American Journal of Applied Sciences 10 (6): 606-614, 2013
ISSN: 1546-9239

doi:10.3844/ajassp.2013.606.614 Published Online 10 (6) 2013 (http://www.thescipub.com/ajas.toc)

Modelling Knowledge Summarization by Evolving Fuzzy Grammar

[1]Nurfadhlina Mohd Sharef, [2]Alfian Abdul Halin and [1]Norwati Mustapha

[1]Department of Computer Science,
[2]Department of Multimedia,
Faculty of Computer Science and Information Technology,
University of Putra Malaysia, Malaysia

Received 2012-10-04, Revised 2012-11-26; Accepted 2013-06-10

ABSTRACT

Summarized text is a simplified and condensed version of the original text containing highlighted information to help the audience get the gist in a short period of time. Typically, text summarization produces abstract or a paragraph-like outputs by omitting details and irrelevant information. However, the text summary can also be produced in a visualized form, such as a chart, graph or table representing a collection of similar cases. The visualized version generates a statistical-like presentation, which often involves numerical and ordinal observation of the gathered knowledge from the text. This requires lexical syntactic understanding of the text. Essential to achieve this goal is topic identification, message analysis/interpretation and knowledge summarization generation. The objective of this study is to model knowledge summarization problem using the evolving fuzzy grammar technique and we focus on metadata generation for producing visualized knowledge summarization. The process comprises of: (i) Identifying the underlying structure of the texts for knowledge summarization, (ii) represent the identified knowledge for summarization manipulation and (iii) presentation of the summarized knowledge. A prototype called FTCat© is developed as a proof of concept and we demonstrate its practicality in summarizing news reports.

Keywords: Text Summarization, Evolving Fuzzy Grammar, Text Mining

1. INTRODUCTION

The rapid growth of information generated daily may result knowledge flooding if not properly organized and managed. One of the ways to contribute in this situation is by having an automatic text summarization tool. Summarized knowledge allows humans to have a quick understanding of the text. It is also regarded as the highlight of the text content and may be represented in the form of a paragraph or visualized. Automatic text summarization contributes in reducing manual analysis of a large collection of documents. These tasks are typically labor intensive but non-trivial. On the other hand, it is important especially in strategic decision making such as financial analysis, judicial verdict, customer relationship management, tactical military and surgery operations.

Automatic summarization is linked closely with text understanding which imposes several challenges comprising of variations in text formats, expressions and editions which adds up to the ambiguities. Researchers in text summarization have approached this problem from many aspects such as natural language processing (Zhang *et al*., 2011), statistical (Darling and Song, 2011) and machine learning (Conroy and Leary, 2001; Xie *et al*., 2003) and text analysis is the fundamental issue to identify the focus of the texts. The information visualization community offers a wide array of algorithms which is also linked to text identification approaches such as by (Liu *et al*., 2012; Ando *et al*., 2005; Kankar and

Corresponding Author: Nurfadhlina Mohd Sharef, Department of Computer Science,
Faculty of Computer Science and Information Technology, University of Putra Malaysia, Malaysia

Mukherjea, 2005; Zhang *et al*., 2011). However, only several (Zhang *et al*., 2011) have focused on the tightly coupled text summarization and visualization. In fact, this helps people to cope with the ever increasing documents and maximize the knowledge acquisition process.

For example, only a high level of knowledge abstraction which requires analysis of a vast amount of documents would be able to answer questions such as "what is the trend of X issue amongst teenagers?", "what are the major topics discussed by community Y in the first quarter of the year?" and "how is the split of feedback on the active topic presented last week?". We hypothesize that the combination of message identifier and information visualization template that can plot the frequency and the evolution of the occurring messages can help in a quick view of the summarized knowledge.

This study models knowledge summarization problem with Evolving Fuzzy Grammar (EFG) technique. The essence is the facilitation of the texts underlying structure which are then transformed into a higher conceptual level called grammar. The fuzzy notion arises because the text to grammar matching process involves uncertainty; there are variations in possible matching and fuzzy membership allow a degree of similarity to be assigned. The learned grammar are then used to identify the portion of texts that matches with the pattern of texts modeled by the grammar. These knowledge are then reproduced in the form of metadata. A visualization template then manipulates the information in the metadata as a means for the represented summarized knowledge. This study is structured as follows. In the next section we discuss the related approaches to text summarization and information visualization. Section 3 describes the applied evolving fuzzy grammar for knowledge summarization while section four concludes the study.

2. RELATED APPROACH

Text summarization is an automated process that produces a summary of the original content (single or multiple documents) and produces the result in the form of a short passage or a list of main sentences from the original document using computational techniques (Thanadechteemapat, 2010). Automatic text extraction is generally established by automatic topics identification across the collection of the document contents and the reproduction of the analyzed information in a condensed manner.

2.1. Text Summarization Techniques

Topic analysis is the main feature in automatic text summarization (Liu *et al*., 2012; Yeh *et al*., 2005). There are various techniques that have been applied in text summarization comprising of (i) statistical approach (Darling and Song, 2011) which can be used to find the main topic by computing the frequency of words or phrases from the original text and used to construct the result, (ii) machine learning approach such as the supervised learning techniques such as genetic algorithm (Xie *et al*., 2003) and hidden markov model (Conroy and Leary, 2001) to build text classifier and unsupervised learning techniques to find keywords and phrases that have similar characteristics with the trained topic classifier, (iii) Natural Language Processing (NLP) techniques focuses on the understanding the context of the documents such as through rhetoric structure analyzing (Chengcheng and Engineering, 2010), latent semantic models (Yeh *et al*., 2005) and latent dirichlet allocation (Darling and Song, 2011; Liu *et al*., 2012) to produce the summary.

The topic analysis method can be further broken into two approaches, namely the sentenced-based and keyword-based. The sentenced-based method range from the automatic hypertext link (Salton *et al*., 1997) to generate intra-document links, i.e., links between various paragraphs (or sentences) of an article. By placing the paragraphs and the intra-document links on a text relationship map, it is possible to visualize the structure of a document. In the keyword-based topic analysis such as employed in (Liu *et al*., 2012), each topic is characterized by a set of keywords or clues (Carenini *et al*., 2007; Geng *et al*., 2006).

Whilst domain-restricted text summarization (Reeve *et al*., 2007; Ando *et al*., 2005) is easier due to the prominence of topics in the homogeneous document collection, open domain text summarization (Nomoto and Matsumoto, 2003) is more challenging, which is topped by the text expression variation challenges.

In contrast to the approaches discussed above, FTCat© produces summarization of the texts in the XML-based metadata form in order to provide a more structured view and quicker understanding to the user. Another distinguishing character is the EFG method utilizes a context free grammar method which requires the terminal grammar containing its predefined set of words to be prepared and utilizes these to transform the text fragments used in building the classifier into grammars. This syntax and shallow semantic representation approach allows more complex matching

and representation compared to fixed keywords as one word can be defined several granularities, for example in an address classifier the word 'restaurant' can be parsed by 'Food Provider' classifier as well as 'Business'.

2.2. Visualization Technique for Summarized Knowledge Viewing

There are two categories of approaches being developed for visualization based text summarization: metadata-based and content-based text visualization. The former method focuses on visualizing the metadata of text documents, which would result to limited coverage due to the structure of the designed metadata while the latter involves deeper analysis of the text such as identifying the relationship between the paragraphs and documents. The graph-based visualization allow navigation of the information by expandable nodes while statistical based display such as graph and charts offers static information visualization.

PubCloud (Kuo *et al*., 2007) used tag clouds for the summarization of biomedical literature queries. Tag clouds are visually-weighted renditions of collections of words (tags) that can be used to represent the concepts in the documents. The clouds are usually formed based on the computed word frequency and displayed in variable font colors and sizes; larger fonts depict most frequently used words. Tag cloud generation also usually involves some text processing including removing uninformative elements such as stop words and stemming.

Visualization supported text analysis allows the user to make connections within entities in the documents where two entities are connected if they appear in one or more documents together such as adopted in Jigsaw (Stasko *et al*., 2007). Four visualization functions are used in Jigsaw: (i) tabular connections view containing multiple re-orderable lists of entities. The connections between entities are shown by coloring related entities and drawing links between them. (ii) a semantic graph view displaying connections between entities which allows analysts to dynamically explore the documents by showing and hiding links and nodes, (iii) a scatter plot view which highlight pair-wise relationships entities and (iv) a text view displaying the original reports with entities highlighted.

FTCat© benefits from the combination of metadata and content based visualization approaches since the EFG method will first analyze the documents looking for portion of texts recognized by the fuzzy grammar classes and generate a metadata reflecting the manipulated knowledge.

3. KNOWLEDGE SUMMARIZATION WITH EVOLVING FUZZY GRAMMAR

3.1. Metadata Generation with Evolving Fuzzy Grammar for Knowledge Summarisation

As a proof of concept, we developed an application called FTCat© which was mainly motivated by the Worldwide Incident Tracking System (WITS) where crime incident data are stored, managed and visualized. **Fig. 1 and 2** show the example of screenshots in WITS. FTCat© aims to automate the human analysts tasks in WITS by automatic text summarization.

FTCat© has been tested on four domains namely product review, medical reports, economical statement and crime incidents. For the economical statement summarization the likelihood of economical index direction (increasing or decreasing) is identified while the summarized knowledge on crime incidents are the type of events (e.g., bombing, armed attack and arson), the number of wounded victims and number of dead victims.

In WITS XML tags are constructed manually by analysts. The WITS system allows information manipulation including filtering and searching information according to event types, wounded count, dead count, weapon types, victim types. Trending can also be tracked, which is displayed in the form of charts and graphs. The system can be used by many parties, ranging from politicians, social analysts, reporters and public.

Although the generated metadata in FTCAT© has a similar structure to the WITS's but the tags in the FTCAT© metadata is flexible and depends on the trained grammars modeled by the user. To allow flexibility in the text summarisation request, the metadata is generated in two versions: per information entry and cumulatively to represent the total collection made available to the system. This is because user might be interested in requesting summarisation from certain time interval and thus the collective information is no more accurate.

This section will illustrate text summarization using samples from crime incidents and show examples of FTCat© interfaces when tested on the four mentioned domains. The EFG method utilizes the underlying text connotation in such a way that lexicalised information is used to transform the text into more meaningful conceptualization, called grammar. **Table 1** shows the example of terminal grammars and their definitions. The terminal grammars are used to transform the selected text fragments for grammar training as shown in **Table 2** so the converted grammar-form can be generated as shown in **Table 3**.

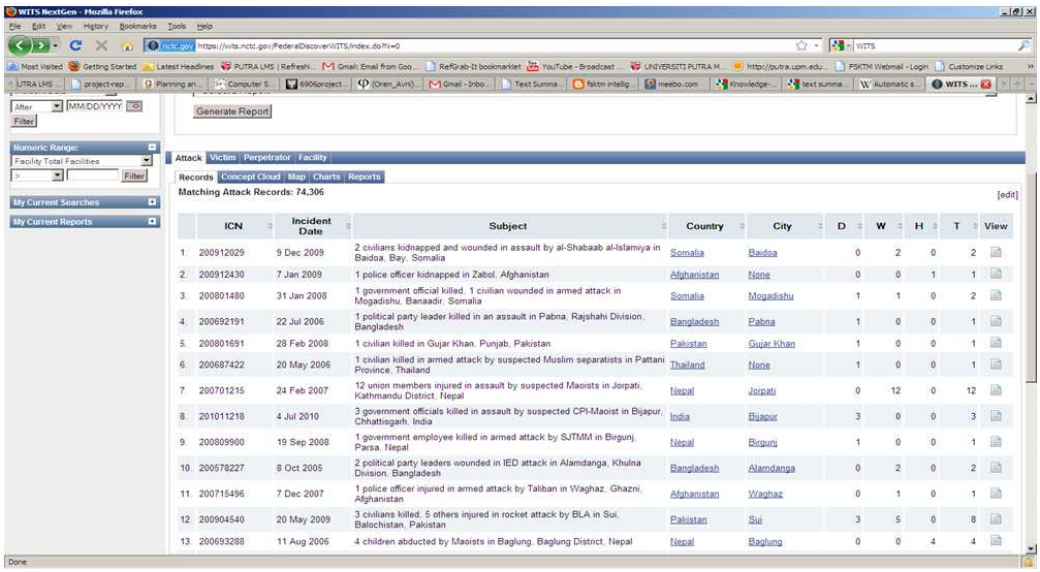

Fig. 1. Tabulated data based on learned categories

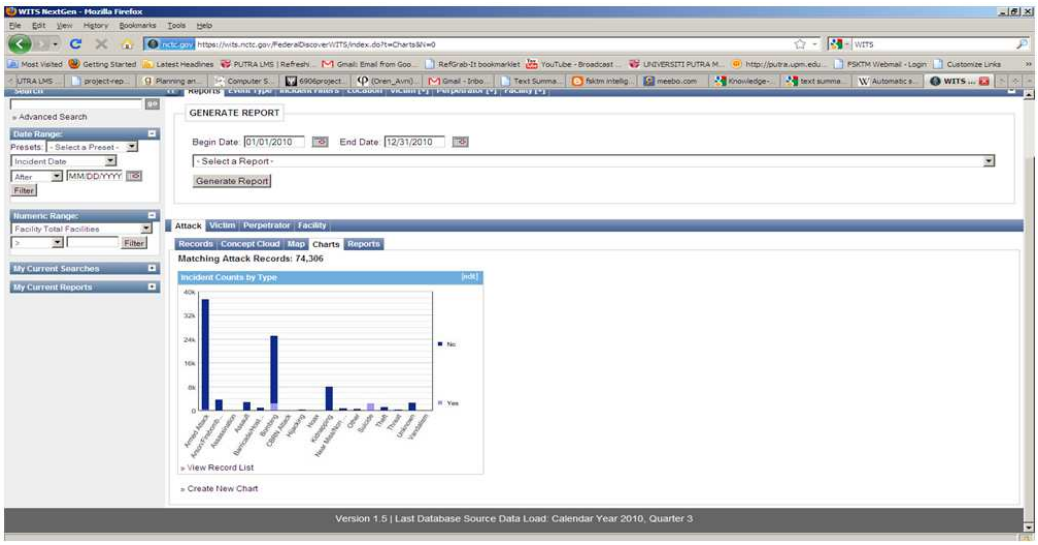

Fig. 2. Chart of incidents record by country

Table 1. Examples of terminal grammars and their definitions

Terminal grammars	Definition
Explosive	bomb, explosive device
Criminal list	assailant
Bomb action	exploded, detonated

The dead count and wounded count detection requires a deeper analysis compared to those performed in event type identification. This is because besides extracting the texts that express dead and wounded messages, numerical processing needs to be performed, as shown in **Table 4**. For text summarization task these information is accumulated and recorded in the FTCat© generated metadata such as shown in **Fig. 3**. As the automatic text summarization should be dynamic, the.

The algorithm for FTCat© is shown in high level pseudocode in Algorithm 1 while Algorithm 2 shows the general algorithm for EFG.

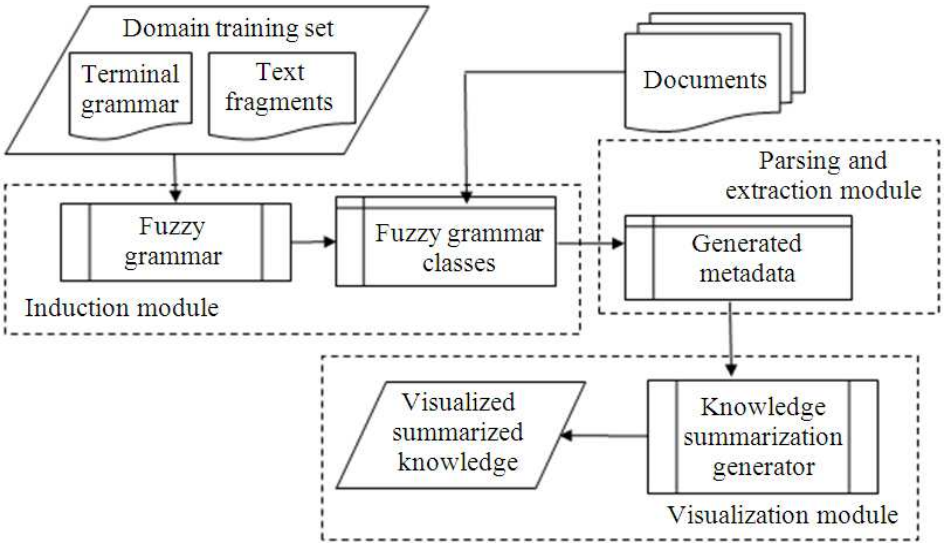

Fig. 3. FTCat© architecture

Table 2. Examples of text fragment examples that can be selected to train the event type grammars

Text fragment examples	XML tag: Event type
- Bomb exploded	Bombing
- Explosion occurred	
- Men detonated a bomb	
- Attacks to occur	Armed Attack
- Attackers threw a grenade	
- Assailants attacked a security vehicle	
- Gunmen killed a member	

Table 3. Examples of grammars for bombing events and the parse-able texts

Grammar examples	Example parses of the grammar
explosive bomb Action	Bomb exploded
[criminal List]	Assailant detonated
Bomb Action explosive	explosive device
Bomb Action explosive	Detonated bomb
bomb_bomb Action	Bomb exploded
[criminal List]	Assailant detonated
Bomb Action explosive	explosive device

Table 4. Examples of grammars for bombing events and the parse-able texts

Data	Dead count	Wounded count
killing two villagers	2	0
beheaded a police official	1	0
killed a lawyer	1	0
injuring 18 others	0	18
no injuries	0	0

The EFG algorithm shows that the fuzzy grammar process should be repeated for each desired fuzzy grammar class. This will also require the preparation of text fragments for training and the terminal grammars that can represent the text fragments. During the visualization execution, the collection of texts will firstly be checked for text that can be recognized by the fuzzy grammar and metadata will be generated (example shown in **Fig. 3**), together with the identified values. Depending on the type of visualization needed, the metadata will be read and manipulated to present information in a suitable form:

Algorithm 1: General FTCat© algorithm
Input: Collection of texts, knowledge summarization request
Output: Metadata, Summarized Knowledge Visualization
Process:
Initialize the trained fuzzy grammars
For each text in the collection and for each of the fuzzy grammar
    execute the text fragment extraction process
    generate metadata
    identify parsed texts according to developed fuzzy grammars
    perform numerical processing if needed
End For
For each of the knowledge summarization request
    Execute metadata manipulation
    Perform numerical processing if trending is needed
End For

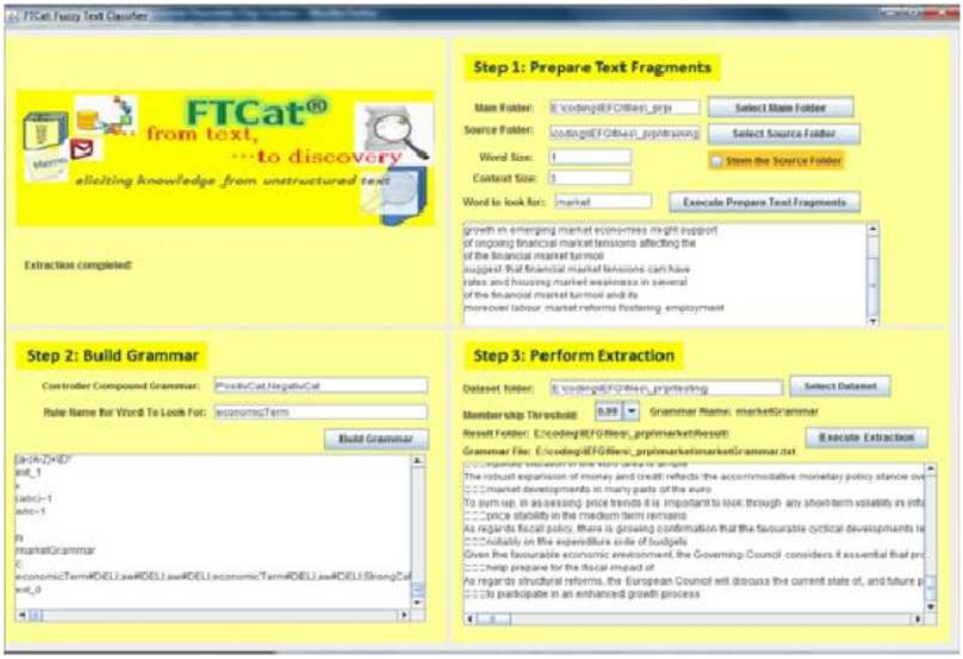

Fig. 4. Main FTCat© interface

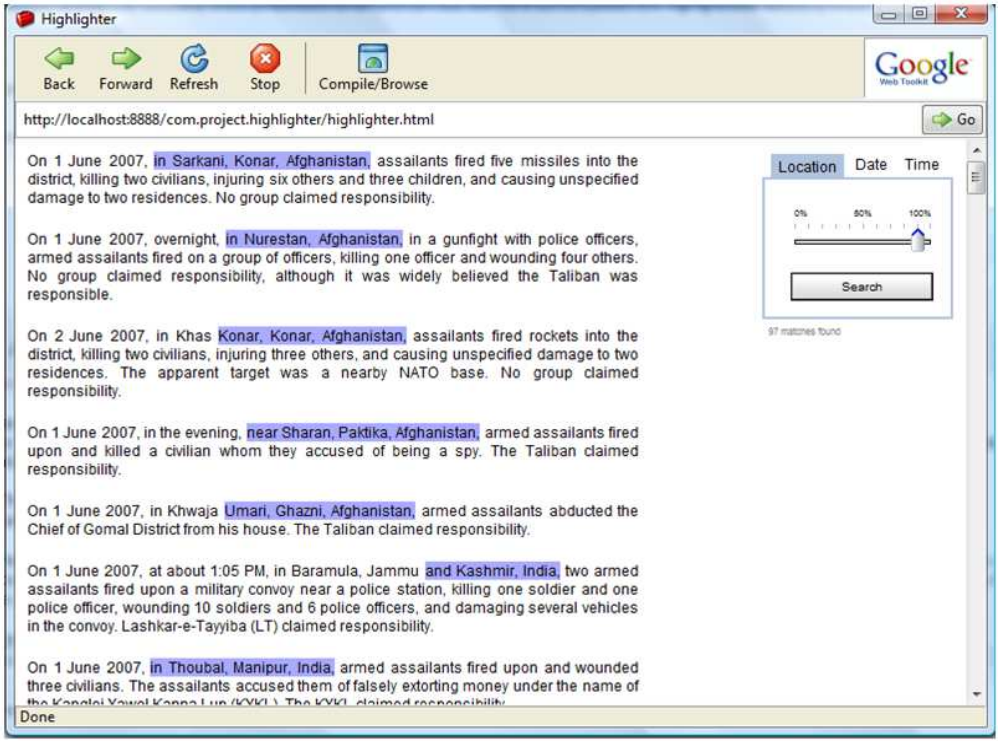

Fig. 5. Screenshot from text extraction function for crime incidents identification

```
<FTCat© >
<CrimeIncidents>
<BombingEvents>169</BombingEvents>
<ArmedAttackEvents>213</ArmedAttackEvents>
<ArsonEvents>67</ArsonEvents>
<WoundedCount>2786</Wounded>
<DeadCount>1656</DeadCount>
</FTCat© >
```

Fig. 6. Example of generated metadata for crime incidents identification

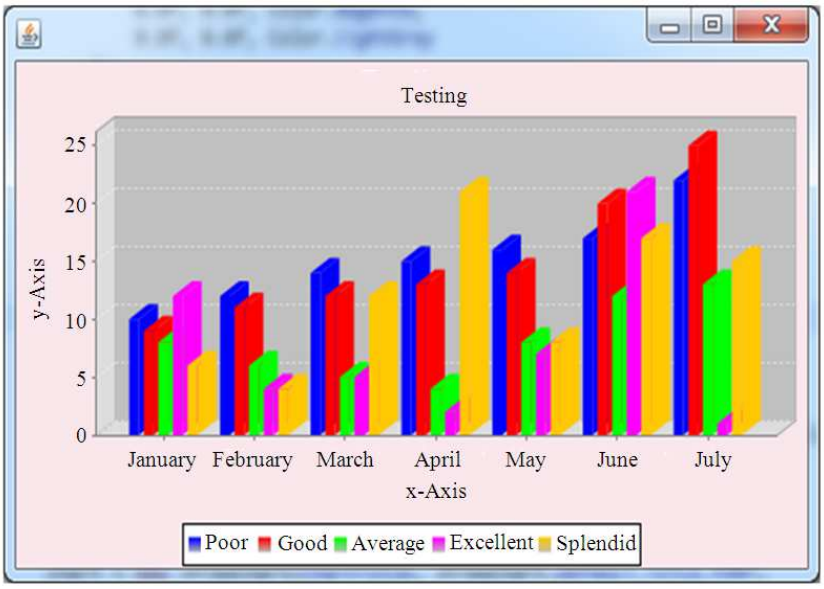

Fig. 7. Example of visualization in FTCat© for product reviews

Algorithm 2: General EFG algorithm
Input: Text fragments for training for each fuzzy grammar classes, terminal grammar, number of fuzzy grammar classes, labels for each of fuzzy grammar classes
Output: Fuzzy grammars
Process:
For each of the fuzzy grammar classes
For each of the text fragments for training
Execute evolving fuzzy grammar
Transform text fragments into fuzzy grammar
If fuzzy grammar is parsed by existing fuzzy grammars Then
Skip
Else
Generalize the existing fuzzy grammar collections by replacing the matching existing fuzzy grammar or adding new fuzzy grammar to the learned fuzzy grammars
End If
End For
Generate metadata by using the labels and assigning suitable values
End For

3.2. Visualization of Metadata-Based Summarised Knowledge

FTCat© consists of three modules (**Fig. 4**) namely induction, parsing and extraction and visualization. The

Induction Module generates the fuzzy grammar based on two inputs namely terminal grammars, which are the lexicon for the text fragments and the selected text fragments; the set of examples used to train the fuzzy grammar classes. For example, to build a fuzzy grammar class for restaurants, the terminal grammars are <restarantName>:={café, restaurant, bistro} and <anyWord>=a-zA-Z. The built fuzzy grammars are <anyWord> <restaurantName>, which will recognize patterns such as 'Summerfield Bistro' and 'Pizzeria Restaurant'. Note the Induction Module may need to be executed several times to build each needed topic classifier.

In the Parsing and Extraction Module, the fuzzy grammars will be used to identify the topics in the supplied documents (**Fig. 4**). In the interface the user can choose the topics and the parsed text fragments are highlighted (**Fig. 5**). The output is the XML-based metadata (**Fig. 6**), comprising of the identified topics and their matching contents. The XML-metadata formed summarized information is manipulated by the Knowledge Summarization Generator so that further processing such as information organization and arithmetic operations can be performed to produce the final visualized display (**Fig. 7**).

4. CONCLUSION

Automatic text summarization could assist in intelligent decision making and mitigate the tedious, manual information organization. Although many text summarization approaches are available, only several researches have focused on the coupling of text-based summarization with a visualization technique. This study models knowledge summarization with EFG by focusing on the coupling of identified important text, utilization of metadata and visualization template. EFG has previously been applied as text extractor and text classifier. The developed FTCat© application extends the existing EFG function for metadata generation which are manipulated by the visualization template to produced summarized knowledge visually. The combination of these components distinguishes this study from others that focus solely on text based or graphical based summarization. As a future work it would be interesting to enrich EFG to incorporate higher level of semantic representation compared to the current shallow semantic approach. This is hypothesized to be more powerful in representing more complex knowledge in text mining applications.

5. REFERENCES

Ando, R., B. Boguraev, R. Byrd and M. Neff, 2005. Visualization-enabled multi-document summarization by Iterative Residual Rescaling. Nat. Lang. Eng., 11: 67-86. DOI: 10.1017/S1351324904003389

Carenini, G., R.T. Ng and X. Zhou, 2007. Summarizing email conversations with clue words. Proceedings of the 16th International Conference on World Wide Web, (WW' 07), pp: 91-10. DOI: 10.1145/1242572.1242586

Chengcheng, L. and I. Engineering, 2010. Automatic text summarization based on rhetorical structure theory. Proceedings of the International Conference on Computer Application and System Modeling, Oct. 22-24, IEEE Xplore Press, Taiyuan, pp: V13-595-V13-598. DOI: 10.1109/ICCASM.2010.5622918

Conroy, J.M. and D.P. Leary, 2001. Text summarization via hidden markov models. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sept. 9-12, ACM Press, New Orleans, LA, USA., pp: 406-407. DOI: 10.1145/383952.384042

Darling, W.M. and F. Song, 2011. Probabilistic document modeling for syntax removal in text summarization. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, (CL' 11), ACM Press, Stroudsburg, PA., pp: 642-647.

Geng, H., P. Zhao, E. Chen and Q. Cai, 2006. A novel automatic text summarization study based on term co-occurrence. Proceedings of the 5th IEEE International Conference on Cognitive Informatics, Jul. 17-19, IEEE Xplore Press, Beijing, pp: 601-606. DOI: 10.1109/COGINF.2006.365553

Kankar, P. and S. Mukherjea, 2005. Text-based summarization and visualization of gene clusters. Proceedings of the ACM Symposium on Applied Computing, Mar. 13-17, ACM Press, Santa Fe, NM, USA., pp: 210-211. DOI: 10.1145/1066677.1066728

Kuo, B.Y.L., T. Hentrich, B.M. Good and M.D. Wilkinson, 2007. Tag clouds for summarizing web search results. Proceedings of the 16th International Conference on World Wide Web, May 8-12, ACM Press, Banff, AB, Canada, pp: 1203-1204. DOI: 10.1145/1242572.1242766

Liu, S., M.X. Zhou, S. Pan, Y. Song and W. Qian *et al*., 2012. TIARA: Interactive, topic-based visual text summarization and analysis. ACM Trans. Intell. Syst. Technol., 3: 25-25. DOI: 10.1145/2089094.2089101

Nomoto, T. and Y. Matsumoto, 2003. The diversity-based approach to open-domain text summarization. Inform. Proc. Manage., 39: 363-389. DOI: 10.1016/S0306-4573(02)00096-1

Reeve, L.H. H. Han and A.D. Brooks, 2007. The use of domain-specific concepts in biomedical text summarization. Inform. Proc. Manage., 43: 1765-1776. DOI: 10.1016/j.ipm.2007.01.026

Salton, G., A. Singhal, M. Mitra and C. Buckley, 1997. Automatic text structuring and summarization. Inform. Proc. Manage., 33: 193-207. DOI: 10.1016/S0306-4573(96)00062-3

Stasko, J., C. Gorg, Z. Liu and K. Singhal, 2007. Jigsaw: Supporting investigative analysis through interactive visualization. Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, Nov. 30-Dec. 1, Sacramento, California, USA., pp: 131-138.

Thanadechteemapat, W., 2010. Discover information and knowledge from websites using an integrated summarization and visualization framework. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Jan. 9-10, IEEE Xplore Press, Phuket, pp: 232-235. DOI: 10.1109/WKDD.2010.109

Xie, Z., X. Li, B.D. Eugenio, P.C. Nelson and W. Xiao *et al*., 2003. Using gene expression programming to construct sentence ranking functions for text summarization. Proceedings of the 20th International Conference on Computational Linguistics, (CL' 03), ACM Press, Stroudsburg, PA, USA., pp: 2-5. DOI: 10.3115/1220355.1220557

Yeh, J.Y., H.R. Ke, W.P. Yang and I.H. Meng, 2005. Text summarization using a trainable summarizer and latent semantic analysis. Inform. Proc. Manage., 41: 75-95. DOI: 10.1016/j.ipm.2004.04.003

Zhang, Y., D. Wang and T. Li, 2011. iDVS: An interactive multi-document visual summarization system. Mach. Learn. Know. Disco. Databases, 6913: 569-584. DOI: 10.1007/978-3-642-23808-6_37