Visualizing Association Rules of Analyzed XML Semantic Structured Text

Z. T. Fayed[*1], M. M. Abdallah[**2]

[*] *Computer Science Department,*
Faculty of Computer and Information Sciences,
Ain Shams University,
Cairo, Egypt.
[1] ZTFayed@cis.asu.edu.eg
[**]Computer Science Department, Faculty of Science,
Minia University,
El-Minia, Egypt.
[2] Mas_m4@hotmail.com

Abstract-***No doubt that the whopping size of the available information on the internet gets the user quite confused when he deals with that bulk of information. The new mining techniques are built to solve that problem and to enable the user to get and understand the information patterns; one of these techniques is information visualization. In this paper a new information visualization model is built upon the results of the association rule technique which is implemented on the XML text-based documents. End results show that in the future we can perceive enormous size of information and get its hidden knowledge in small time.***

Key words: ***XML, Information Visualization, Classification, Association Rules***

1 INTRODUCTION

Recently, the emergence of the semantic web technologies (e.g. RDF, OWL…) and text mining techniques galvanize the researchers and the organizations to publish their information in as so-called Linked Open Data in a semantic structured form to conform to these new techniques. Nevertheless, the novice user could hardly deal with those techniques easily. The visual data mining or information visualization refines the semantic information and visualize it into simple logic forms

The information visualization which is the exploit of interactive representation of the information and knowledge (e.g. entity relations and classification rules) tend to amplify the user cognition of that information. Due to the computer breakthrough, now it is possible to transform automatically the written information into diagrams and visual forms which make the cognition process simpler and practical. Moreover the visualization process can be made dynamic, interactive, and can be integrated into large process of sense making and creation.

In spite of the fact that the text mining branch is not relatively new, albeit so far there is no dependable systematic means to represent the mining and analysis results to the novice users. In this paper we propose a new model for mapping between the text mining techniques (e.g. classification technique) and the information visualization. In other words, that model epitomizes the extracted information and knowledge from the text using the visualization technique to facilitate and amplify the cognition of this knowledge. Besides we argue the possibility of creating a visualized index which represents the term-topic relations which in addition gives good insights into the high frequent terms.

2 RELATED WORK

An accretion number of research and commercial systems are using visualization techniques to assist in supporting investigative analysis. Analyst's Notebook [1]invents a new semantic visualization model to help analysts with investigations. Nodes in the graph are entities of semantic data types such as person, event, etc. Despite the system enable us to import text files and do automatic layout, its cardinal application is helping the analysts in manually crafting case charts.

Oculus Info Inc. equips a suite of systems with diverse aspects of investigative analysis. TRIST [2]enables the analysts to formulate, refine, organize and execute queries over appreciable document collections. User interface of TRIST is a multi-pane scene that provides alternative perspectives to search results encompass clustering, trend analysis, comparisons, and other alternatives. Information retrieved using TRIST can be consumed into the SANDBOX system [3], analytical sense-making environmentwhich helps to sort, organize, and analyze large amounts of data. The system's main point is to amplify human's insights with computational linguistic, analytical functions, and by motivating the analyst to make thinking more incisive. The system maintains interactive visualization techniques embrace gestures for posing, moving, and grouping information, additionally templates for building visual models of information and visual evaluation of evidence. An assessment experiment of the SANDBOX system disclosed that analysts using the system did higher quality analysis in less time than using standard tools.

Within the scope of knowledge-interest visualization, many authors have contrived concepts which motivate the visualization process by making use of formalized knowledge. Wang et al. [4] give bright study of how knowledge “advances" through the visualization process by unpack many alternative conversion processes. Nevertheless, information on how to assign these processes in generic InfoVis[5] systems to help users in visualizing semantic models, is missing. Chen et al.[6]figures a high-level knowledge-based infrastructure by analogy with the visualization system, which extracts information from data and exploits it together with predefined expert knowledge to configure the visualization process.

3 Association Rule Prelude

From the text mining techniques our model visualize the association rules results. Due to its ability exhibit the relation between the frequent terms occurs within each topic. The definition of an association rule varies with disciplines and implementations.

Association rule mining finds interesting associations and/or correlation relationships among appreciable set of data items. Equally the association rules show attribute’s value conditions that occur frequently together in a given dataset [7]. The preambles necessary to understand for performing data mining on any data are discussed below:

Let$\{T_1, T_2 \ldots T_m\}$be a set of terms. Let S be a set of sentences where each sentence P is a set of terms such that P$\subseteq$ T. Each sentence is associated with an identifier (SID). Let A, B be two sets of terms. A sentence P is said to contain A if and only if A $\subseteq$ P. an association rule is an implication of the formA$\Longrightarrow$B,

where,$A \subset T$, $B \subset T$, and $A \cap B = \phi$ (1)

Support (s) and confidence (c) are two measures of rule interestingness. They respectively reflect the usefulness and certainty of the discovered rule. A support of 2% of the rule A$\Longrightarrow$B means that A and B exist together in 2% of all the sentences under analysis. The rule A$\Longrightarrow$B having confidence of 60% in the sentence set S means that 60% is the percentage of sentences in S containing A that also contains B.

A set of terms is referred to as a termset. A termset that contains k terms is a k-termset. The occurrence frequency of a termset is the number of sentences that contain the termset. If the relative support of a termset t satisfies a prescribed minimum support threshold, then t is a frequent termset. The association rule mining can be viewed as a two-step process:

1- *Find* all frequent termsets: Each of these termsets will occur at least as frequently as a predetermined minimum support count.

2- *Generate* strong association rules from the frequent termsets: The rules must satisfy minimum support and confidence. These rules are called strong rules [7].

4 Visualizing association rules

The proposed model is implemented to support the ongoing text mining and visualization research [8][9] [10] on large document corpora. The cornerstone is to study the relationships and correlation among topics that are used to portray a corpus. The goal is to detect substantial association rules within a corpus such that the turnout of a set of topics in an article alludes to the presence of another topic. For instance, one might comprehend in headline news that whenever the words “Labor market” and “inflation” occur, it is highly probably that the unemployment is also mentioned. We elucidate the results using news Reuter’s corpus with more than 12,901 articles. Numerous experiments are conducted on Reuters Corpus Volume 1 (RCV1) to evaluate the efficiency of the proposed model. Figure 1 illustrates the structure of the proposed visualizing model and clearly demonstrates the main components of that model.

One of the best information visualization applications is the ability to interact with the chart components. The arrow in figure 1 which is titled with “Using visualization interaction” implies that the user can communicate and reach the documents by interacting with one of the chart components. Chart component could be a term or a rule was extracted using the association rule analysis.

The following procedure demonstrates the functions and sequence of each component in Figure 1.

1- Process and consume the document corpora to extract the single sentences, the main relations among these sentences, and hence their associated categories;

2- Perform NLP processing on the extracted sentences from step 1;

3- Carrying out the association rules mining technique onto the outcome of step 2 to investigate the correlation relationships between the sentences and the associated topics;

4- Visualize the results; thetree graph isa popular technique to depict items associations. The nodes of the tree represent the terms, and the edges represent the associations between these terms. Figure 2 shows a tree instance for the association rules. An association graph can quickly turn into an interlaceddisplay with as few as a dozen rules.

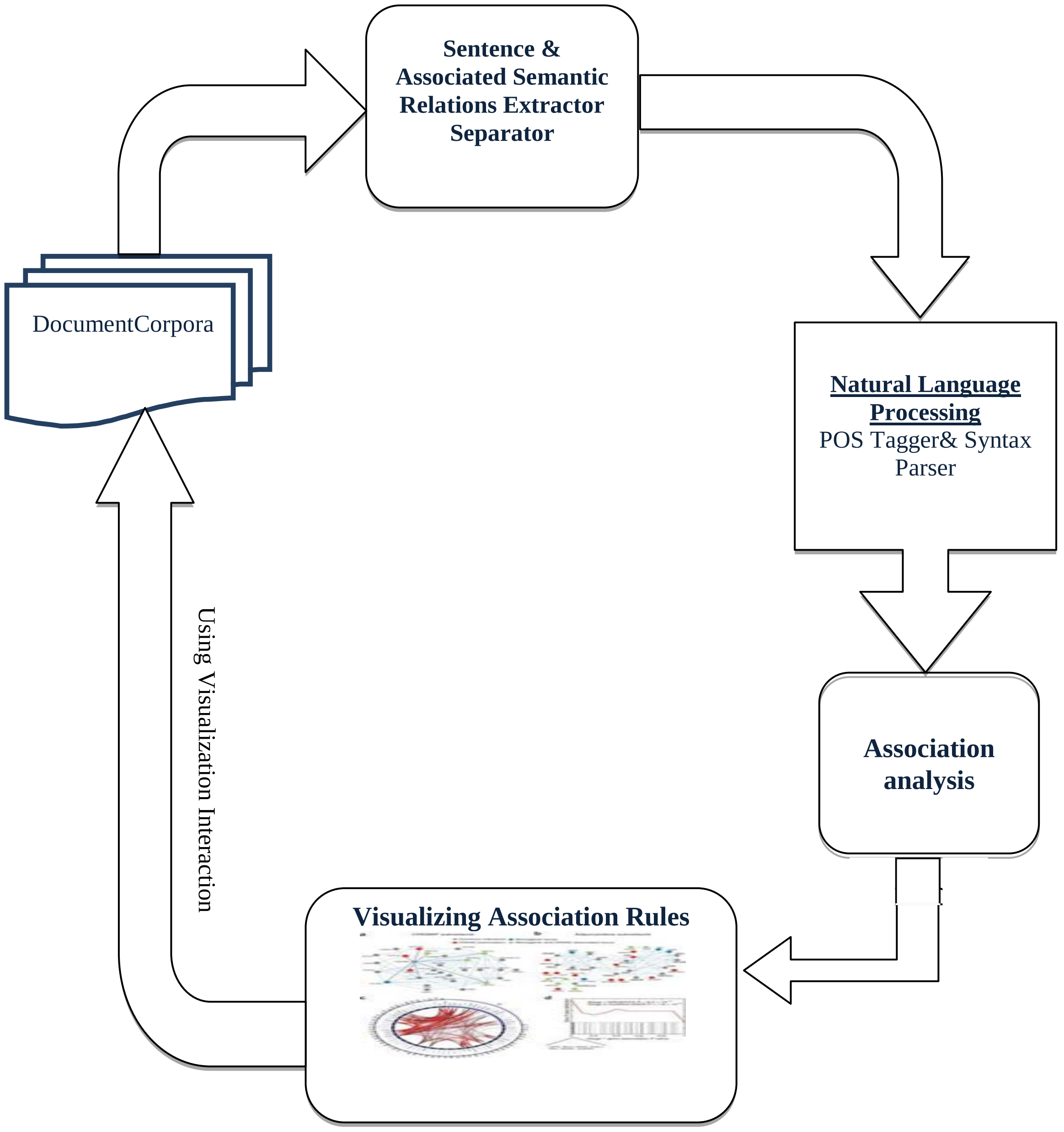
Sentence & Associated Semantic Relations Extractor Separator
Natural Language Processing
POS Tagger& Syntax Parser
Association analysis
Visualizing Association Rules
Using Visualization Interaction
DocumentCorpora

Figure 1: the proposed visualizing model.

The following algorithm in figure 3 adumbrates the main procedure used to extract the association rules and visualize it for amplifying user cognition.

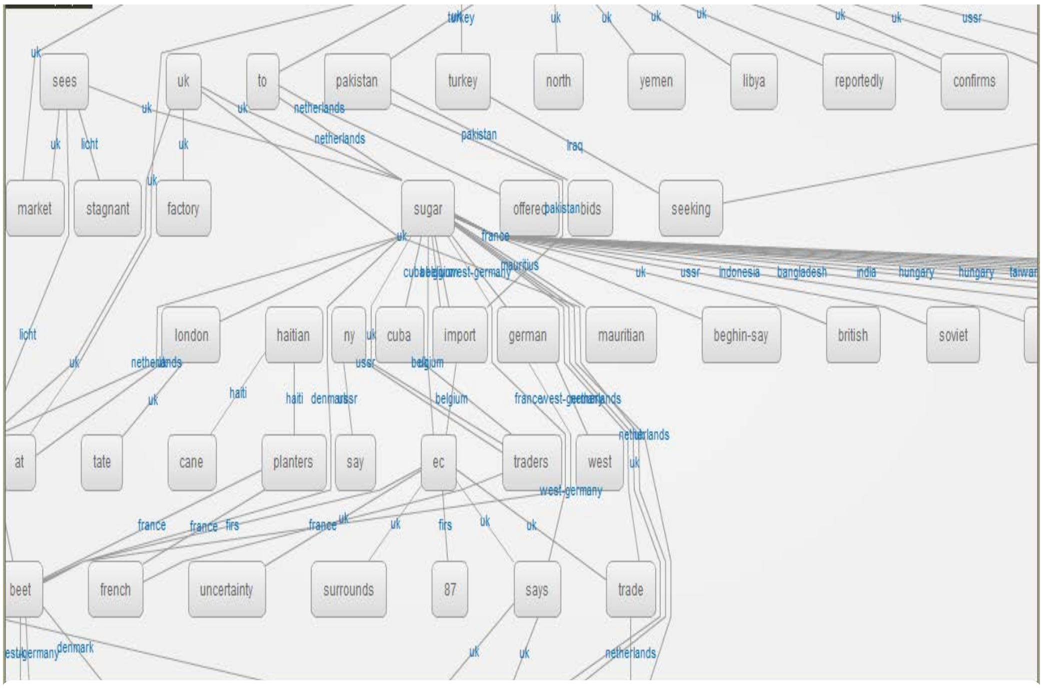

Figure 2: A sub tree represents association rule results between text terms

```
1. Read the document corpora (D)
2. For each document (d) do
      2.1 Extract the d sentences (s)
      2.2 Set the topic (T) for each s
      2.3 For each s do
            2.3.1 Identify the parts of speech (pos) of s
            2.3.2 Do semantic processing for each s
      2.4 Analyze the results of 2.3.1 and 2.3.2 steps using the
          association rule mining technique
3. Visualize the results of step 2.4 in a graph (g)
4. Reach the specific document corpora using g components
```

Figure 3: steps of extracting the association rules and visualize it

To boost navigation, zooming and panning is activated using mouse wheel interactions. Besides, the Graph View also equips a ‘Circular Layout’ command that repositions all the visible nodes close around a large circle in the view. Nodes connecting to more than one node are drawn inside the circle. Thus, the set of nodes easily noticeable inside the circle represents a more highly connected network of terms that may be related in considerable ways and liable should be examined more closely.

5 EXPERIMENTAL RESULTS

A. dataset description

The experimental setup conducted on a data set has 12,902 XML documents from the Reuters 21578 dataset [11]. There are 9031 documents in the training set, and 3870 documents in the test set. Out of the 5 category sets, the topic category set contains 135 categories, but only 90 categories have at least one document in the training set. These 90 categories were used in the experiment.

In the dataset, the text directly is analyzed, rather than, using metadata associated with the text documents. This clearly demonstrates the effect of using concepts on the text categorization process. The stop words are removed from the concepts that are extracted by the proposed model. The extracted concepts are stemmed using the Porter stemmer algorithm [12].

B. Perceptual statistical evaluation for information visualization

Infrequently it may be conducive to use statistical discovery techniques to learn about some class of visualization techniques. Assume that we wish to carry out an analysis of how many data dimensions can be relocated to visual texture. The first distinct question is: How many perceptually distinguished texture dimensions are there? And hence: How can we productively map data dimensions to them? If the answer cannot be detected in the research, one way to continue is to use a kind of statistical data-mining method to find the answer. We might ask people to assort textures in as many various ways as we can think of (e.g., roughness, regularity, or fuzziness). Then apply a statistical technique to find out how many dimensions there really are in the subjects' responses.

In visualization, multiple regression is a statistical technique that can be used to find out whether it is attainable to foreseesome response variable from display particulars. For example, the time which it takes to judge the shortest path in a node–link diagram could be predicted from the number of link crossings in the diagram and the bendiness of the path[13]. Judgment of the associated connected terms diagram enables the user to predict and get insights into the terms which often tend to be mentioned associated frequently with specific topics.

The general strategy for building simple cognitive models is viable to the problems of graph aesthetics. If we explore a task such as locate the shortest path between two nodes, we can measure how the time to carry out such a task depends on various factors, such as the path length, the continuity of the path and the number of edge crossings on the path.If such anapproach can createreliable results, we can measure the cognitive cost of the number of edges leaving inner nodes and eachcrossing in the graph. Cognitive costs enable us to understand how to get more done while conserving as much of your mental reserve as possible.The results could then be used to implementoutstanding layouts to support a set of tasks that are in prospect in the use of a node-link diagram.

Ware. C et.al developed a stable paradigm to measure the cognitive cost of the information visualization diagrams [13]. That paradigm measures the time to grasp the shortest path between two nodes in a spring layout graph. Their method implicated generating a large number of graphs each with a unique shortest path between two specified nodes, in which the following factors varied:

- correct value for the shortestpathlength (spl):3, 4 or 5
- number of crossing edges on the shortest path (cr)
- total number of crossed edges in the graph drawing (tcr)
- the number of edges branching fromnodes along the shortest path (br)

The graph instance in figure 4 exemplifies the Ware C. et al paradigm.

Applying the analysis paradigm onto the graph in figure 4 by extracting the spl, cr, tcr, and brfactors for the path between the nodes (Food and Inflation) will result the following:

spl=4, cr= 1 , tcr= 6, and br= 5;

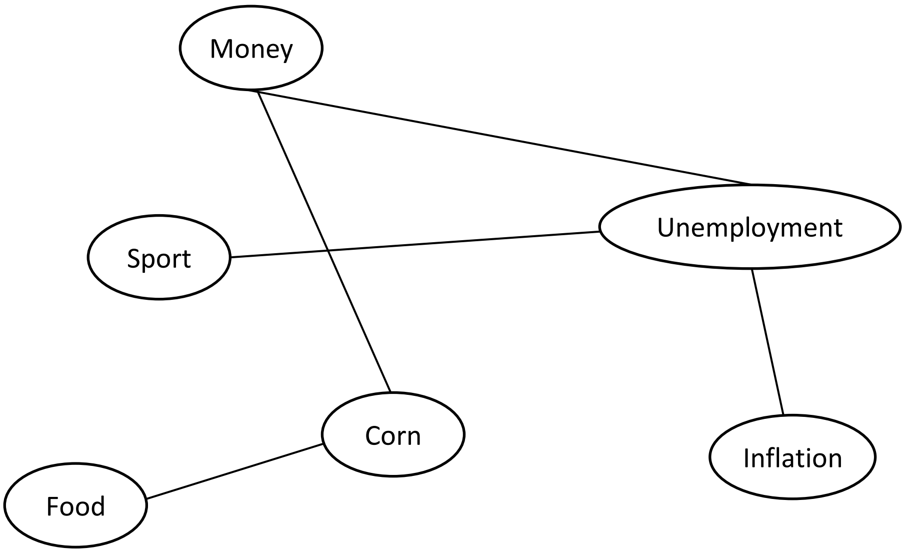

Figure 4: A graph instance exemplifies for explaining multiple regression

By observing and recording these factors for all graphs which visualize the association rules we get a table containing the factor values. Thereafter implementing the multiple regression analysis on that table to examine the factors and hence indicate the extent to which dependent value (i.e. response time rt) correlates with the independent variables (i.e. spl, cr, tcr, and br).

C. Multiple regression analysis (result analysis)

Though, linear correlations are imperfect for convenient interpretation, as there are many significant correlations between the independent variables themselves (i.e. cr, tcr, and br). We need to assure that the interior relationships between the independent variables are ‘factored out’, so we can match the relative contribution of each variable independent of its relationship to the other variables.

Performing multiple regression analysis on the graph collection, with response time as the dependent variable, the following equation relates the response time to the shortest path length (spl), the number of crossing edges on the shortest path (cr) and the number of edges branching from nodes along the shortest path (br). However the other independent variables were not significant.

$$rt= 5.072 + .722\ spl -.014\ cr -.14\ br \qquad (2)$$

It is best to measure “How good a fit is the least square regression equation (2) for our given data?” For answering this question the coefficient of multiple determinations (R^2) is examined [14]. The coefficient of multiple determinations is the value that can indicate the extent to which the dependent value correlates with the independent variables on the left hand side in equation (2).Besides it is a scalar that is defined as the Pearson correlation coefficient between the predicted and the actual values of the dependent variable in a linear regression model that includes an intercept.

$R^2$for the extracted equation (2) = 0.767, which means that about 76.7% of the variation in the response variable (i.e. response time (rt) as a cognition measurement to the visualization graph) can be explained from the least-squares equation (2) and the corresponding joint variation of the variables spl, cr, and br taken together. The remaining 100% - 76.7%=23.3% of the variation in rt is due to random chance or possibly the presence of other variables not included in this regression equation.

Figure 5 showing the scatter plots graph of the average response timefor each graph drawing over all subjects.

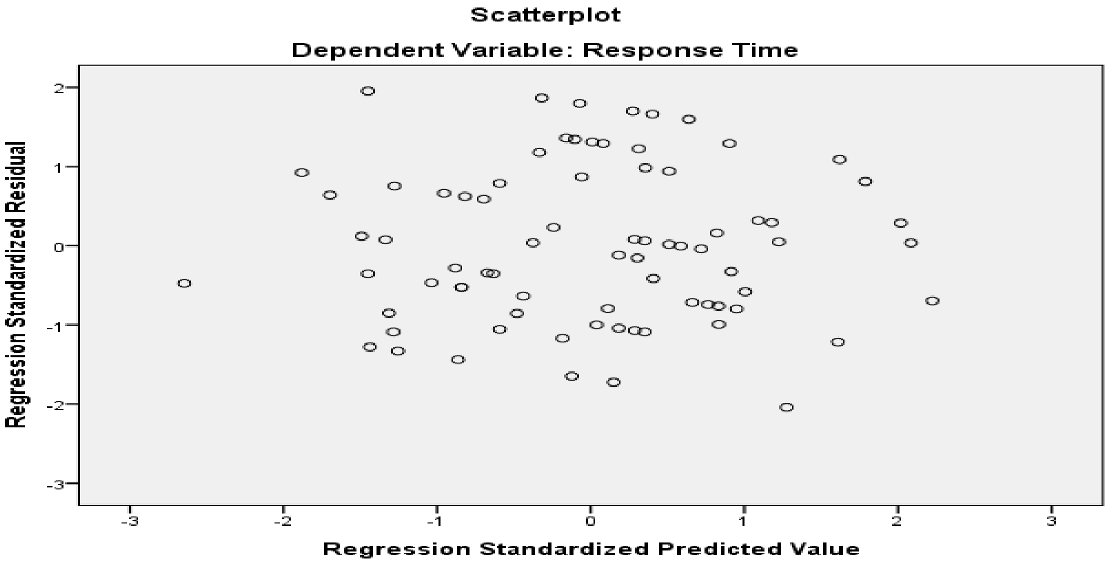

Figure 5: Scatter plots graph of the average response time for drawing over all subjects

6 CONCLUSION AND FUTURE WORK

As a contribution to graph layout research we visualize the results of association rule text mining technique. The proposed model used the tree graph as a popular technique to depict item associations. Nodes of the tree represent the terms, and the edges represent the associations between these terms.The most significant results of this work is the finding that the shortest path length (spl), number of crossing edges on the shortest path (cr), and the number of edges branching from nodes along the shortest path (br) participate in amplifying the user cognition and take cognizance of the visualized graph.For future work we will carry out the classification and naïve Bayes technique; hence visualizethe results of them as we did with the association rules.

References

[1] (2013, Dec.) i2 – Analyst’s Notebook. [Online]. http://www.i2inc.com

[2] D Jonker, W Wright, D Schroh, P Proulx, and B Cort, "Information triage with TRIST," in International Conference on Intelligence Analysis, McLean, USA, 2005.

[3] C Görg, Z Liu, N Parekh, K Singhal, and J Stasko, "Jigsaw meets Blue Iguanodon – The VAST 2007 Contest. Proceedings of IEEE," in Symposium on Visual Analytics Science and Technology , Sacramento, pp. 201-202, 2007.

[4] X. Wang et al., "Defining and applying knowledge conversion processes to a visual analytics system.," Computers & Graphics, vol. 33, no. 5, pp. 616–623, October 2009.

[5] J. D. Fekete, "The infovis toolkit. In Information Visualization, 2004. INFOVIS 2004. ," in IEEE Symposium on infovis , Austin, TX, pp. 167-174, 2004.

[6] M. Chen et al., "Data, information, and knowledge in visualization.," in Computer Graphics and Applications, vol. 29, pp. 12-19, 2009.

[7] H. Jiawei and K. Micheline, Data mining: Concepts and Techniques.: Morgan Kaufmann Publishers, 2000.

[8] H. Beth, H. Michelle, H. Susan, and W. Paul, "Visualizing the Full Spectrum of Document Relationships.," in the Fifth International Society for Knowledge Organization, France, 1998.

[9] Thomas, J., Cook, K., Crow, V., Hetzler, B., May, R., McQuerry, D. & Wong, P. C. Human—Computer Interaction with Global Information Spaces—Beyond Data Mining. In Digital Media: The Future. Springer London, pp. 32-46, 2000.

[10] Bornelöv, S., Enroth, S., & Komorowski, J. Visualization of Rules in Rule-Based Classifiers. In Intelligent Decision Technologies. Springer Berlin Heidelberg, pp. 329-338, 2012.

[11] Reuters-21578 Test Collections. [Online]. http://www.daviddlewis.com/resources/testcollections/reuters21578/

[12] M. F., Porter, "An algorithm for suffix stripping," Electronic library and information systems, vol. 14 , no. 3, pp. 130-137, 1980.

[13] C. Ware, H. Purchase, L. Colpoys, and M. McGill, "Cognitive measurements of graph aesthetics," Information Visualization, vol. 1, no. 2, pp. 103-110, 2002.

[14] T. Keith, Multiple regression and beyond. Boston: Pearson Education, 2006.

BIOGRAPHY

MohammedMasoudAbdallahis an Assistant Lecturer in Computer Science Department at Faculty of Science, MiniaUniversity, Egypt. He has done M.A. in computer science from Minia University in 2010. He was among university toppers. He is also pursuing his PhD in Computer Science. Mohammed is devoting his research work in field of XML Text Mining. He has developed a number of research projects in field of Text Mining including classifying and clustering the XML text-based documents and visualizing the mining results etc.

ZakiTaha Ahmed Fayed, professor, received undergraduate degree in Computer Engineering in 1976 from Ain Shams University, Egypt, and Master Graduate degree in Computer EngineeringinIdentify the character of the speaker through the analysis of voice tag 1991from Ain Shams University, and Ph. D. degree in Computer Engineering from Ain Shams University, in 1997.

تجسيد قواعد إرتباط نص الـ XML الهيكلى الدلالى المحلل

أ.د / زكى طه فايد
أستاذ بكلية الحاسبات و المعلومات جامعة عين شمس
ZTFayed@cis.asu.edu.eg

محمد مسعود عبدالله
مدرس مساعد- كلية العلوم جامعة المنيا
Mas_m4@hotmail.com

خلاصة:

هذه المقالة تقدم طريقة جديدة لتجسيد و تصوير قواعد إرتباط المكونات النصية لملفات الـ XML. هذه المكونات عبارة عن الكلمات المفتاحية الدلالية لكل جملة على حدة و من خلال التدقيق فى هذه القواعد يمكننا توقع و التنبؤ بمدى إرتباط كل كلمة مفتاحية بالأخرى.

تجسيد و تصوير قواعد الإرتباط يكون بإستخدام تقنيات تصور المعلومات. هذا التجسيد يعمل على تضخيم إدراك المشاهد بحيث يستطيع إدراك و فهم قواعد إرتباط نصوص الـ XMLبسهولة و يسر و فى أقل وقت بمجرد مشاهدة هذا التجسيد لقواعد الإرتباط.