with respect to the design requirements are discussed in Section 5. Section 6 presents insights gained and lessons learnt during the design and development efforts in this research project. Section 7 describes future research avenues and concludes the paper.

2. Motivational Scenarios

Two typical scenarios are presented below for illustrating the motivations of this research study. These scenarios are selected from the perspectives of the (potential) investors and the issuer/underwriter involved in an IPO process. The (potential) investors refer to the ultimate buyers of the stock in the open market; the issuer is the firm that releases the stock to the market; while the underwriter is the agency that represents the issuer in the IPO process. The prospectus serves as the most credible source for the issuer/underwriter providing relevant information to the potential investors.

1. "Underpricing" and "overpricing" are two common phenomena in the IPO process. For instance, Facebook Inc. is a representative example of "overpricing" – the stock price has descended significantly since the initial offering. When a version of the prospectus is released, the market (i.e. potential investors) reacts to it by speculating the price of the final offer based on the information revealed in the document. It has been emphasized in the literature that analyzing informative contents within the prospectus would be beneficial in understanding the price changes in an IPO process [2]. The challenges are – *How to identify indicators for pricing changes from the implicit information hidden in the textual content of IPO prospectuses? Can these indicators be aligned with other price estimation indicators?* The average investors lack of expertise and domain knowledge from retrieving and comprehending such information; thus, a decision support framework can be useful to assist them in identifying aforementioned indicators, and ultimately useful in better estimating the value of the stock for investment decisions.

2. On the other hand, the underwriters representing the issuing firms also need a normalized knowledge structure as a reference for composing the prospectus documents. Underwriters play a very significant role in the IPO process [7]. Despite the attitudes and writing patterns among different underwriters, a prospectus of good quality requires certain expertise and insights into the domain. For example, compared to a veteran underwriter, a novice underwriter may lack adequate skills and experience, which could possibly lead to an unsuccessful IPO. In order to provide knowledge/decision support to the underwriters when assessing an IPO prospectus, a self-organized, frequently updating, and domain-specific knowledge structure can be a valuable resource.

3. Design Requirements

In regards to addressing aforementioned issues, existing studies in related disciplines have raised the need for a decision support framework to assist decision makers/stakeholders during the IPO process. Considering these domain-driven motivational needs, the following requirements ought to be satisfied within the design and development of a text analytics system:

- **To be able to create a formalized repository containing relevant domain knowledge:** IPO documents (i.e. prospectus) embed domain knowledge of distinct aspects, such as organizational descriptions, managerial information, competitiveness, business policies, and expectations of the future [2]. A normalized conceptualization should be created as a vault for such information in order to provide references/guidelines for analytical activities. This conceptualization should also be able to evolve by adding newly-discovered knowledge into it (semi-) automatically in an iterative fashion.
- **To identify relevant concepts/entities related to the IPO process from the prospectus:** Prospectus documents might contain several different expressions (i.e. "customers" and "buyers" might refer to the same entity) or in different parts of speech (i.e. "issue" and "issuance") of the same term. These different mentions of the same entity should be identified and then associated to the belonging concept as its instances correctly. Further, the extent of the mentions (i.e. the

span of the representation of a particular entity in the text) should also be determined to reduce information redundancy.

- **To discover relations between extracted concepts from the IPO prospectus:** Different types of relations may exist between entities in the textual content of the IPO documents. For instance, a term namely "granting award" might be a particular type of another global concept "employee benefit plan". Sometimes such relations are not explicitly defined in the textual content; however, these relations bear crucial information for investment decisions. A such, the proposed framework should be able to extract such entity relations.
- **To extract and reason about hidden semantic information in the prospectus:** Other than relations mentioned earlier, other types of linkages (i.e. co-dependency, causal links, etc.) may also exist between entities in the form of patterns. The proposed framework should be able to dig through the implicit information embedded in the textual contents of the prospectus able to extract such patterns and recognize potential linkages within them.
- **To support 'delta' analysis and contextual analysis of IPO prospectus:** During the time span from the release of the first version of the prospectus to the final offering, several versions of the prospectus are issued to the public. The content changes are posited to be meaningful for knowledge discovery purposes, which can be analyzed through the so-called 'delta' analysis of prospectus versions. Further, a prospectus typically contains several sections (i.e. *Prospectus Summary, Risk Factors, Management Discussion & Analysis*, etc.), and the appearances of same terms in different contexts (i.e. section, subsection, paragraph, sentence, etc.) may imply different semantic meanings requiring context-based analysis. Accordingly, the proposed framework should support both 'delta' analysis and contextual analysis.
- **To present the discovered knowledge in a normalized form and to support user-defined queries:** In addition to the aforementioned requirements regarding the knowledge discovery process itself, another important issue is that of presenting the results from such analyses to the end users. The proposed framework should support presenting output in the form of a normalized conceptualization; and allow the users to execute queries for question-answering purposes.

4. Related Work

4.1. Role of the prospectus in IPO process

As stated earlier, disclosing certain key information regarding the issuing firm is mandatory according to US federal regulations before any stock can "*go public*". Multiple versions of the prospectus are typically generated in this compliance process. Given that the prospectus is the most accessible and explicit source for the investor community interested in obtaining information regarding the issuing firm, analytical studies related to prospectus analysis have been conducted over the recent decades [7], [9], [10], [11], [12], [13].

Traditionally, researchers have shown interest in understanding the IPO process and the subsequent activities (i.e. post-issuance price changes, etc.) through identifying and analyzing the quantitative facts released in prospectus, as well as the roles of all involved parties [10].

Prior research that investigates IPO performances (both *ex-ante* and *post-hoc* of the stock offering) can be categorized into two types: the first type focuses on identifying the correlations between activities in the IPO process and its performance; while the second type concentrates on interpreting the performance with the roles of the involved parties in the process. For instance, Bhabra and Pettway [1] have conducted an empirical study – containing a classification and a regression analysis – to illustrate the correlations between the IPO process, subsequent stock returns (as a measure of the IPO performances), and the subsequent outcomes. Jain and Kini [14] have investigated the lifecycle of the IPO process and the firms' post-issue status. Similar studies can also be found in [4], [11], [13], [15], [16]. Some researchers have also argued that the involved parties' activities in the IPO process and attitudes toward the IPO process, such as issuers, underwriters, etc., have effects on the IPO pricing and performance [7], [9], [12].

Although abovementioned studies have provided interesting insights, mining and extracting informative contents from the prospectus has started to gain attention from researchers. For instance, Deumes [17] conducted a comprehensive content analysis in order to understand the disclosure of risks and the volatility of the stock price in the future of the firm. Arnold et al. [6] used an ambiguity model (based on the occurrences of the keywords) on the risk factors section of the prospectus and discovered that the ambiguous information regarding risks has a negative effect on the investors' decisions. Similar work is available in [2]. Although, without a formalized conceptualization of the domain knowledge, extant studies focus on merely discovering the relevant concepts but ignoring the relations between them and the context in which these concepts are mentioned. Moreover, there is no normalized representation of the results which can be

used to directly update the domain knowledge base. A decision support framework is designed and developed (as a prototype system) to leverage such issues.

4.2. Ontology-Based Information Extraction (OBIE)

In order to design an artifact to fulfill aforementioned requirements, IE, particularly OBIE, has been chosen as the design methodology for the proposed framework, which falls into a wider domain of text mining. Various definitions of IE can be found in the literature. Based on a synthesis study, IE can be defined as a process consisting of 4 steps [18], [19], [20]:

1) Isolating different textual elements in the natural language document(s);

2) Identifying different mentions of a particular class of concepts, relations, or events for a pre-defined purpose;

3) Extracting information regarding such concepts, relations or events;

4) Representing the extracted knowledge in a formalized structure.

To further understand the role of IE, let us imagine a spectrum with one pole as Information Retrieval (fetching relevant information for certain purposes) and the other pole as textual understanding (completely understanding the content of the documents). IE should be located in the middle of the spectrum [18] – which means only the relevant parts of the documents are processed in IE. Three types of knowledge are extracted in IE systems, namely entities, relationships between entities, and properties describing entities [21]. Although IE systems are typically designed for a particular purpose (thus have distinct structures), some of the common modules across different IE systems can include: text zoning, preprocessing, lexical analysis, filter, parsing, semantic interpretation and disambiguation, co-reference resolution, and template generation [22].

A widely-accepted definition of OBIE is that an OBIE system is a system that extracts particular types of knowledge from semi-structured/unstructured natural language texts and provides outputs guided by ontologies. OBIE is a particular type of IE which highly relies on the ontologies, which serve as the formal and explicit representation of domain knowledge, provide not only the guidelines for the extraction processes, but also the format and standardization for representing the outputs [23]. Particularly in knowledge-based IE systems, ontological information embedded serves as the basis for crafting extraction rules/patterns [24]. OBIE systems aim at processing unstructured/semi-structured natural language sources with the guidance from the ontology – the knowledge within the ontology provides assistance for the concept annotation/disambiguation purposes. To the best of our knowledge, there is no existing IE system for analyzing the prospectus documents.

5. Analytical Framework

The architecture and key features of the analytical framework are demonstrated next.

5.1. Demonstration

The analytical framework contains three modules, namely information extraction module, semantic reasoning and learning module, and analytics module, respectively. The architecture of the framework is depicted in Figure 1.

5.1.1. Information Extraction Module. The information extraction module, with an information extractor as the kernel, contains several activities, such as: pre-processing the prospectus, annotating the prospectus based on the pre-defined domain ontology, and populating the ontology with extracted entities and relations. With respect to the design requirements mentioned earlier, the pre-processing of the prospectus should contain segmentation of the document (for contextual analysis) and change tracking (for Delta analysis). Further, the extractor provides capabilities such as prospectus entity recognition, entity relation recognition, coreference resolution, and so forth, as shown in Figure 1.

5.1.2. Reasoning and Learning Module. The core of the reasoning and learning stage is the reasoner. In this stage, the extracted knowledge is used as the source for inferencing and ontology learning. The inferencing function, including subsumption-based reasoning, and semantic reasoning, is facilitated through the population/instantiation of the prospectus ontology.

Subsumption reasoning is employed for inferring classes and individuals in the hierarchical structure of the prospectus ontology – which are used in domain rule integration. The semantic reasoning relies on a set of logical/semantic rules which represents the domain knowledge. Such rules can be used to infer informative assertions with the help from the rule engine. Each of the rules is in the "IF-THEN" form; new knowledge is added to the domain knowledge base when the premises of the rules are met. Additionally, we adopt ontology learning techniques in this particular stage to automatically update the prospectus ontology from the results of the information extraction module.

5.1.3. Analytics Module. The third module in the analytical framework is the analytics module, which enables user-defined queries to retrieve relevant information from the domain knowledge base. A specific semantic querying language needs to be selected for realizing function. The kernel of this stage is the ontology query engine.

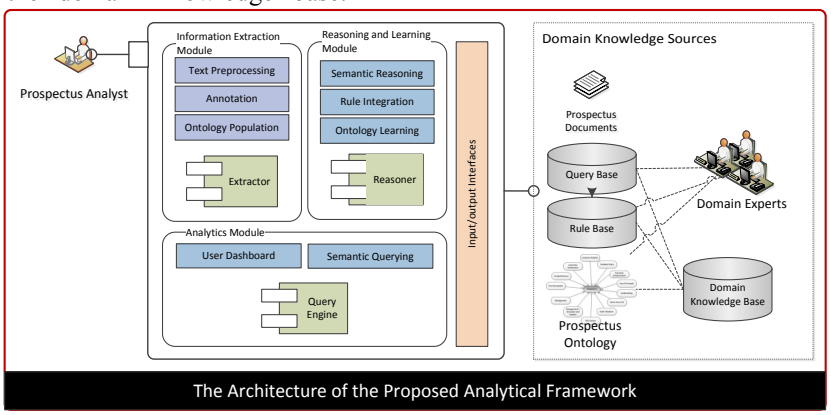

Figure 1. The Architecture of the Proposed Analytical Framework

5.1.4. Prospectus Ontology. The pre-defined ontology is constructed by domain experts including the authors. The initial ontology has 13 main classes: *Competitiveness, Corporate Governance, Customer Relation, Dividend Policy, Executive Compensation, Firm Description, Management, Management's Discussion and Analysis, Risk Factors, Stock Valuation, Terms of an IPO, Underwriting,* and *Use of Proceeds*. Several relevant keywords have been attached with each class. A partial view (containing the relevant classes) of the prospectus ontology is shown in Figure 2. The ontology serves as the core component in the proposed framework for both the annotation and population purposes. Protégé 3.5 beta [25] has been used for developing the Prospectus ontology in the OWL (Web Ontology Language) format.

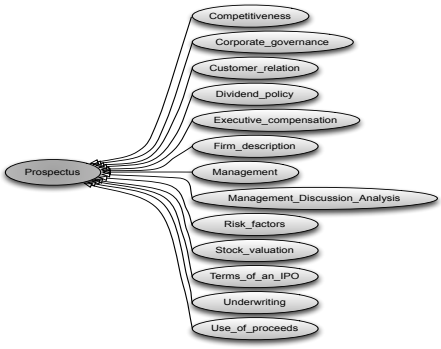

Figure 2. Overview of Prospectus Ontology

The implementation of the framework is discussed next.

5.2. Implementation of the Framework

5.2.1. Data Collection of the IPO Prospectus. The IPO prospectus documents are accessible through the SEC EDGAR database maintained by the SEC[26]. All documents required by the SEC are archived in this database. Users are allowed to search the documents by different fields (such as company names, Central Index Key (CIK), etc.) and then download the desired documents. The prospectus is coded as "S-1" and the amendment to the prospectus is coded as "S-1/A". In the prototype development stage, we have downloaded the prospectus and all its amendments from 5 companies in the Information Technology (IT) and service industries, which respectively are: *Newegg, Facebook, LinkedIn, Delta Airlines*, and *Google*. Among different formats of the prospectus, we have selected the HTML format in order to maintain consistency. These companies are selected as a convenience sample given that they provide S-1 documents in the desired format.

5.2.2. Implementation Environment of the Analytical Framework: GATE. The relevant information within the prospectus documents are annotated by a text-processing application built on "General Architecture for Text Engineering" (GATE). GATE is a comprehensive architecture supporting NLP engineering and providing IE implementations such as *tokenizers, sentence splitters, part-of-speech taggers, gazetteers, pattern-matching grammars, stemmers,* and *co-reference resolution* [27]. Some of these implementations have been modified to support the information extraction module in our prototype system. A grammar rule language named JAPE (Java Annotations Patterns Engine) is used in the preprocessing, annotation, and the ontology population

activities. Essentially, JAPE rules are finite-state transducers. Each transducer contains a different type of rules, while each rule has a Left-hand Side (LHS) and one (or multiple) Right-hand Side (RHS). The LHS is used for pattern-matching while the RHS could be used for annotation and other aforementioned purposes [28].

5.2.3. Pre-processing of the Prospectus Documents. Text pre-processing is one of the most important phases in most IE systems, and it provides annotations and other information for subsequent use. Particularly in our project, document segmentation and

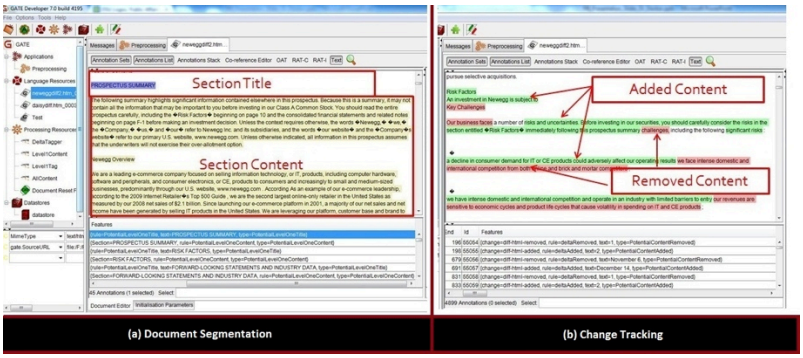

Figure 3. Pre-processing of IPO Prospectus

change tracking are two of the most significant activities in the pre-processing stage. In this case, different JAPE rules are coded and stacked as a pipeline in order to facilitate these activities.

The segmentation of the documents enables the later contextual analysis. In this phase, contextual information is added along with the extracted knowledge to the Prospectus ontology. The JAPE rule for such function first identifies the section titles and then the section contents – then the relative position of the extracted entities is determined. The prospectus contents in the prototype are identified at the section level (i.e. *prospectus summary, risk factors*, etc.). The implementation of such function in GATE is shown in Figure 3 (a).

As mentioned earlier, multiple versions of a given prospectus are released during the IPO process. Between these different versions, we posit that changes in content would have potential associations to certain phenomena in the IPO process (e.g. changes in offer prices). To investigate this proposition, tracking content changes across different versions of the prospectus is crucial in our analytical framework – it is also enabled in the pre-processing stage. A freeware tool named *daisydiff* [29] has been employed for comparing different versions of the prospectus; then JAPE rules are encoded to annotate the changed content and the model of change (*added/removed*, etc.). The implementation of change tracking in GATE is illustrated as above (Figure 3 (b)).

5.2.4. Semantic Annotation of the Prospectus Documents Based on the Prospectus Ontology. The semantic annotation of the prospectus documents, following the phases of entity recognition, relation recognition, and the attributes identification, is implemented in GATE. The ontology-based gazetteers (provided within GATE) in combinition with a plug-in to GATE, namely APOLDA (Automated Processing of Ontologies with Lexical Denotations for Annotation), have been applied to provide semantic annotation of documents following the paradigm of lexicon-based annotation [30]. The APOLDA plugin is more suitable dealing with a large amount of concepts (classes) with less textual representations, while the ontology-based gazetteers are more useful when the ontology has fewer classes. The results of the semantic annotation of the prospectus document are shown in Figure 4.

5.2.5. Instantiation of the Prospectus Ontology. The semantic annotations of the discovered entities consist of several characteristics (termed as features in Figure 4(a)) such as: the name of the belonging ontological class, the URI of the ontological class, the span of the annotation (in the form of start and end nodes), annotated text, the location of the text (in which sections), and the change type of the text if available (e.g. added, removed). Another set of JAPE rules are coded to populate the Prospectus ontology by adding the annotations (in the form of ontology instances) and aforementioned features (in the form of datatype properties). The output of the instantiation process is depicted in Figure 4(b) below. The left part of Figure 4(b) includes ontological classes (pre-defined, same as those in Figure 2) and ontology instances. The right part of Figure 4(b) are the features picked from the semantic annotations in the prospectus documents.

5.2.6. Rule-based Reasoning. A large body of semantic reasoning and querying languages is available in the extant literature. For the seamless interoperability with selected tools (Protégé and GATE), SWRL (Semantic Web Rule Language) has been chosen as the basis for enabling semantic reasoning in the proposed framework. SWRL "allows the users to write rules expressed in terms of OWL concepts to reason about OWL individuals" [31]. The rules are composed in an antecedent-consequent pattern – the antecedent is termed as the head of the rule while the consequent is named as the body of the rule. Both the head and body consist of smaller blocks, namely atoms. SWRL rules, along with the existing assertions in the knowledge base, could be used for inferring new assertions. These rules are used to derive relationships between existing entities, or inferring new entities in the ontological structure. Implementing such rules in the domain rule base enhances the domain knowledge base with rich contextual information derived from the prospectus documents. In addition, such contextual information provides the basis for enabling analytics for decision-supporting purposes. The SWRL tab is provided within Protégé 3.5 beta for the users to compose and apply the SWRL rules. To execute such rules, a rule engine needs to be employed as shown in Figure 1. In this framework, we have selected the Jess rule engine for executing encoded reasoning rules. Jess is a rule engine based on the Java platform. To utilize Jess, the logic should be defined in the format of Jess rules or XML (Extensible Markup Language) and the data that the rules operate on [32]. The Jess rule engine can easily be installed on the analyst's machine and then provide support for the execution of SWRL rules in Protégé 3.5 beta.

5.2.7. Ontology Querying. The Semantic Query-enhanced Web Rule Language (SQWRL) is built on SWRL and thus select for enabling the semantic querying functionality in the proposed framework. A SQL (Structured Query Language)-like query language is provided to facilitate querying abilities for retrieving knowledge from OWL [33]. The SWRL tab in Protégé also provides the interface for the users to cast such queries.

5.3. Validation of the Proposed Framework

In this section, the processes and metrics used for evaluating the proposed artifact are discussed. Other than adopting traditional IE metrics (e.g. *Recall*, *Precision*, and *F-measure*, etc.), we have also considered some evaluation metrics specific to our project. Here, we present the validation of the proposed framework against the design requirements discussed earlier.

5.3.1. Validation Against Design Requirements. In addition to validating the artifact against abovementioned metrics, we need to investigate that the proposed artifact has addressed all the design requirements listed in Section 3.

- With the prospectus ontology created, we have established a formalized conceptualization for knowledge in this specific domain. All the relevant concepts mentioned earlier, along with the relations between them and the properties describing them would be included in the ontology through ontology instantiation and learning. An automated population mechanism has also been developed in order to update the domain knowledge within the ontology.
- With the assistance of GATE (particularly ontology-based gazetteers and APOLDA), various textual representations of relevant entities can be extracted from prospectus document, as well as the extent of mentions.
- With the subsumption reasoning enabled by the ontology, the relations between relevant domain concepts could be identified. Also, the semantic reasoning (via SWRL) facilitates extracting other hidden linkages (e.g. causal, temporal, etc.) between the concepts.
- With the ontology population and learning mechanism, it is possible to decrease manual effort in the annotation, reasoning, and analytics activities.
- The Delta analysis and the contextual analysis are well supported in the proposed framework, as required. A freeware tool named *daisydiff* has been employed for comparing different versions of the prospectus; then a JAPE rule is coded to annotate what have been changed and how they have been changed (e.g. *added/removed*, etc.) in the contents. On the other hand, a different JAPE rule has been written for dividing the prospectus documents into different sections (e.g. *prospectus summary, risk factors, MDA*, etc.) in order to enable contextual analysis.
- With the prospectus ontology created, the extracted knowledge is incrementally added to it in the form of ontological classes, relations, or properties. Further, the results from the reasoning and querying activities are also represented in these forms. Thus, the outcomes from the artifacts are in formalized representations and could be directly imported to the domain knowledge base.
-

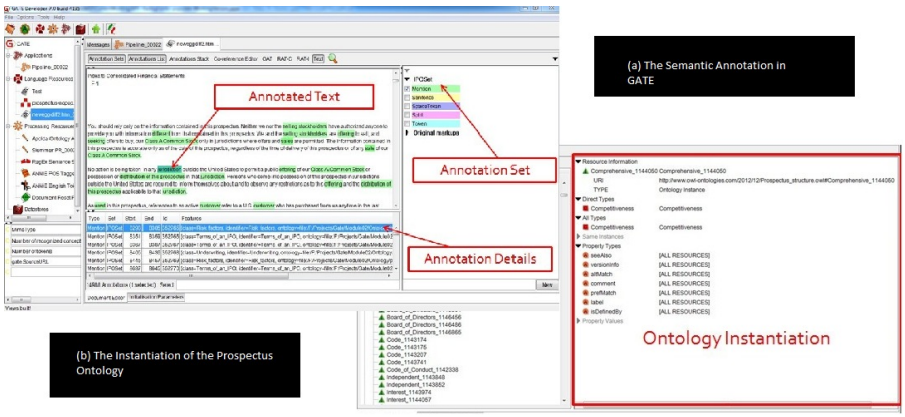

Figure 4. Semantic Annotation and Ontology Instantiation

6. Lessons Learned

During the design and implementation of the proposed framework, several issues have been identified that need to be concerned in relevant future studies. These issues might lead to future research opportunities in various domains. They may also be further developed into guidelines for developing future similar applications. Lessons learned from this project could be categorized into two groups: technique-oriented and application-oriented.

6.1. Technique-oriented Issues

Technique-oriented issues refer to the issues that are related to the technique we used to develop the proposed framework, which is in particular, OBIE.

Firstly, ontology learning has become increasingly important with respect to OBIE process, due to the domain knowledge and expertise required in constructing and managing ontologies. Average end users would benefit from (semi-)automated ontology learning by reducing these barriers via the development of generic guidelines, the adoption of ontology learning achievements in OBIE applications, and the approaches that increase the accuracy and efficiency of ontology learning in OBIE [34]. In the proposed framework, we incorporate ontology learning in the reasoning and learning module – we believe that by enabling such functionality in the proposed framework, it would be capable of updating the domain knowledge base by adding new knowledge nuggets in the context of IPO.

Secondly, ontology integration in the context of OBIE should receive more attention. Current OBIE systems often rely on a single ontology (e.g. only the prospectus ontology has been used in the proposed framework), yet typically a business scenario usually involves multiple ontologies (e.g. An "organization" ontology is used to depict the organizational structure of a firm, and an "activity" ontology is used to reflect the tasks within the firm's business processes). Additional efforts should be devoted to develop new ontology integration approaches in order to leverage issues such as ontology collaboration, conflicts handling, and so forth [35], [36], [37]. A new type of ontology, namely "*bridge ontology*", has been developed for such purposes [38].

Thirdly, novel, project-specific evaluation metrics need to be developed in OBIE projects – current metrics are mostly adopted from the Information Retrieval (IR) field (such as the ones used in this project: *recall, precision,* and *F-measure*). New measures that are exclusive to the process of OBIE, and are derived from the goals of the project, should be developed in order to provide both cross-sectional and case-wise evaluations to OBIE systems.

Other than aforementioned issues, attention should be given to developing intelligent user interfaces to lower the adoption barriers, enhancing change control toward ontologies in OBIE projects, improving the performance of co-reference resolution through the use of ontologies, and so forth.

6.2. Application-oriented Issues

The application-oriented issues fall into two main categories: the issues regarding applying the proposed framework in the context of IPO, and the issues regarding adopting similar OBIE applications to other domains.

As discussed in Section 2, the major use cases of the proposed framework include: estimating the short-term price offerings in the IPO process, and providing referential support for the underwriters when composing an IPO prospectus. In regards to the aspect of price estimation, the proposed framework would enhance the studies toward understanding the pricing phenomena (such as the "*underpricing*"

phenomena and the anormalies in price offerings, etc.) by providing: 1) a formalized conceptualization that is more reliable and efficient when comparing to the manual efforts of the domain experts; 2) a mechanism that enables the antomated processing of the prospectus (which is highly suggested as one of the most significant future steps in [1]) and more thorough knowledge discovery within the prospectus with consideration of cross-version and cross-section analyses (comparing to the SVM-based method proposed in [2]). On the other hand, the prospectus ontology would become an rich domain knowledge base via the ontology learning and ontology instantiation features from the proposed framework. Thus, we will be able to derive a set of practical guidelines from the prospectus ontology and use them for providing decision support to the underwriters when assessing prospectus documents.

At a more global level, applications similar to the proposed framework could be used in a variety of domains, including bioinformatics, medicine, semantic web, and process analytics (e.g. process mining, policy enforcement, business intelligence).

7. Conclusion

IPO prospectus is one of the most reliable sources from the point of view of exchanging crucial information between the investors and the issuers/underwriters and as such, understanding the phenomena in the IPO process holds significant value for knowledge discovery purposes. Such information is disclosed via the issuance of the prospectus – which contains voluminous textual information across different versions. In this paper, we have proposed a formalized framework to facilitate the knowledge discovery process in order to align the hidden link between the changes in the informative contents of the prospectus and the (short-run) trends of the price offerings in the market. The proposed framework is constituted by three major modules: the Information Extraction module, the Reasoning and Learning module, and the Analytics module. The implemented IE module, which prepares the prospectus, provides the semantic annotations to the prospectus based on the pre-defined ontology, and updates the ontology with the discovered information in an iterative and incremental fashion. The overall framework, implementation of the Information Extraction Module, and preliminary results are demonstrated in this article, along with a brief evaluation against design requirements to validate the functionalities and efficiencies of the proposed framework. The proposed framework can be useful for the average potential investors to identify the indicators of the price changes from the textual contents in the IPO prospectus. Also, the self-organized prospectus ontology can serve as a knowledge repository for assisting the underwriters with assessing IPO prospectus. We believe that both researchers and practitioners can derive value from the proposed framework.

7.1. Directions for Future Research

As for the future steps, first we need to fully implement the proposed framework by adding the other two modules to the IE module. The ontology learning function in the Reasoning and Learning module is the key feature which is novel – incorporating ontology learning, including developing in IE applications would increase ease-of-use from the end user's perspective. On the other hand, the Analytic module facilitates querying and displaying the discovered knowledge to the users based upon their demands. We also need to enhance domain knowledge base by adding more prospectus documents (as testing data set), updating the Prospectus ontology (via ontology learning mechanism), and constructing the rule base/query base (with the help from the domain experts). Third, since the ultimate goal of the proposed framework is to supporting the pricing decisions in the IPO process, we plan to use the framework on the testing data set to better understand the "underpricing" phenomenon in the IPO process. This can be achieved by building on extant studies by aligning the informative contents to the "underpricing" phenomenon [1], [2] through applying the proposed framework as an alternative analytical tool.

References

[1] H. S. Bhabra and R. H. Pettway, "IPO Prospectus Information and Subsequent Performance," The Financial Review, vol. 38, no. 3, pp. 369–397, Aug. 2003.

[2] K. W. Hanley and G. Hoberg, "The Information Content of IPO Prospectuses," Review of Financial Studies, 2012.

[3] V. Babich and M. J. Sobel, "Pre-IPO Operational and Financial Decisions," Management Science, vol. 50, no. 7, pp. 935–948, Jul. 2004.

[4] M. Lowry and G. W. Schwert, "Is the IPO pricing process efficient?," Journal of Financial Economics, vol. 71, no. 1, pp. 3–26, Jan. 2004.

[5] S. Varshney and R. Robinson, "IPO Research Symposium Review," Journal of Economics and Finance, vol. 28, no. 1, pp. 56–67, 2004.

[6] T. Arnold, R. P. H. Fishe, and D. North, "The Effects of Ambiguous Information on Initial and Subsequent IPO Returns," Financial Management (Blackwell Publishing Limited), vol. 39, no. 4, pp. 1497–1519, 2010.

[7] S. P. Ferris, G. Q. Hao, and S. M. Liao, "The Effect of Issuer Conservatism on IPO Pricing and Performance," Review of Finance, pp. 1–45, 2012.
[8] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," Journal of Management Information Systems, vol. 24, no. 3, pp. 45–77, 2008.
[9] R. B. Carter, F. H. Dark, and A. K. Singh, "Underwriter Reputation , Initial Returns , and the Long-Run Performance of IPO Stocks," The Journal of Finance, vol. LIII, no. 1, pp. 285–312, 1998.
[10] C. M. Daily, S. T. Certo, D. R. Dalton, and R. Roengpitya, "IPO Underpricing: A Meta-Analysis and Research Synthesis," Entrepreneurship: Theory & Practice, vol. 27, no. 3, pp. 271–296, 2003.
[11] J. R. Ritter, "The Long-Run Performance of Initial Public Offerings," The Journal of Finance, vol. 46, no. 1, pp. 3–27, 1991.
[12] P. Roosenboom and J. Thomas, "How Do Underwriters Value Initial Public Offerings? An Empirical Analysis of the French IPO Market," Contemporary Accounting Research, vol. 24, no. 4, pp. 1217–1243, Dec. 2007.
[13] H. I. Silverman, "Qualitative Analysis In Financial Studies : Employing Ethnographic Content Analysis," Journal of Business & Economics Research, vol. 7, no. 5, pp. 133–136, 2009.
[14] B. a. Jain and O. Kini, "The Life Cycle of Initial Public Offering Firms," Journal of Business Finance and Accounting, vol. 26, no. 9–10, pp. 1281–1307, Nov. 1999.
[15] Y. Kim and A. Heshmati, "Analysis of Korean IT startups' initial public offering and their post-IPO performance," Journal of Productivity Analysis, vol. 34, no. 2, pp. 133–149, Apr. 2010.
[16] R. Rajan and H. Servaes, "Analyst following of initial public offerings," The Journal of Finance, vol. 52, no. 2, pp. 507–529, 1997.
[17] R. Deumes, "Corporate Risk Reporting: A Content Analysis of Narrative Risk Disclosures in Prospectuses," Journal of Business Communication, vol. 45, no. 2, pp. 120–157, Apr. 2008.
[18] D. Appelt, "An introduction to information extraction," Artificial Intelligence Communications, vol. 12, no. 3, pp. 161–172, 1999.
[19] J. Cowie and W. Lehnert, "Information Extraction," Communications of the ACM, vol. 39, no. 1, pp. 80–91, 1996.
[20] R. Grishman, "Information extraction: Techniques and challenges," Information Extraction A Multidisciplinary Approach to an Emerging Information Technology Lecture Notes in Computer Science, vol. 1299, pp. 10–27, 1997.
[21] S. Sarawagi, "Information Extraction," Foundations and Trends in Databases, vol. 1, no. 3, pp. 261–377, 2008.
[22] J. R. Hobbs, "The generic information extraction system," in Proceedings of the 5th conference on Message understanding, 1993, pp. 87–91.
[23] D. C. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," Journal of Information Science, vol. 36, no. 3, pp. 306–323, Jun. 2010.
[24] A.-P. Manine, E. Alphonse, and P. Bessières, "Information Extraction as an Ontology Population Task and Its Application to Genic Interactions," in 2008 20th IEEE International Conference on Tools with Artificial Intelligence, 2008, pp. 74–81.
[25] "Protege 3.5 Beta Release Notes," 2012. [Online]. Available: http://protegewiki.stanford.edu/wiki/Protege_3.5_Beta_Release_Notes. [Accessed: 01-Sep-2012].
[26] SEC, "SEC EDGAR," 2013. [Online]. Available: http://www.sec.gov/edgar.shtml.
[27] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: an architecture for development of robust HLT applications," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, no. July, pp. 168–175.
[28] H. Cunningham, D. Maynard, and V. Tablan, "JAPE: a Java Annotation Patterns Engine," 2000.
[29] G. Van den Broeck and D. Dickison, "Daisydiff," 2011. [Online]. Available: https://code.google.com/p/daisydiff/.
[30] C. Wartena, R. Brussee, L. Gazendam, and W.-O. Huijsen, "Apolda: A practical tool for semantic annotation," in Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA '07), 2007, pp. 288–292.
[31] M. O. Connor, H. Knublauch, S. Tu, B. Grosof, W. Grosso, and M. Musen, "Supporting Rule System Interoperability on the Semantic Web with SWRL," The Semantic Web – ISWC 2005 Lecture Notes in Computer Science, vol. 3729, pp. 974–986, 2005.
[32] Sandia-National-Laboratories, "Jess, The Rule Engine for the Java Platform," 2008.
[33] M. O'Conner and A. Das, "SQWRL: a Query Language for OWL," in Proceedings of OWLED 2009 OWL: Experiences and Directions. Sixth International Workshop, 2009, vol. 23, no. Owled, pp. 3–10.
[34] A. Gómez-Pérez and D. Manzano-Macho, "An overview of methods and tools for ontology learning from texts," The Knowledge Engineering Review, vol. 19, no. 03, pp. 187–212, Jun. 2005.
[35] L. Reeve and H. Han, "Survey of semantic annotation platforms," in Proceedings of the 2005 ACM symposium on Applied computing - SAC '05, 2005, pp. 1634–1638.
[36] D. C. Wimalasuriya and D. Dou, "Using multiple ontologies in information extraction," in Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09, 2009, pp. 235–244.
[37] M. M. Wood, S. J. Lydon, V. Tablan, D. Maynard, and H. Cunningham, "Populating a Database from Parallel Texts Using Ontology-Based Information Extraction," in in Natural Language Processing and Information Systems, 2004, pp. 254–264.
[38] B. Xu, P. Wang, J. Lu, Y. Li, and D. Kang, "Bridge Ontology and Its Role in Semantic Annotation," in Proceedings of the International Conference on Cyberworlds (CW '04), 2004, pp. 329–334.