Outlier detection in financial statements: a text mining method

S. S. Kamaruddin[1], A. R. Hamdan[2], A. Abu Bakar[2] & F. Mat Nor[3]
[1]*College of Arts and Science, Universiti Utara Malaysia, Malaysia*
[2]*Faculty of Information Technology & Science, Universiti Kebangsaan Malaysia, Malaysia*
[3]*Graduate School of Business, Universiti Kebangsaan Malaysia, Malaysia*

Abstract

This paper presents a text mining methodology to extract outlying knowledge from a collection of financial statements. The main idea is to extract relevant financial performance indicators and discover implicit textual description of the indicators. The extracted information was represented using a network language i.e. conceptual graph. Outlier mining was performed on the conceptual graph representation using a deviation based method. Experiments were carried out to evaluate the effectiveness of the proposed method. Results show that the proposed method is able to excerpt outlying knowledge from the financial statements with accuracy comparable to human experts.

Keywords: text mining, information extraction, conceptual graphs, outlier mining in text, deviation based outlier mining method.

1 Introduction

In recent years, there has been a continuous interest among the data mining community towards outlier detection due to its potential in discovering rare and interesting patterns from datasets. The importance of outlier detection is highlighted in diverse applications ranging from fraud detection and threat warning systems to systems that are dedicated to discover new knowledge from unusual patterns in data. With the growing amount of textual documents generated today, enormous attention should be focused on retrieving outlying patterns from text. Outlier detection in text aims at discovering implicit

WIT Transactions on Information and Communication Technologies, Vol 42, © 2009 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA090081

knowledge that distinctively deviates from the general information contained in textual data. Finding extreme patterns in text is a non-trivial problem due to its high dimensionality.

A broad spectrum of techniques from various disciplines such as statistics, machine learning, data mining, information theory and spectral decomposition are normally employed to solve the problem [1]. However, most researchers tend to approach the problem by extracting attributes from text prior to applying standard statistical or data mining techniques to detect outliers from the data. It is not feasible to implement some of these approaches on a very large document collection. As a contrast, we propose a three stage text mining method that is capable of effectively achieving the goal of identifying text outliers. In the first stage, the target sentences are extracted with an information extraction system. Each extracted sentence is parsed and represented as conceptual graphs (CGs) in the second stage. In the final stage a deviation detector matches all CGs to assign dissimilarity scores and ranks the scores to identify the top outlying sentences.

Our proposed method is based on the deviation based outlier mining method, but instead of using the sequential exception technique introduced in [2], we present a novel standard based deviation detection technique. In this technique a standard CG with embedded synonyms is introduced and a set based dissimilarity function is derived to calculate the difference between the compared CG with the standard CG. Evaluation of the proposed method shows similar performance with expert ranking given the same sentences. The rest of the paper is organized as follows. Section 2 presents previous work in the area. In section 3, we explain the proposed method implemented in this work. Section 4 is devoted to empirical evaluation and results. We end the paper with conclusions in Section 5.

2 Related work

Outliers in text are typically known as novelty detection, anomaly detection and deviation detection. In this section we discuss some related works in this area by focusing firstly on the extraction of relevant information from text, secondly on the representation of text and lastly on the outlier detection method. In text mining, extraction of relevant information and representation of the text are two crucial issues. Relevant information extraction will reduce the search space and filter the text from vast amounts of irrelevant information. Accurate representation of text will determine the accuracy of the mined information.

Earlier work in information extraction depended on a hand-crafted template as reported in [3, 4]. Recent study on information extraction [5, 6] concentrated on relation extraction where general natural language processing (NLP) tasks, such as tagging and parsing, were used before domain specific relation extraction was performed on the parsed tree. In our work, the datasets are a collection of financial statements. Financial statements are produced with fonts, colours and layout to make them readable, but it is difficult to infer their complex internal structure. For this reason, we have adopted a rule-based multi-pass strategy as

proposed in [7]. Using this method we employed a specific rule-based method to recognize the noun phrases and performance indicators and then to efficiently extract the related sentences from the financial statements.

Extracted information needs to be represented in a more structured manner to facilitate mining tasks. There are various methods for text representation, among which are vector based representation and N-grams as reported in [8-10]. However, these methods represent words in isolation without considering the context in which the words occurs. Researchers in the field are beginning to give importance to richer representation schemes, such as network language, that capture the relation between words in order to yield results that are more promising. Following these developments, a number of network languages were employed to model the semantics of natural language and other domains. Some examples can be found in [11-14].

In our research we utilize the conceptual graph, a particular network language proposed by Sowa and Way [15] and show how this formalism may be used to represent knowledge and mine outliers. Among the advantages of using conceptual graph formalism are, firstly, its expressions are similar to natural language. Secondly, they are adequate to represent accurate and highly structured information beyond the keyword approach and thirdly, both semantic and episodic association between words can be represented using CGs [15]. Various works have used CG representation to capture the structure and semantic information contained in free text. Among them are works reported in [16-18]. All these works employed conceptual graph representations of respective texts to perform assorted tasks such as information retrieval, concept learning and creating text databases.

The outlier detection methods typically discern among the statistical approach, the distance-based approach, the classification-based approach or the clustering-based approach. Statistical approaches as explored in [19] require prior knowledge of data distribution hence they are considered as unsuitable for the high dimensionality of text data. We review some other methods that were particularly used in discovering text outliers. Distance-based approaches were explored in [20] where N-grams terms frequency distribution were created and the dissimilarity between two document vectors were computed by measuring the angle between two vectors using the cosine function. The distance-based method was also used in [10] to find outliers in text. These methods are acceptable; however, distance becomes less meaningful with the increase in the dimensionality of data sets.

Classification-based methods, such as Neural Network, Naïve Bayes and Support Vector Machine are explored in [21]. Although these methods offer promising results, they are only applicable if we can clearly distinguish the differences between outlying classes and normal classes for the training data set. To overcome this limitation, many researchers divert their attention to clustering-based methods, such as the Expectation Maximization algorithm, as explored in [9, 22]. According to this method, outliers are data items that do not belong to any clusters. These approaches are apparently slow since it is well-known that the computational complexity of most clustering algorithms is exponential as the

size of data increases. Therefore, it is not suitable for high dimensionality of text data. Furthermore, most frequently outliers are the by product of clustering, therefore clustering algorithms are not optimized to finding outliers compared to other methods. In addition, most cluster-based algorithms rely on some distance computation between data items. The work reported in [23] in particular shows the clustering of CGs to detecting outliers. However, the approach involves organization of CGs into hierarchies and demands complex clustering algorithms. On the other hand, a deviation based outlier mining method offers linear complexity as reported in [2, 24] and is desirable if the differences between the normal and abnormal data are not so evident as in the text data.

3 The proposed text outlier mining method

The proposed text outlier mining method involves three stages. We have adopted a hybrid approach that combines the rule-based method with the computational linguistics-based method that is specifically deep parsing. The rule-based method defines a set of rules for possible textual relationships while deep parsing deals with the structure of entire sentences, hence it is more accurate [12]. As shown in fig. 1, the financial statements are first processed by an information extraction system to extract the relevant performance indicators. Secondly, the extracted sentences are parsed to obtain the syntactic structure of the sentences to facilitate the process of transforming them into a conceptual graph representation. In the final stage the generated CGs are matched to detect the outlier. We detail these stages in the following sections.

3.1 Information extraction

The aim of information extraction is to excerpt relevant information and filter out the irrelevant information from the lengthy financial statements. The challenge in this task is to ensure that the extractor is tailored to the special needs of extracting the financial performance indicators together with their respective textual descriptions. The developed extractor scans the financial statements to identify and extract key phrases and relevant sentences. We have performed multi pass scans on the financial statement using an integrated development environment named *VisualText* with the help of the NLP++ programming language. In this component, the raw text was first tokenized into units of alphabetic, numeric, punctuation, and white space characters. Then, a joining operation is performed on the resulting tokens. This operation is needed because it is necessary to join some tokens in order to consider them as one group, for example numbers, percentages and dates.

Next, the documents were zoned into paragraphs, headers, sentences, and table zones. Zoning facilitates the searching process where the search space can be reduced by directly focusing on certain headers. This further improves the process of finding the required information. We perform noun phrase recognition on the zoned documents to identify important phrases. A list of noun phrases is given as input to perform this process. We also provide a list of financial

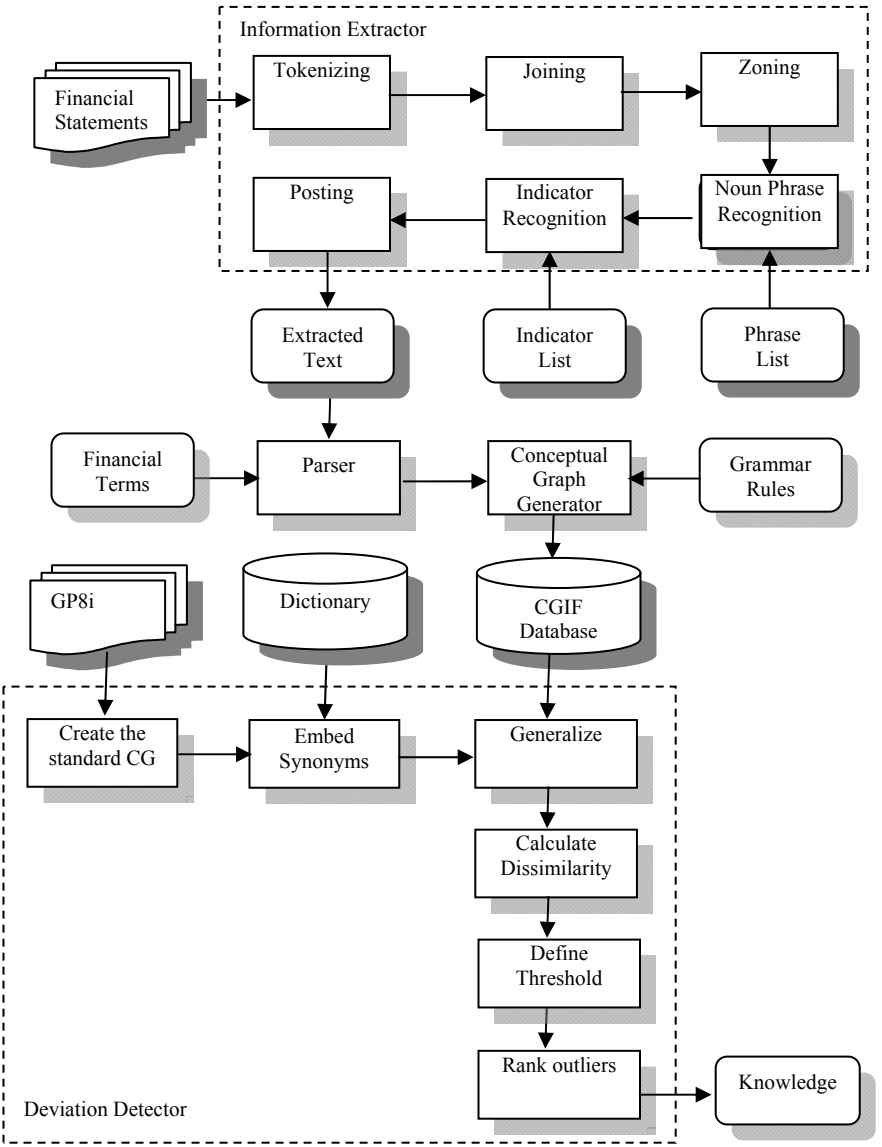

Figure 1: Text outlier mining method.

indicators to be searched and extracted. The results are extracted text that contains relevant performance indicators and phrases for further processing. The extractor was evaluated to measure its performance. The performance is reported in section 4 of this paper.

3.2 Parsing and conceptual graph generation

The parsing was implemented using the Link Grammar Parser (LGP) [25], 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. The benefit of using LGP is that there exists a structure similarity to CGs, hence it is easier to map the obtained structure to CGs [26]. Suchanek et al. [27] reported that the LGP provides a much deeper semantic structure than the standard context-free parsers. The parser is able to identify the syntactic level of sentence decomposition and categorizes the phrase into the following: S, which represents sentences; NP, which represents Noun Phrases; VP, which represents Verb Phrases; and PP, which represents Preposition Phrases.

This structure was traversed from its roots to generate the conceptual graphs. One prominent advantage of representing information with conceptual graphs is the ease of interpretability. In addition, the conceptual graph inherits the mathematical foundation of graph theory. This gives an extra advantage for further processing. The general grammar rules were used by the CG generator in the process of traversing the constructed sentence structure. Using this method the generator successfully identified nouns, verbs and adjectives, which were later built into concepts, whereas the prepositions were transformed into relations. The process is explained in detail in our previous paper [28]. The results of the generator were reformatted into a list of concepts and relation predicates following the formalism implemented in Conceptual Graph Interchange Format (CGIF) as shown in fig. 2. This representation induces structure to the documents and makes it easier to perform mining tasks on them. The CGs were labelled with G_{iyx} where i represents performance indicators, y represents financial years and x represents the respective number of sentences extracted. The G_{iyx} were stored in the CGIF database.

G_{321} = (c1.the_year.).(c2.net_profit.*).(c3.dividing.*).*
(c4.calculated.).(c10.the_financial_year.*).(c5.issue.*).(c6.ordinary_s*
hares.).(c7.the_number.*).(c8.the_Bank.*).(c9.Basic_Earning_per_sh*
are.),(by,c3,c7).(for,c2,c1).(obj,c3,c2).(by,c4,c3).(during,c5,c10).(in,c*
6,c5).(of,c7,c6).(is,c8,c4).(of,c9,c8).}

Figure 2: The CGIF for CG representing performance indicator 3, financial year 2002 and sentence 1.

3.3 Deviation-based outlier mining method

A deviation-based outlier mining method involves the analysis of the main characteristics of an object in a group. Then it compares the characteristics with the rest of the objects in the group. Any object that "deviates" from these characteristics is regarded as an outlier. The first step in this method is to create the standard CG for all extracted performance indicators. For this purpose, we have utilized a predefined standard produced by the Malaysian Government

Authority, Bank Negara Malaysia (BNM). The standard, named GP8i, is a guideline for the specimen reports and financial statements for licensed Islamic banks. The GP8i was analyzed and standards related to the performance indicators were extracted and parsed. Then, the CGs were created from these standards and were given the label SG_i, where i is the number of performance indicators.

In the next step the generated SG_i were embedded with synonyms. We refer to a predefined dictionary extracted from *Wordnet* to accomplish this purpose. We than obtained all the CGs to be compared from the CGIF database. We perform a generalization function on the CG. This is done by matching the concepts from the G_{iyx} with the concepts and synonyms of the SG_i. The matched concepts were renamed accordingly and their identifiers were updated. The next step in our method is to perform a matching process of the G_{iyx} and SG_i with a dissimilarity function. The degree of dissimilarity of the compared conceptual graph, G_{iyx}, to a given standard conceptual graph, SG_i, is calculated with the following dissimilarity function:

$$D_{(Giyx\,,\,SGi)} = \frac{n(\mathrm{G_{iyx}} \cup \mathrm{SG_i}) - n(\mathrm{G_{iyx}} \cap \mathrm{SG_i})}{\mathrm{n(G_{iyx}} \cup \mathrm{SG_i})} \tag{1}$$

It is based on the Jaccard distance dissimilarity measure. It indicates that the dissimilarity between any two CGs is the ratio of the size of their union minus the size of their intersection to the size of their union. We based our dissimilarity function on the Jaccard distance because the conceptual graphs are in set format and we do not have to change the sets into vectors to use cosine distance or change it into points to use Euclidean distance; instead, we represent sets as sets and employ the Jaccard distance measure. Using this dissimilarity function, the identical CGs have a dissimilarity of 0, completely dissimilar CGs have a score of 1 in such a way that for G_{iyx}, the SG_i of any performance indicator i is:

$$D_{(Giyx\,,\,SGi)} = \begin{cases} 1 & if\,(\mathrm{G_{iyx}} \cap \mathrm{SG_i}) = \varphi \\ 0 & if\,(\mathrm{G_{iyx}} \cup \mathrm{SG_i}) = (\mathrm{G_{iyx}} \cap \mathrm{SG_i}) \\ 0 > D < 1 & otherwise \end{cases} \tag{2}$$

Once the dissimilarity scores are calculated, the process is followed by a threshold definition and ranking. The result of the whole process is the top n outlying sentences from the collection of financial statements.

4 Empirical evaluation and results

A number of experiments were set to evaluate the proposed text outlier mining method on a corpus which contains a collection of real-world financial statements of a local Islamic bank for a period of 9 years (2000 – 2008). The corpus contains approximately 163,000 words. The important performance indicators that were considered relevant were Total Assets, Share Capital and

Net profit/loss. Our method extracted 30 sentences describing these performance indicators and these sentences were parsed and transformed into conceptual graphs. We evaluate the performance of the extractor using the following precision and recall measures.

$$\text{EPrecision} = \frac{\text{Re}trieved \;\&\; \text{Re}leveant}{\text{Re}trieved \;\&\; \text{Re}levant + \text{Re}trived \;\&\; Irrelevant} \tag{3}$$

$$\text{ERecall} = \frac{\text{Re}trieved \;\&\; \text{Re}levant}{\text{Re}trieved \;\&\; \text{Re}levant + \text{Re}levant \;\&\; Notretrieved} \tag{4}$$

$$\text{F-measures} = \frac{2 \times E\,\text{Pr}ecision \times E\,\text{Re}call}{E\,\text{Pr}ecision + E\,\text{Re}call} \tag{5}$$

The Information Extractor revealed high precision scores of 94% and recall scores of 87%. The F-measures score was 90% which is considered high for an information extraction system. For the next experiment, we implemented the dissimilarity measures as discussed in section 3 to calculate the dissimilarity scores between standard CG and the extracted sentences CGs. As a control for the experiment, we provide the same sentences to an Assistant Audit Manager of an Islamic bank to rank the sentences with scores between 1 to 10, where 1 indicating the most dissimilar sentence and 10 representing the most similar sentences. We then ranked our dissimilarity scores the same way. Fig. 3 presents the rankings for all CGs. The graph shows that almost all CGs were ranked similar to the expert judgment.

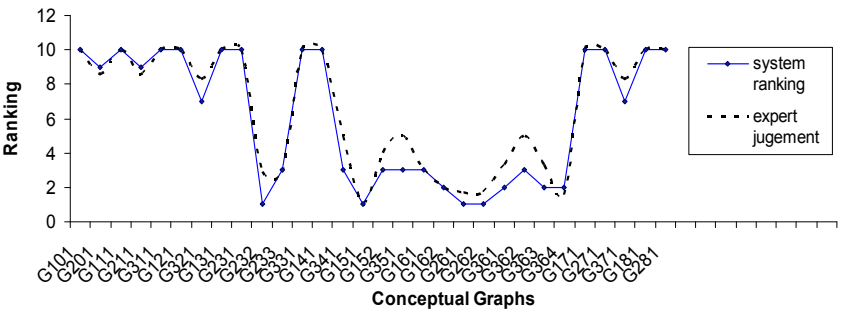

Figure 3: Ranking of conceptual graph similarity compared to expert judgement.

We calculated the ranking precision for each performance indicator with equation (6) and the Average Ranking Precision (ARP) with equation (7).

$$\text{RP}_i = \frac{\text{Correctly ranked CG}\,(\text{G}_i)}{\text{Total CG}\,(\text{T}_i)}\text{, where } i = \text{performance indicator} \tag{6}$$

$$\text{ARP} = \frac{\text{Correctly ranked CG (G)}}{\text{Total CG (T)}}\ ,\ \forall \text{performance indicator} \qquad (7)$$

Table 1 presents the ranking precisions. The highest precision of 90% was recorded for the Net profit/loss performance indicator. However, one important point to remember is the number of conceptual graphs generated within each performance indicator is different; therefore lower precision values may be recorded for lower number of conceptual graphs. Nevertheless the proposed method has produced an average ranking precision of 82% which is considered acceptable.

To further evaluate our results we have calculated the values of Return on Assets (ROA) and Return on Equity (ROE) using the numerical values of total assets, share capital and net profit/loss where ROA = Total assets/Net profit(loss) and ROE = Share capital/net profit(loss). We than compared the resulting values with an averaged sentence ranking for each financial year and present the results in a 2 y-axis line graph as shown in fig. 4.

The figure shows the trend of the ROA and ROE values. Both ROA and ROE values show a drop in the bank's performance for 2005 and 2006. Similarly our proposed method ranked the sentences extracted in these years as outliers. The results show that detecting outlying sentence can give insight knowledge on why the banks performance is low in the mentioned years. Table 2 presents one example of outlying sentence and explanation on why it is considered as outliers.

Table 1: Ranking precisions.

Measurement	Scores
$RP_{\text{Total Asset}}$	75%
$RP_{\text{Share Capital}}$	77%
$RP_{\text{Net profit/Loss}}$	90%
ARP	82%

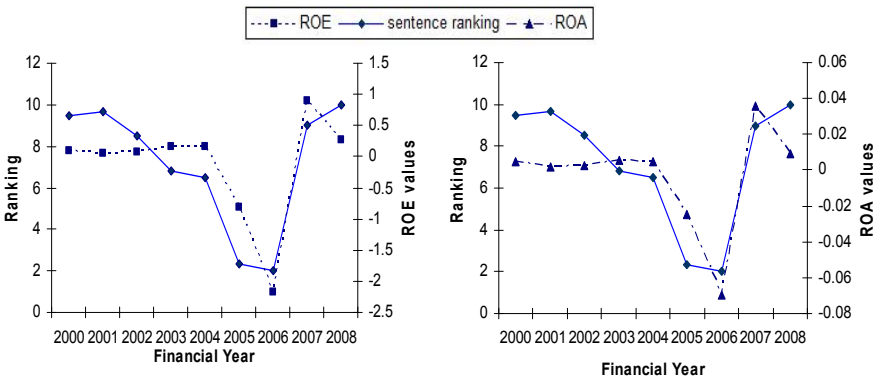

Figure 4: Sentence ranking, ROA and ROE values for various financial years.

Table 2: Example of outlier.

Outlying Sentence	Explanation
For the FYE2006, the Bank reported a higher total income of RM960.63 million compared to FYE2005 but a one-off provision of RM1.48 billion for non-performing financing (NPF) resulted in a loss before tax and zakat of RM1.28 billion, while net loss amounted to RM1.30 billion.	This sentence is considered outliers because in this year the bank recorded the greatest loss of 1.3 billion due to non performing financing, which is considered an abnormal event.

5 Conclusion

In this paper we have presented a text outlier mining method using conceptual graphs to detect deviations from financial statements. We have proposed a deviation based outlier mining method and illustrate with experiments how this method can be employed to yield a promising result. Our method successfully extracted relevant performance indicators with a high accuracy. The outlier mining produces similar result to expert judgment and when compared against financial ratios such as ROA and ROE the method successfully highlights the outlying sentences and the performance trends.

One prominent benefit of our proposed method is in the introduction of a standard CG in the matching process to reduce the complexity of the existing clustering based method from exponential to linear complexity. Other benefits include the process of embedding synonyms into the standard CG. This process discards the rigidness of matching exact words therefore it is suitable for matching sentence with different terms but represents similar meaning. Besides that, the proposed method only requires a function that can rank the degree of dissimilarity between sentences. This is considered extremely suitable for large text database.

References

[1] Chandola, V., A. Banerjee, and V. Kumar, *Outlier Detection - A survey*. 2007, Department of Computer Science and Engineering, University of Minnesota: Minneapolis, USA. pp. 1-53.

[2] Arning, A., R. Agrawal, and P. Raghavan. A Linear Method for Deviation Detection in Large Databases. *in Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pp. 164-169, 1996.

[3] Andersen, P.M., et al. Automatic extraction of facts from press releases to generate news stories *in Proceedings of the third conference on Applied natural language processing*, Trento, Italy., pp. 170-177, 1992.

[4] Kornfeld, W. and J. Wattecamps. Automatically Locating, Extracting and Analyzing Tabular Data. *in Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne Australia, pp. 347-348, 1998.

[5] Mangassarian, H. and H. Artail, A general framework for subjective information extraction from unstructured English text. *Data & Knowledge Engineering*. **62**: pp. 352-367, 2007.
[6] Culotta, A., A. McCallum, and J. Betz, *Integrating probabilistic extraction models and data mining to discover relations and patterns in text*, in *Human Language Technology Conference of the North American Chapter of The Association of Computational Linguistics (HTL/NAACL)*. 2006.
[7] Meyers, A., *Multi-Pass Multi-Strategy NLP*, Text Analysis International, Inc.
[8] Agyemang, M., K. Barker, and R.S. Alhajj. Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams. *in 2005 ACM Symposium on Applied Computing*, ACM: Santa Fe, New Mexico, USA, pp. 482-487, 2005.
[9] Miller D.J., B.J., A Mixture Model and EM-Based Algorithm for Class Discovery, Robust Classification, and Outlier Rejection in Mixed Labelled/Unlabeled Data Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **25 (11)**: pp. 1468 - 1482, 2003.
[10] Miller R.C., M.B.A. Outlier Finding: Focusing User Attention on Possible Errors. *in Proceedings of the 14th annual ACM symposium on User interface software and technology* Orlando, Florida 2001.
[11] Cimiano, P., A. Hotho, and S. Staab, Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *Journal of Artificial Intelligence Research*. **24**: pp. 305-339, 2005.
[12] Fürst, F.e.e. and F. Trichet. AxiomBased Ontology Matching. *in KCAP'05*, Banff, Alberta Canada, 2005.
[13] Rajarama, K. and Ah-Hwee Tan. Mining Semantic Networks for Knowledge Discovery. *in Third IEEE International Conference on Data Mining (ICDM'03)*, pp. 633, 2003.
[14] Bales, M.E. and S.B. Johnson, Graph theoretic modelling of large-scale semantic networks. *Journal of Biomedical Informatics*. **39(4)**: pp. 451-64, 2006.
[15] Sowa, J.F. and E.C. Way, Implementing a semantic interpreter using conceptual graphs. *IBM J. Res. Develop*. **30(1)**: pp. 57-69, 1986.
[16] Gonzalez, J.A., L.B. Holder, and D.J. Cook. Graph based Concept Learning. *in Proceeding of the Fourteenth Annual Florida AI Research Symposium*, pp. pp. 377-381, 2001.
[17] Hensman, S. and J. Dunnion. Automatically Building Conceptual Graphs using VerbNet and WordNet, *in Proceedings of the 2004 international symposium on Information and communication technologies ISICT '04* Trinity College Dublin, 2004.
[18] Karalopoulos, A., M. Kokla, and M. Kavouras. Geographic Knowledge Representation Using Conceptual Graphs. *in 7th AGILE Conference of Geographic Information Science*, Heraklion, Greece, 2004.
[19] Montes-y-Gómez, A. Gelbukh, and A. López-López, Mining the news: trends, associations, and deviations. *Computación y Sistemas*. **5(1)**, 2001.

[20] Agyemang, M., K. Barker, and R.S. Alhajj. WCOND-Mine: Algorithm for Detecting Web Content Outliers from Web Documents. *in Proceedings of the 10th IEEE Symposium on Computers and Communications (ISCC 2005)*, EEE Computer Society: Murcia, Cartagena, Spain, 2005.

[21] Manevitz, L.M. and M. Yousef, One-Class SVMs for Document Classification. *Journal of Machine Learning Research 2*: pp. 139-154, 2001.

[22] Cooley, R., B. Mobasher, and J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web. *in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI97)*, Newport Beach, CA, USA, 1997.

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

[24] Xie, C., Z. Chen, and X. Yu, *Sequence Outlier Detection Based on Chaos Theory and Its Application on Stock Market* Lecture Notes in Computer Science: Fuzzy Systems and Knowledge Discovery. Vol. 4223/2006. 2006: Springer Berlin / Heidelberg. 1221-1228.

[25] Sleator, D. and D. Temperley. Parsing English with a link grammar. *in 3rd Int. Workshop of Parsing Technologies*, 1993.

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

[27] Suchanek, F.M., G. Ifrim, and G. Weikum. Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents. . *in SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[28] Kamaruddin, S.S., A.R. Hamdan, A.A. Bakar, and F.M. Nor. Conceptual Graph Formalism for Financial Text Representation. *in Proceedings of International Symposium of Information Technology (ITSIM08)*, Kuala Lumpur, Malaysia, pp. 1-6, 2008.