Conceptual Graph Formalism for Financial Text Representation

Siti Sakira Kamaruddin
College of Arts & Science, Universiti Utara Malaysia, 06010 Sintok, Kedah
Faculty of Technology & Information Science, Universiti Kebangsaan Malaysia, 43600 Bangi , Selangor
sakira@uum.edu.my

Abdul Razak Hamdan
Faculty of Technology & Information Science, Universiti Kebangsaan Malaysia, 43600 Bangi , Selangor.
arh@ftsm.ukm.my

Azuraliza Abu Bakar
Faculty of Technology & Information Science,Universiti Kebangsaan Malaysia, 43600 Bangi , Selangor.
aab@ftsm.ukm.my

Fauzias Mat Nor
Faculty of Economy & Business, Universiti Kebangsaan Malaysia, 43600 Bangi , Selangor.
fauzias@pkrisc.cc.ukm.my

Abstract

We present an approach to automatically transform a financial text into conceptual graph formalism. The approach exploits the constituent structure of sentences and general English grammar rules to perform the transformation. We suggest face validation and traces as the evaluation method to be performed on the resulting formalism to validate its accuracy. We also discuss the potential manipulation and application of the constructed conceptual graph database.

1. Introduction

Natural language text are flooding out the repository of world knowledge to a greater extend as opposed to structured databases. The great challenge in this area is to represent text into a more reliable representation to facilitate future retrieval and processing. Most previous work in this field focuses on vector representations, which has a statistical basis and represents words in isolation ignoring the sequence in which the words occur [1].

Researchers in the field are beginning to give importance on a richer representation scheme, which is proven to yield promising results. Following these developments, a number of network languages were employed to model the semantics of natural language and other domains. In our research we utilize conceptual graph [2], a particular network language proposed by John F. Sowa and show how this formalism may be used to represent meaningful knowledge from financial text.

The nature of financial text is different from the natural conversational language. It contains variety of morphologies and synonyms. However, the financial terms and jargons of the domain are normally repeated through out the text. This gives an advantage to the process of analyzing them. Besides that, further empirical observation on the financial text reveals that there are rare cases of semantic ambiguities occurring and the words and phrases in the text can be classified into limited number of groups. These groups follow a specific syntactic pattern similar to general grammar rule. Therefore, we have incorporated the grammar rule in the process of transforming the sentence structure of financial text into graphical knowledge representation scheme. This grammar rule is explained detailed in section 3.2 of this paper.

As a possible solution for the discussed research problem, this work is concerned on the efficient representation of financial text using conceptual graphs. The remainder of the paper is structured as follows. The details regarding conceptual graph

978-1-4244-2328-6/08/$25.00 © 2008 IEEE

formalism is explained in Section 2. Some related works are also presented in this section. Section 3 discusses the approach used to generate the conceptual graphs. Validation of the conceptual graphs is described in Section 4. Section 5 and 6 describe the discussions and conclusion respectively.

2. Conceptual graph formalism and related works

Conceptual Graphs (CG) are finite, connected, bipartite (Involving two elements: concepts and relations) graphs. A graph is comprised of a set of vertices or nodes and edges. As a contrary to other network languages, the edges are not labeled. These edges can be weighted to show its importance and to facilitate further manipulation of the graphs but in this work we favor simple graphs because it is sufficient enough to efficiently represent the problem domain.

In our work the CG represents relations between words. The vertices represent either concepts or conceptual relations and the edges are connections between them. The advantages of using conceptual graph formalism are first, it simplifies the representation of relations of any arity compared to other network language that used labeled arc. Second, its expressions are similar to natural language. Third, they are adequate to represent accurate and highly structured information beyond the keyword approach [3] and fourth, both semantic and episodic association between words can be represented using CGs [2].

Karalopoulos *et al.* [4] presents a simple method of creating a general form of CG and processing geographic text to match that general form. As a result they have generalized the geographic text into uniform CG to be used for further processing. The nature of the geographic text makes the method presented in their work feasible. Hensman & Dunnion [5] used CG representation for indexing XML documents. The information about the index is embedded as a meta-tag in the document. They have presented a two step approach; identifying semantic roles of sentences then using the role together with semi-automatically compiled domain specific knowledge to construct CGs. The accuracy of their work depends on the existence of words in the linguistic resources that they used.

To summarize, graph based representation are versatile and flexible. It is suitable for a wide range of problems in natural language processing, information retrieval and text mining. There are various other works that have used CG representation to capture the structure and semantic information contained in free text [6-10]. Like numerous other researchers, we are convinced that the financial text can be efficiently represented using conceptual graph formalism.

3. Representation of financial text as conceptual graphs

This section describes how the financial texts were transformed into CGs. Figure 1 illustrates the processes and components involved.

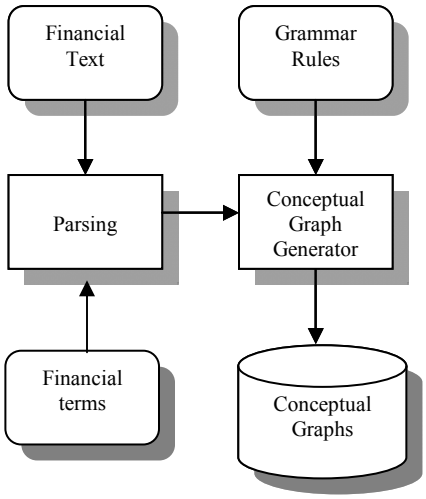

Figure 1. Process of transforming text into conceptual graphs

The process begins with the collection of financial texts being parsed by the parser. Additional financial terms were incorporated during parsing. The results from the parser which is the sentence structure were input into a conceptual graph generator. The generator uses grammar rules to construct the CGs. Some samples of textual documents relating to financial domain were collected from annual reports of companies to be used in this work. Figure 2 shows an extract of the samples.

BUSINESS PLAN STRATEGY

The Malaysian economy recorded a growth of 5.4 percent in 1999 and is expected to improve to 6 percent in the year 2000 and to be further sustained in the year 2001. The growth is attributed to higher domestic demand brought about by the recovery in private investment and consumption, sustained external demand or export and low interest rate regime.

The recent increase in oil price will likely slow down the global economic performance and affect Malaysia's export performance, build up inflation and increase interest rate. In the banking industry, loan growth is gradually picking up and liquidity is still ample.

Figure 2. A sample of financial text

978-1-4244-2328-6/08/$25.00 © 2008 IEEE

3.1 Parsing

The parsing was implemented using the Link Grammar Parser (LGP) [11] a formal grammatical system to produce syntactical relations between words in a sentence. We have incorporated additional financial terms in the parser's dictionary to cater for the special needs arising in the problem domain as discussed in section 1 of this paper. We propose to use the LGP because; there exist a structure similarity to conceptual graphs hence it is easier to map the obtained structure to conceptual graphs [12]. Suchanek *et al.* [13] reported that the LGP provides a much deeper semantic structure than the standard context-free parsers. Figure 3 shows the linguistic structure produced after parsing the example text using LGP.

Original sentence (before parsing)

"The Malaysian economy recorded a growth of 5.4 per cent in 1999"

Linguistic structure (after parsing)

```
[S [NP Malaysian economy NP]
    [VP recorded
        [NP [NP growth NP]
          [PP of
                [NP 5.4 percent NP] PP] NP]
        [PP in
             [NP 1999 NP] PP] VP]
```

Figure 3. Sentence structure

The parser has identified phrases and has categorized the phrase into: S which represents sentences; NP represents Noun Phrases; VP represents Verb Phrases and PP represents Preposition Phrases.

3.2 Conceptual graph generator

The sentence structure generated during the previous steps shows the syntactic level of sentence decomposition. This structure was traversed from its roots to generate the CG. The following grammar rules were the basis of the CG generator.

S → NP VP
NP → [DET] [N-MOD]…NOUN [PP]
N-MOD → ADJ | NOUN
VP → VERB [NP] [PP]
PP → PREP NP

Additional abbreviations used in this rules can be defined as follows: PREP represents prepositions, DET represents determiners, N-MOD represents noun modifier and ADJ represents adjectives. The square brackets represents options and the '…' suggests that there could be more of the previous combination of sentence structures. Based on this grammar rules, a decision process model for CG generator was developed and the model was further used to implement the generator. Figure 4 presents the model.

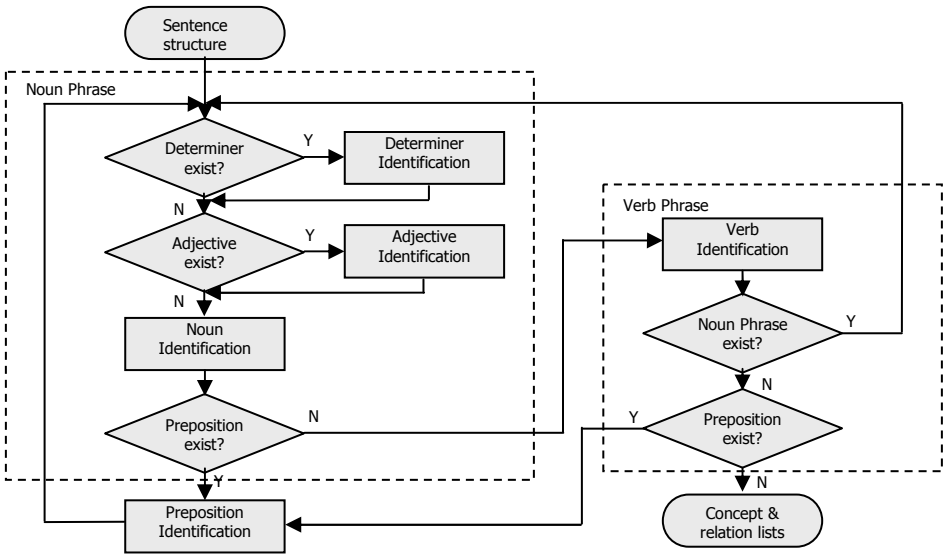

Figure 4. Decision process model for conceptual graph generator

978-1-4244-2328-6/08/$25.00 © 2008 IEEE

The model comprises steps to be taken in the process of analyzing the sentence structure. The identification of noun, verbs and adjectives will assist in building the concepts while prepositions are used to identify the relationship between the built concepts. The determiners role can be ignored for the simple CG that were constructed in this work. The results of the generator were reformatted to a more readable form and were stored in the CG database. Figure 5 shows some of the CG generated in its linear form for the sample text in Figure 1.

```
[recorded] –
        (agnt) -> [Malaysian economy]
        (obj) -> [growth] -> [of] -> [5.4 percent]
        (in) -> [1999]
[expected] –
        (agnt) -> [Malaysian economy]
        (to) -> [improve] –
                (to) -> [6 percent]
                (in) -> [year 2000]
[attributed] –
        (agnt) -> [growth]
        (to) -> [higher domestic demand] –
                    (is) -> [brought about] –
                      (by) -> [recovery] –
                           (in) -> [private investment
                                          and consumption]
                      (by) ->[sustained] –
                                (agnt) ->[external demand
                                             or export]
                      (by) ->[low interest rate regime]
[slow down] –
        (agnt) -> [recent increase] -> (in) -> [oil price]
        (obj) -> [global economic performance]
[effect] –
        (agnt) -> [recent increase] -> (in) -> [oil price]
        (obj) -> [Malaysia's export performance]
        (obj) -> [built up inflation]
        (obj) -> [increase interest rate]
```

Figure 5. Sample of generated conceptual graphs

4. Conceptual graph validation

The produced CGs have to be evaluated to determine that they are reasonable and correctly represent the semantic of the sentences. It should be able to substitute the sentence without jeopardizing the meaning of the original sentence. The ultimate aim of the evaluation process is to have conceptual graphs containing set of well-defined and non-ambiguous semantic relations. Hence, it includes determining if the appropriate detail and aggregate relationships have been used for the intended purpose, and if appropriate its structure, logic, and mathematics can be accurately derived from the CG representation.

As the CGs in this work are produced by extensive use of syntactic information obtained form the process of parsing, the results are highly correlated with the produced sentence constituents. If the constituents are correct then the resulting conceptual graphs is guaranteed to be correct because of two reasons; first the parser is proven to be effective in correctly identifying sentence structures [14]; second, the algorithm that generates it is based on the CG generator decision process model, which was developed from standard grammar rules.

However, we will perform a final evaluation on the resulting conceptual graphs to further validate its accuracy. The validation technique that will be used in this work is face validation and traces. Face validation refers to evaluation using experts on the problem domain. We will randomly pick the produced CG's together with its corresponding text and have the experts examine the CGs to determine if it is correct and reasonable. We will also employ traces where each concept is tracked through the CGs to determine if its relations were correctly defined and the accuracy was maintained.

5. Discussions

The representational language employed in this work is based on conceptual graphs. They belong to a network language and they became popular for their visual capability of the knowledge they represent. The domain expert are aware that numerous hours spent on reading and understanding textual documents can be decreased by representing them in a more structured formalism as proposed in this work. It is consented by most that conceptual graphs eases interactions between human and the knowledge base.

Besides benefiting from the ease of interpretability, conceptual graphs can also be manipulated for further processing because it inherits the mathematical foundation of graph theory. More formally a conceptual graph $G = (V, E)$. V is partitioned into two disjoined sets consist of Vc a set of concept nodes and Vr a set of relation nodes, $e \in E$ is an ordered pair that connects an element of Vc to an element of Vr. New graphs can be created either by generalizing or specializing from existing graphs. A number of operations such as projection (graph matching), unification (join), simplification, restriction and copying can be performed on the produced CG.

978-1-4244-2328-6/08/$25.00 © 2008 IEEE

The most widely used operations on the produced CG are graph clustering and graph matching and the result are established for various purpose. For example, clustering method can be performed on CG to detect regularities. In [15], the researcher showed the potentials of CG to be used as a special indexing scheme for text collection and able to assist in the discovery of trends, association rules and deviations. Representing text with CG formalism eases the process of comparing information contained in text by performing CG matching. One such application of CG matching is in semantic search. In [16], the authors proposed a CG matching algorithm that detects the semantic similarity between concepts and relations to improve the precision and recall in the searching process.

Additional information such as descriptions and the organization of the graphs into hierarchies of abstraction can help to reduce the search space and facilitate further analysis. CG is proven to be competitive with and more expressive than the logic-based method [17]. CGs are also utilized in software engineering discipline to represent user requirements, build design specification, proposing frameworks and model verification [9]. In the medical field, various medical text are transformed into CGs such as the work reported in [7] and [14], where the CG were used to capture the structure and semantic information contained in free text medical documents.

In this section, we do not intend to review exhaustively the application of CGs but rather as an illustrative view of the potentials of CGs to further justify the reasons for representing financial text with this formalism. Our aim for using this representation scheme on financial text is towards performing text mining and subsequently detect deviations in the financial texts in order to extract relevant knowledge of the outlying and contradicting financial reporting. Eventually any extraordinary financial reporting will be detected and treated as a possible new knowledge.

In order to detect the deviations among a set of conceptual graph, we will perform similarity measures on the CGs by comparing both concept and relation nodes. For this purpose we will use the conceptual graph clustering method proposed in [18]. Using this method, the overlap between two conceptual graphs is measured by calculating conceptual similarity and relational similarity. The expectation of our future work is to achieve the same results as proven in [18], i.e. regarding CG as an index of the text collections; detection of patterns not only rare graphs; and to be able to visualize the deviations from different level of generalization.

6. Conclusions

The work presented here focuses on the representation of financial text using conceptual graphs. Developing an automatic generator for transforming a financial sentence structure into the corresponding conceptual graph representation breaks many limitations and obstacles in the extraction of financial text and facilitates the implementation of financial text mining programs.

This research contributes to different areas such as natural language processing, information retrieval, and text mining, which benefits from accurate representation of text contents. Furthermore, it lays the foundation for the exploitation of conceptual graph's potentials in order to identify and formalize homogeneity and heterogeneity between financial text and further facilitates the process of manipulating a knowledge base of financial conceptual graphs.

7. References

[1] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," *ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations*, vol. 2, 2000.

[2] J. F. Sowa and E. C. Way, "Implementing a semantic interpreter using conceptual graphs," *IBM J. Res. Develop*, vol. 30, pp. 57-69, 1986.

[3] I. Ounis and M. Pasca, "A Promising Retrieval Algorithm For Systems based on the Conceptual Graphs Formalism," presented at Proceedings of the Intrernational Database Engineering and Applications Symposium IDEAS'98, Cardiff, Wales, UK., 1998.

[4] A. Karalopoulos, M. Kokla, and M. Kavouras, "Geographic Knowledge Representation Using Conceptual Graphs " presented at 7th AGILE Conference of Geographic Information Science, Heraklion Greece, 2004.

[5] S. Hensman and J. Dunnion, "Automatically Building Conceptual Graphs using VerbNet and WordNet, ," presented at Proceedings of the 2004 international symposium on Information and communication technologies ISICT '04, Nevada, USA 2004.

[6] T. Amghar, D. Batistelli, and T. Charnois, "Reasoning on aspectual-temporal information in French within Conceptual Graphs," presented at Proceedings. 14th IEEE International Conference on Tools with Artificial Intelligence, 2002. (ICTAI 2002). Washington DC, USA, 2002.

[7] S. Chu and B. Cesnik, "Knowledge representation and retrieval using conceptual graphs and free text document

978-1-4244-2328-6/08/$25.00 © 2008 IEEE

self-organisation technique," *International Journal of Medical Informatics*, vol. 62, pp. 121-133, 2001.

[8] F. e. Fürst and F. Trichet, "AxiomBased Ontology Matching," presented at 3rd International Conference on Knowledge Capture KCAP'05, Banff, Alberta, Canada., 2005.

[9] R. Hill, S. Polovina, and M. Beer, "From Concepts to Agents: Towards a Framework for Multi-Agent System Modelling," presented at Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems (AAMAS'05), The Netherlands, 2005.

[10] C. M. Jonker, R. Kremer, P. V. Leeuwen, D. Pan, and J. Treur, "Mapping visual to textual knowledge representation," *Knowledge-Based Systems*, vol. 18, 2005.

[11] D. Sleator and D. Temperley, "Parsing English with a link grammar," presented at 3rd Int. Workshop of Parsing Technologies, Tilburg, The Netherlands. 1993.

[12] L. Zhang and Y. Yu, "Learning to Generate CGs from Domain Specific Sentences. ," presented at In Proceedings of the 9th International Conference on Conceptual Structures (ICCS 2001), LNCS 2120, Stanford, CA, USA, 2001.

[13] F. M. Suchanek, G. Ifrim, and G. Weikum., "Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents. ," presented at SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, 2006.

[14] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, "Approaches to text mining for clinical medical records," presented at Proceedings of the 2006 ACM symposium on Applied computing, Dijon, France, 2006.

[15] Montes-y-Gómez, A. Gelbukh, and A. López-López, "Mining the news: trends, associations, and deviations," *Computación y Sistemas*, vol. 5, 2001.

[16] J. Zhong, H. Zhu, J. Li, and Y. Yu, "Conceptual Graph Matching for Semantic Search," presented at Proceedings of International Conference on Conceptual Structures, Borovets, Bulgaria, 2002.

[17] J. A. Gonzalez, L. B. Holder, and D. J. Cook, "Graph based Concept Learning," presented at Proceeding of the Fourteenth Annual Florida AI Research Symposium, Florida, USA, 2001.

[18] M. Montes-y-Gómez, A. Gelbukh, and A. López-López, "Detecting Deviations in Text Collections : An Approach using Conceptual Graphs," presented at Proc. MICAI-2002:Mexican International Conference on Artificial Intelligence, Mexico, 2002.

978-1-4244-2328-6/08/$25.00 © 2008 IEEE