An Ontology for Representing Financial Headline News

Sehl Mellouli, Faouzi Bouslama, Aichath Akande

Management Information Systems Department, Université Laval, Quebec, Canada

Abstract

This paper presents the development of an ontology to represent financial headline news. This ontology is developed using the On-To-Knowledge methodology where the focus is on the design steps of the Knowledge Meta Process. This development is part of an ongoing project which aims to design a virtual stock market simulator based on multi-agent systems. The proposed ontology has 31 concepts and includes 201 attributes. The testing results conducted on reliable headline news show that 99% of these headline news can be properly represented by the attributes of the right category in the ontology. Unreliable headline news characterized by news having uncertainties, incompleteness, ill-definition, or imprecision cannot be represented by the proposed ontology. Approaches for representing these unreliable headline news are discussed.

Key words: Ontology, Financial headline news, Knowledge Meta Process methodology, Reliable headlines, Uncertainties

1. Introduction

The stock markets have always been complex to understand and to predict. Recently, the "Sub-prime crisis" collapsed most of the world's financial markets. Before it occurs, there were probably various indicators collected from sources of information which could have signalled the crash. If these news were properly extracted, analyzed, classified and represented, the tendencies of the financial market would have been predicted and decisions-makers would have been alerted beforehand. For example, eighteen months before the crisis happened, the real-estate market in the US lost around 20% of its value and the American Federal Reserve gradually raised its intervention rate from 1.25% to 5.25% between 2004 and 2006 [1].

In seeking to describe the basic categories and relationships of entities and concepts of any stock market, one can think of creating a knowledgebase of the domain. By using an ontology, one can represent the financial information with all of its complexity and relationships. This paper is an attempt to develop a well-structured ontology to represent various types of financial headline news. There are already ontologies being used to represent some aspect of the financial domain. In one application [2], a financial ontology has been developed to describe symbols of stock markets. In another application [3], called the Financial Exchange Framework' (FEF), an ontology has been designed to model the Financial Exchange Theory (FET). FET proposes a base for the definition of a several financial entities including international financial products and financial transactions. Though these financial ontologies describe part of the financial aspects, they lack the description of the structure of the financial news.
Moreover, these ontologies do not provide the type and the characteristics of the financial news. To develop a well-structured ontology, this paper proposes to use the On-To-Knowledge Methodology [4]. The Tom Grubers five principles [5] for ontologies design are used in the validation process.

The design and testing of the ontology are conducted using headline news extracted from Reuters. The ontology has 31 concepts and 201 attributes. A sample of 227 headline news composed of 136 reliable news and 91 unreliable news is used for testing. Unreliable headline news are characterized by uncertainties, incompleteness, ill-definition or imprecision. The tests are conducted in two stages. First of all, the category of the headline news to be tested is identified, then in a second stage the information in the body text making the headline news is mapped into the attributes of the identified category. There are two sets of attributes in the proposed ontology: general or common attributes shared by all categories, and specific attributes for each sub category. A test is successful if the information of the headline news is properly mapped to the attributes.

The testing results conducted on reliable headline news show that 99% of these headline news can be properly represented by the attributes of the right category in the ontology. Unreliable headline news cannot be represented by this ontology since it does not formally model any unreliable concept. However, this paper provides an insight on how to model unreliable headline news for further expansion of the proposed ontology. This research work is part of an ongoing more comprehensive project which aims to design and develop a virtual stock market simulator based on multi-agent systems. Agents have all to gain from having a common understanding of the financial news. Agents can therefore take informed decisions on selling or buying stocks. The research work described in this paper focuses on the development of the ontology for financial news representation. Forthcoming papers will show how agents can use this ontology to comprehend and process any financial news.

The rest of the paper is organized as follows. Section 2 introduces ontologies and the Knowledge Meta Process methodology. Section 3 shows the development process of the financial headline news ontology. Section 4 gives the testing results. Finally, Section 5 is a discussion and a conclusion.

2. Ontology Development

Ontologies have been widely used in Knowledge Engineering (KE) and Artificial Intelligence (AI) to structure the concepts of a domain. An ontology is represented as a knotted tree of linked concepts or classes. A class can have subclasses which represent concepts more specific than the superclass. The properties of each concept describe the characteristics and the attributes of the concept and the attributes restrictions. An ontology and the individual instances of the classes constitute a knowledge base.

There are several techniques to design and develop ontologies [4] [6]. Each of them encompasses a number of fundamental rules and specifications to guide developers in the design of proper ontologies. In this paper, the "On-To-Knowledge Methodology" [4] is chosen for two reasons. Firstly, it conforms to the IEEE 1074-1995 standard [7] which helps ultimately achieve quality ontologies. Secondly, the methodology is appropriate as the developed ontology is part of an ongoing project to design a virtual stock exchange market in which intelligent agents are used to trade stocks. Based on the available financial news, these agents are expected to take decisions on the buying and selling of stocks. Therefore, they are expected to be able to automatically extract and interpret data from these news. Moreover, one of the core ideas of using On-To-Knowledge is to be able to extract meta data semi-automatically from documents and subsequently to use them for querying and browsing [4].

The On-To-Knowledge methodology is divided into two distinct processes: the Knowledge Meta Process focusing on development, and the Knowledge Process relating to the post implementation process of the ontology. This research focuses on the Knowledge Meta Process as the present work is in a developmental stage. The Knowledge Meta Process consists of five main steps [4]: Feasibility study, Kick-off, Refinement, Evaluation, and Application & evolution. The feasibility study step consists of identifying the resources and the potential solutions. At the end of the study, a decision has to be taken on whether it is possible to continue or not. The Kick-off phase provides a semi-formal description of the ontology. The refinement step consists of refining the results of the Kick-off phase. Then the ontology is evaluated. Many techniques have been developed for ontologies design evaluation such as the Tom Gruber's "Five principles" which includes
clarity, coherence, extendibility, minimal encoding bias, and minimal ontological commitment [5]. Finally, the ontology is deployed and can evolve over time. Each of these steps requires to be validated before moving to the next step. Each of these steps requires to be validated before moving to the next step.

The development of the proposed ontology involved the expertise of two IT professors, a Master student, and an expert in the financial domain, all based at Universite Laval. It took the developers five months of partial time to complete their work.

3. Financial news ontology.

The first step of the Knowledge Meta Process methodology is the project scope and feasibility study. This development requires the use of information sources on financial news. Several of these sources are available. In this research study, Reuters is chosen as a source of information since Universite Laval provides a direct access to Reuters financial news. As the financial domain covers a wide range of sub-domains that are plagued with news having precise and complete as well as ill-defined and uncertain pieces of news, it is expected that the ultimate ontology be complex and quite involved.

The second step in the Knowledge Meta Process is the financial news extraction and kick-off. The information to be represented by the ontology needs to be delimited. To this end, a corpus of one thousand financial headline news is extracted from the Reuters’ site. These news are selected from four key sectors of the Canadian economy which are the telecommunication, the auto industry, hydrocarbon (oil) distribution, and the raw materials sectors, respectively.

For each sector, a set of different news related to one particular organization or a company are extracted. The collected headline news were published between January 2007 and September 2008. The headlines are then categorized by sectors and by companies. These headline news are neither in a standard format nor on the same subject. For example, the following financial news which reads "11-Feb-2008 15:35 CNW & Aliant launches public awareness and education campaign to mark global Safer Internet Day" is about a social implication of the company Bell Aliant. There are various types of headline news. Some relates to the social activities, some are of economic nature and others depict the financial news. Moreover, these headline news refer to different concepts and data types, and one of the challenges in developing an ontology for these news is how to discover those concepts for fitting these different headline news into a single ontology. The building of this ontology is shown in the following steps.

The third step is the development of the ontology. This step is part of the "Kick-off" and "Refinement" steps of the Knowledge Meta Process methodology. A combined approach is used in the development of the ontology. The first attempt in the development of the ontology is to organize the news in a homogeneous manner. Hence, the way to proceed is to identify the semantic of the news and to group them under different subjects.

The first attempt of this organization provided four categories in which the financial headline news can be grouped. Figure 1, shows the four categories of news:

- Social category: These headline news are related to social external implications of a firm which have no direct economic effects on its revenue such as the voluntary work, the support for associations which defend humanitarian causes, lawsuits in which a firm is implicated, etc. Here is an example of headline news in this category: Aliant launches public awareness and education campaign to mark global Safer Internet Day.
- Economic category: These headline news are related to investments and strategies of a firm in order to improve its services and increase its capital and profits. The following extract is an example of headline news in this category: Toronto stocks set to fall on inflation jitters.
- Financial category: These news are related to financial stock markets behaviours and tendencies. Example of news extract in this category is ASTRAL MEDIA INC < ACM a.TO > Q3 SHR C$0.67.
- Environmental category: These headline news represent both the ecological news, and the decisions and actions that a firm takes to help save the environment. Example of news extract in this category: Program details announced for ecoRevolution Conference- Industry experts to present at international symposium on business and the environment in Moncton, NB on November 27.

A second round of analysis is done by refining each category. The refinement results in either the creation of new categories or the merging of existing ones. For example, the Environmental headline news can be considered as part of Socials headline news since they are related to social actions of firms linked
to the environment. Financial headline news are also considered as Economic headline news because they represent the evolution of companies shares traded on stock markets which have a direct impact on the capital, the profits and the assets of those organizations. As for the Economic category, more groups in the form of subclasses are created where a clear distinction is made between news related to mergers, acquisitions and agreements of a firm, stock-markets evolution, products and services, etc. Figure 2 illustrates the graphical representation of the different hierarchies of these economic subclasses.

The news in the sub-classes are further analysed to reveal new groups. In fact, in the "Stock Market" category for example, two distinct types of headline news can be distinguished. On one side, there are headline news which provide specific information on the evolution of stock markets such as the NYSE (New York Stock Exchange) or the TSX (the Toronto Stock Exchange), and on the other side, there are other headline news specific to firms’ daily, weekly, monthly or annually financial results. Thus, two news classes are created in the "Stock Market" to represent this knowledge. Figure 3 shows the Stock Market’ class and subclasses.

The financial news concepts are gradually refined by adding more samples of headline news in the study. Figure 4 shows the results of these refinements which depicts the overall structure of the proposed financial headline news ontology.

After the identification of the concepts of the ontology, the focus shifts to their attributes. General information and key words which characterize each headline news are then identified. In fact, common attributes to all headline news have been noticed. These attributes are the name of the company related to the news, the date of extraction, the symbol of the source name, and the category to which the news belongs to.

For example, this headline news “14-Feb-2008 15:35 CNW Aliant launches public awareness and education campaign to mark global Safer Internet Day” can be mapped to the common attributes as follows:

- Company name: Bell.
- Date of extraction: 14-Feb-2008
- Source name: CNW
- Category related to: Socials
After the identification of general attributes of financial news, the focus is on particular and specific attributes related to the subclasses. For "Stock Exchange" sub-category for example, the specific attributes actor, verb action, complement are identified. The following example of headline news extracted from Bell associated news ("Toronto stocks set to fall on inflation jitters") is identified as belonging to the "Stock Exchange" category. Moreover, the information inherent in this news can be mapped to the specific attributes as follows:

- actor: Toronto stocks
- verb action: set to fall
- complement: on inflation jitters

Another example of headline extracted from Astral which states Toronto stock index seen rising on record oil price is mapped as follows:

- actor: Toronto stocks
- verb action: seen rising
- complement: on record oil price

When considering the "Agreement" category, all the attributes of "Stock Exchange" category are also true. In addition the attribute partner name is added. For example, the attributes of the following headline news which reads Astral joins CHUM in Canada digital radio project are mapped as follows:

- actor: Astral
- verb action: joins
- partner name: CHUM
- complement: in Canada digital radio project

Here is another example showing the mapping of the headlines information to the specific attributes of the "Agreement" category. The headline news extracted from Bell reads Alliant and Solution Inc Sign multi-Year contract for Alliant Hotspot Management:

- actor: Alliant
- verb action: sign
- partner name: Solution Inc
- complement: multi-Year contract for Alliant Hotspot Management

The values that the attributes can take are usually restricted. For example, the values that the hour attribute can take are between 0 and 24. After the identification of general attributes of financial news, the focus is on the particular and specific attributes related to the subclasses. For "Stock Exchange" sub-category for example, the specific attributes "actor", "verb action", "complement" are identified. For Agreements, all of these attributes are also valid plus the attribute "partner name".

This process is carried on while using all the news related to each class. Finally, the classes, subclasses and attributes in the ontology are closely checked to guarantee the comprehensiveness and integrity of the ontology. The final ontology is composed of 31 concepts and includes 201 attributes.

4. Tests and evaluation of the ontology

4.1. testing results

The testing phase follows the design and development of the proposed ontology. A sample of 227 financial headlines news related to different firms and sectors of economy is used in the testing process. Here are some samples of headline news used in the testing:

- 17-Jun-2008 08:30 RTRS Air Canada reduces fall and winter capacity in response to record price of fuel.
- 02-May-2008 17:03 RTRS UPDATE 3-Toronto stocks climb 200 pts on robust resources
- 03-Jun-2008 13:30 CNW CGI announces Open Source software to support EPA’s National Environmental Information Exchange Network

By conducting a thorough investigation of the headline news used for testing, one can notice that these news can be split into two categories: reliable headline news category and an unreliable headline news category. Reliable headline news are the headlines that are clear, precise and concise. Unreliable headline news can be characterized by any of the following features: uncertainties, incompleteness, ill-definition, or imprecision. An uncertain piece of news is deduced by the lack of information to ascertain whether it is true or false. An imprecise news gives a true information but data is missing to accurately describe it. An incomplete news do not have all the parts for its completeness whereas an ill-defined news does not make a sense or it is not properly defined. Out of the total number of headline news used for testing, there are 91 unreliable headline news which are distributed as shown in Figure 5: 50 are imprecise, 2 are uncertain, 30 are ill-defined and 8 are incomplete.

There are 136 reliable headline news out of total of 227 headline news used for testing. They constitute around 60% of the total of the headline news. The unreliable headlines makes the remaining 40%, a quite considerable percentage of the total. These unreliable headline news are not used in the testing process of the ontology for two reasons. First, the on-
ontology is built based only on reliable news. Second, the ontology in its current format does not include representation of any concepts related to the four features of unreliability such as the uncertainties or the incompleteness. This challenging work is an ongoing future extension of the proposed ontology.

Figure 6 shows the distribution of the reliable headline news in their respective four categories: Socials, Economics, Politics and Announcement. As shown, economics news constitutes almost half of the news sample. This is due to the fact that the economic category has the most subcategories which are Stock exchange, Companies financial results, Currency, Investment, Mergers acquisitions agreement, Products, and Services.

The tests are conducted as follows. Take for example the following financial news which was extracted from Air Canada associated news: “UPDATE 3-Toronto stocks climb 200 pts on robust resources”.

First of all, the category of the headline news to be tested is identified. This example represents a news in the Stock Exchange category since it relates to “Toronto stocks”. The objective of the testing will then shift the focus to the mapping of the headline news inherent information to the category’s attributes. Therefore, the pieces of information in the headline news are checked whether they map to the attributes of the identified category. A test is successful if the information of the headline news is properly mapped to the attributes. For this example, the mapping of the information to the attributes of the Stock Exchange category is as follows:

- **News:** 02-May-2008 17:03 RTRS UPDATE 3-Toronto stocks climb 200 pts on robust resources
- **News attributes:**
  - Date of extraction: 02-May-2008
  - Source name: RTRS
  - Company name: AIR CANADA.
  - Category related to: Stock exchange
- **Stock exchange attributes:**
  - actor: Toronto stocks
  - verb action: climb
  - complement: 200 pts on robust resources

The test results show that the ontology can represent 99% of reliable headline news. It is worth noting that some financial headline news could not be properly represented by the proposed ontology. For example, the following financial news “BMO RAISES ASTRAL MEDIA <ACMa.TO> PRICE TARGET TO C$51 FROM C$46, KEEPS OUTPERFORM” has two verbs “ Raises” and “Keeps”. The ontology cannot represent a news having multiple verbs in its structure.

The ontology is only a static representation of the financial headline news domain. As stated earlier, in the testing process, the category of each headline news is a priori identified. However, in reality, some news may belong to two or more different categories. The proposed work in this paper does not deal with such news. It will be considered in future work when the ontology will be automatically populated. The population of an ontology requires the understanding of each piece of information so that it can be mapped to the right concept. It necessitates additional capabilities such as natural language processing to solve the automatic ontology population.
4.2. Ontology structure evaluation

The Tom Gruber’s five principles emphasises the important aspects that an ontology needs to observe. The five principles are presented hereafter, and a justification on how the proposed financial news ontology fulfills these principles is given.

- **Clarity:** "An ontology should effectively communicate the intended meaning of defined terms. Definitions should be objective. Where possible, a complete definition (a predicate defined by necessary and sufficient conditions) is preferred over a partial definition (defined by only necessary or sufficient conditions). All definitions should be documented with natural language." [8] The proposed ontology fulfills this criterion. Indeed, the name given to the defined concepts and attributes characterize the financial domain, but their definition is independent of the financial context. Indeed, for example the Economic, Environment, Politics or Stock market concepts can be associated with another domain such as the Banking domain for example. Similarly, the names of the attributes are also independent of the domain. Indeed, the attributes Company name and Date of extraction can be associated with any firm evolving in any industry sector.

- **Coherence:** "An ontology should be coherent: that is, it should sanction inferences that are consistent with the definitions. If a sentence that can be inferred from the axioms contradicts a definition or example given informally, then the ontology is incoherent". [8] Each financial headline news represented by the attributes of a category belongs to that category and can not be in another one.

- **Minimal encoding bias:** "The conceptualization should be specified at the knowledge level without depending on a particular symbol-level encoding. An encoding bias results when, a representation choice are made purely for the convenience of notation or implementation. Encoding bias should be minimized, because knowledge-sharing agents may be implemented in different representation systems and styles of representation". [8]. The definition of the ontology, the concepts and the attributes are all independents of the tool Protege 2000 [9] used to develop the ontology. Indeed, any other ontology development software can be used to represent the same classes, attributes and instances defined in this financial headline news ontology without biasing the comprehension and the interpretation of the represented knowledge.

- **Minimal ontological commitment:** "An ontology should require the minimal ontological commitment sufficient to support the intended knowledge sharing activities. An ontology should make as few claims as possible about the world being modeled, allowing the parties committed to the ontology freedom to specialize and instantiate the ontology as needed". [8]. As both tests results reveal, the financial news ontology covers a minimal ontological commitment. Indeed, since the rate of representation of reliable news is 99% and not 100%, and since the rate of reliable news versus unreliable ones is 60% and not 100%, the financial headline news ontology can be specialized as needed.

- **Extendibility:** "An ontology should be designed to anticipate the uses of the shared vocabulary. It should offer a conceptual foundation for a range of anticipated tasks, and the representation should be crafted so that one can extend and specialize the ontology monotonically". [8] Additional attributes can be easily added to the financial news ontology without changing its basic structure. Indeed, if the exact hour, minutes, seconds and microseconds of extraction, respectively, are relevant information, attributes representing them can be added to the News attributes. Also, as the tests show a 40% of unreliable news in the sample chosen to test the ontology, it is clear that the financial news ontology can be extended to model these particular news.

5. Conclusion and Discussion

This paper presented the development of a well-structured ontology for the representation of financial headline news. This ontology included 31 concepts and 201 attributes. The testing was conducted using a set of 227 samples of news composed of 136 reliable and 91 unreliable headline news, respectively. The proposed ontology successfully represented 99% of reliable news. For unreliable headline news, the capabilities of the proposed ontology must be extended to represent the complexity emanating from unreliable headline news. Hence, there is a need to extend this research work to include the representation of knowledge that can be imprecise, ill-defined, incomplete or uncertain.

Several modeling techniques have been applied in the development of ontologies to represent unreliable knowledge. These models are generally based
on Bayesian Networks (BN). The OntoBayes Model [10], the probabilistic Description Logic Program (PDLP) [11] and PR-OWL [12] are some examples of these approaches. The OntoBayes Model [10] is an ontology-driven uncertainty model which uses probability and dependency-annotated Web Ontology Language (OWL) to represent uncertain information in Bayesian Networks structures. The OntoBayes model provides an extension of ontologies which has the capability of capturing uncertainty knowledge about concepts, properties and relations in domains and supports reasoning with inaccurate information. The PDLP model combines two approaches; the probability and the description logic. This model applied to Logic programming and algorithms techniques provides a fast reasoning of systems such as the Tourism Ontology Uncertainty Reasoning system (TOUR). PR-OWL is an extension of OWL to represent probabilistic ontologies. It provides a framework for authoring probabilistic ontologies and is based on the Bayesian first-order logic called Multi-Entity Bayesian Networks. These models require the prior knowledge of all of the likely situations of the domain under investigation. But in this world full of unpredictable and volatile knowledge, what should constitute the unrealistic part in order to take a decision that must be taken at a precise moment?

In fact, the "belief-function" approach models human decision behaviour and value judgment more naturally than the traditional probabilistic approach [13]. This theory provides a more realistic representation of the world. It was proposed to model quantified belief and "it provide alternatives to the models based on probability functions or possibility functions" [14]. The power of this theory comes from its ability to represent any form of uncertainty: full knowledge, partial ignorance and total ignorance (and even probability knowledge) [15]. It also allows to express total and partial ignorance. Moreover, it is a generalization of the Bayesian inference. Under the belief function, one can evaluate not only the state of nature but also all possible combinations of these states of nature. Another significant advantage of belief functions theory is the combination rule (Dempster’s Rule) to model dependencies between variables and to combine information obtained from multiple sources. Consequently, the theory of beliefs functions to represent the unreliable financial headline news will constitute the subject of investigation for the authors to carry on with full comprehensive development of the ontology.

References