



INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES



Visualizing Military Explicit Knowledge using Document Clustering Techniques

Zuraini Zainol, Afiqah M. Azahari, Sharyar Wani, Syahaneim Marzukhi, Puteri N.E. Nohuddin and Omar Zakaria

To Link this Article: <http://dx.doi.org/10.6007/IJARBSS/v8-i6/4307>

DOI: 10.6007/IJARBSS/v8-i6/4307

Received: 23 May 2018, **Revised:** 19 June 2018, **Accepted:** 29 June 2018

Published Online: 09 July 2018

In-Text Citation: (Zainol et al., 2018)

To Cite this Article: Zainol, Z., Azahari, A. M., Wani, S., Marzukhi, S., Nohuddin, P. N. E., & Zakaria, O. (2018). Visualizing Military Explicit Knowledge using Document Clustering Techniques. *International Journal of Academic Research in Business and Social Sciences*, 8(6), 1127–1143.

Copyright: © 2018 The Author(s)

Published by Human Resource Management Academic Research Society (www.hrmars.com)

This article is published under the Creative Commons Attribution (CC BY 4.0) license. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this license may be seen

at: <http://creativecommons.org/licenses/by/4.0/legalcode>

Vol. 8, No. 6, June 2018, Pg. 1127 - 1143

<http://hrmars.com/index.php/pages/detail/IJARBSS>

JOURNAL HOMEPAGE

Full Terms & Conditions of access and use can be found at
<http://hrmars.com/index.php/pages/detail/publication-ethics>



INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES



Visualizing Military Explicit Knowledge using Document Clustering Techniques

Zuraini Zainol^{1*}, Afiqah M. Azahari², Sharyar Wani³, Syahaneim
Marzukhi⁴, Puteri N.E. Nohuddin⁵ and Omar Zakaria⁶

^{1,2,4,6} *Department of Computer Science, Faculty of Defence and Science Technology, National Defence
University of Malaysia, Sungai Besi Camp 57000 Kuala Lumpur, Malaysia*

³ *Cyber Security Centre, National Defence University of Malaysia, Sungai Besi Camp
57000 Kuala Lumpur, Malaysia*

⁵ *Institute of Visual Informatics, National University of Malaysia, Bangi 43600 Selangor, Malaysia*

Abstract

Speed of decision making, increased operations tempo and enhanced situational awareness are some of the essential characteristics required by United Nations Peacekeeping forces. These can be efficiently achieved by leveraging with enhanced information processing technologies. The process of discovering essential information from the explicit knowledge and its digitization producing web and electronic documents will produce a more effective and efficient decision making environment, enhancing the value of the decision making process. The current work deals with discovering and visualizing useful patterns and knowledge especially in unstructured text, within the domain of United Nations Peacekeeping operations. Text analytics is a powerful technique that helps in achieving the aforementioned goals. Text Analytics of Unstructured Data (TAUD) framework is utilized for analyzing, discovering and visualizing essential patterns and knowledge within the military text documents. The framework focuses on data collection, pre-processing, text analytics and visualization using the hierarchical cluster analysis and K-Means algorithm. The results upon verification indicate that the technique has successfully extracted important knowledge areas or information required by the troops before being deployed in the ground operations. It will substantially help the military commanders and training officers to have an easy and efficient access to all the essential military knowledge before and during deployment on an assignment. Hence, it may lead to an efficient and enhanced situational awareness and decision making during a peacekeeping mission.

Keywords: *Text Analytics, Document Clustering, Framework, Peacekeeping, Visualization, Decision Making*

Introduction

Peacekeeping is one of the among a range of activities undertaken by the United Nations (UN) to maintain international peace and security throughout the world (Hultman et al., 2013). UN is an

international organization founded in 1945 after the Second World War and currently made up of 193 Member States. The mission and work of the UN are guided by the purposes and principles contained in its founding Charter, to promote international cooperation and prevent future conflicts. The role of these peacekeeping forces is to prevent future conflict, support stabilization and rebuilding efforts and not to engage in combat ("United Nation," 2017). Peacekeeping operations have evolved from simply monitoring ceasefires to protecting civilians, disarming ex-combatants, protecting human rights, promoting the rule of law, supporting free and fair elections, minimizing the risk of land-mines, etc. Currently, United Nations has multiple active peacekeeping operations around the world. Peacekeeping is one of the central activities to UN, assisting host countries having difficulties due to conflict and transitioning to peace ("United Nation Peacekeeping," 2017).

UN Peacekeepers come from the Member States, from all around the world, with diverse cultural backgrounds. UN Peacekeepers are an amalgam of civilian, military and police personnel that work together to pursue peacekeeping mandates. However, military personnel are the backbone and the most visible component of a peacekeeping operation. UN military personnel can be called upon to: protect civilians and UN personnel, monitor a disputed border, monitor and observe peace process in post-conflict areas, provide security across a conflict zone, assist in-country military personnel with training and support, assist ex-combatants in implementing the peace agreements ("United Nation Peacekeeping," 2017). Therefore, in order to implement their mandated tasks, the military personnel need to prepare adequately before deployment. This preparation covers every aspect of UN Peacekeeping such as equipment and training amongst others like protection of civilians, use of force and the rules of engagement etc. In addition, the military personnel also focus on conduct and discipline in each training item. Thus, the knowledge regarding training and equipment, vital information and in-depth understanding of the mission including previous and current conflicts, are vital to them. The main challenges are to select the "right" information from various sources that can be transformed into meaningful knowledge. The extracted knowledge can be used to assist the military personnel in decision making and pre-deployment preparation.

This work investigates the explicit knowledge i.e. documented knowledge. One example of the explicit knowledge is UN Peacekeeping Training Manual that can be downloaded from the UN website ("United Nation Peacekeeping," 2017). These guidelines describe the core roles and scope of operational employment activities for UN Military Experts on mission in UN Department of Peacekeeping Operations (DPKO). Further, these guidelines link the roles of military personnel to the skill sets that have been identified by DPKO across all UN peacekeeping missions. These training manuals also cover important topics such as weapon training, military training according to UN operating techniques, safety measures and precautions, specialized training areas and exercises.

The advancement of technology has led to a convenient platform in order to re-produce all the explicit knowledge into better sources such as digitized documents, web documents and electronic documents. For example, the digitized documents are easier to manage including keeping, copying and reproducing, both documents and knowledge. Additionally, there are many document management techniques available to improvise methods for handling and analyzing content of documents. Since the task of the military personnel is varied and crucial, therefore, the explicit knowledge is vital to them. It includes important information and in-depth understanding of the mission. Thus, an efficient technique is needed (e.g. data mining technique) to extract the knowledge and identify hidden information pattern of the mission from various sources e.g. data from reports, briefings, books, guidelines, notes and information.

Based on various sources and types of data, it may be either classified as unstructured or structured. Different source of unstructured data include photos, maps from a GPS device, videos, audio, emails, tweets, and text messages from various sources, etc. Structured data on the other hand is typically the data that is stored in rows and columns of a database (Galbraith, 2014). Data mining technique helps finding knowledge from various re-sources and types including raw and unprocessed data. Using data mining techniques, knowledge can be extracted through the process of sorting and extracting from a large pool of data (Sumathi & Sivanandam, 2006). Usually, the data from various sources and types will store in the data warehouse, whereas the knowledge is extracted using the data mining technique. Here, the data warehouse will store data from various sources and transactions, whereas the knowledge is extracted using the data mining technique such as text mining for discovering useful patterns and knowledge in unstructured text documents. From the literature, there were numerous data mining techniques have been used for knowledge discovery such as classification, clustering, association and prediction (Sumathi & Sivanandam, 2006). The aim of this paper is to determine an efficient text mining technique for analyzing the documents to discover and extract useful knowledge from military peacekeeping documents.

Military Explicit Knowledge

Knowledge is the cornerstone of information revolution and is widely related to process data or aptitudes gained amid a learning procedure. It is often defined as “a fluid mix of framed experience, contextual information, values and expert insight that provides a framework for evaluating and incorporating new experiences and information” (Davenport & Prusak, 1998; Lau, 2016). Choosing the appropriate information from varied sources and transforming them into competent knowledge poses a significant challenge. Documented knowledge is referred to as “explicit” while knowledge which is obtained from experience is known as tacit knowledge. Implied learning or tacit knowledge requires sound judgment and is often stored mostly in the human’s mind (Yusof et al., 2016). Tacit knowledge can neither be codified nor expressed. On the other hand, explicit knowledge is often based on scholarly accomplishments and hence expressible (Lau, 2016). Neither of the two types seems to be utilized to their full potential (Smith, 2001). The aforementioned categorization remains same to the military domain. Knowledge that can easily be articulated, codified, stored and retrieved using different media forms comprises of explicit military knowledge. Media forms include doctrines, tactics, techniques, manuals etc. Explicit knowledge is often expressed in a formal and systematic manner. Furthermore, it can be easily transferred, communicated and shared by many people. Normally, it has been documented in books, papers, training manuals, standard operation procedures, etc. (Alkhred et al., 2018; Nohuddin & Zainol, 2014). The training manual of United Nations Peacekeeping Force also known as ‘Training Guidelines for National or Regional Training Programmes’ serves the role of explicit knowledge for the experimentation procedure. The manual discusses some fundamental subjects relating weapon training, military procedures, safety and precautionary measures, specialized training, etc.

Text Analytics for Knowledge Discovery & Visualization

Text Mining (TM) is a data mining technique that helps in mining operations form text documents. It is also referred as Text Analytics (TA). The process of discovering valid, novel, useful and meaningful patterns from text is known as Knowledge Discovery in Text (KDT) (Feldman & Dagan, 1995). Data mining and TM differ from each other over the type of data used. While data mining only deals with structured data, TM covers both unstructured and semi-structured data (Desai, 2015). TM is

frequently used for finding hidden knowledge behind unstructured text documents. It is an important technique extending data mining, encompassing other disciplines such as information retrieval, statistics, natural language processing, etc. for the process. Typically, information retrieval systems help in finding domain related text documents, natural language processing helps to analyze human language so that computers understand it as humans do, and information extraction deals with extraction of structured data from unstructured sources (Rathi et al., 2010). The analysis process involves various techniques such as intelligent analysis, data mining and knowledge discovery in text. Therefore, TA is often used in many application domains such as healthcare for investigating and clinical decisions (Massey, 2015), social media for preventing crime and fraud detection, biomedical for discovering new drugs, etc. Thus, text mining aims to marshal unstructured data and derive useful insights to solve real-world problems (Godbole et al., 2010).

Typical TM tasks include text categorization, text clustering, and concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e. identifying relations between entities). On the other hand, the analytical process involves various techniques (Chakraborty et al., 2014) such as information retrieval, exploratory analysis includes topic extraction, cluster analysis, concept extraction, summarization, categorization, sentiment analysis, content management and ontology management. TM comprises several processes: collection of dataset, text parsing, text filtering, text parsing and text mining. Figure 1 illustrates and summarizes the TM process flow.

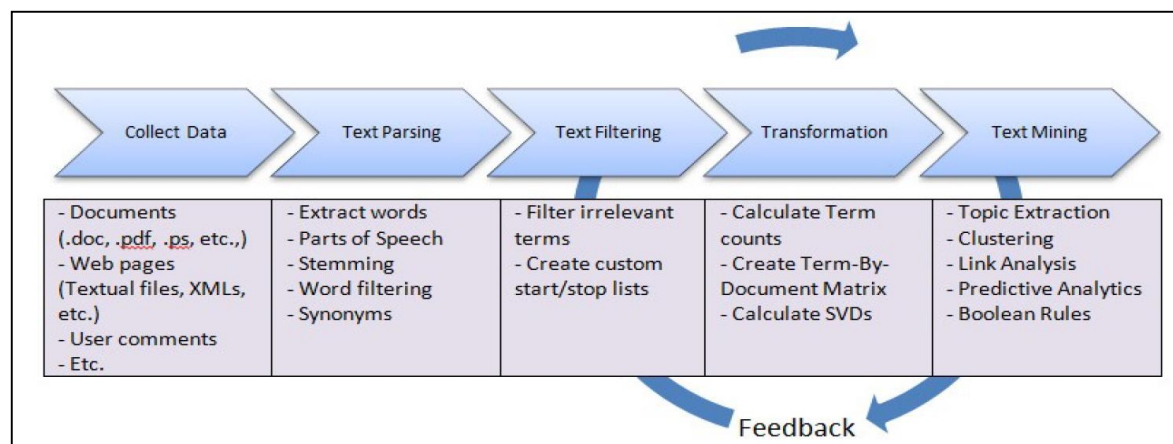


Figure 1. Text mining process flow adopted from (Chakraborty et al., 2014).

Clustering involves grouping similar objects, text, documents, etc. This technique has been practiced by humans for many thousands of years, to group objects according to defined features and similarity (Rodriguez & Laio, 2014). Clustering technique has been applied in several areas such as biology (Rodriguez & Laio, 2014), market share (Mukherjee & Bandyopadhyay, 2015), spatial data (Esmine et al., 2015) (high-dimensional data), trend analysis (Amin et al., 2014; Nohuddin et al., 2018), image processing (Hatamlou, 2013) and document clustering (Nohuddin et al., 2015; Zainol et al., 2017).

Classification of similar text documents into groups is often referred as document clustering. The documents that are semantically relevant and readily understandable are grouped together in form of specific contexts or topics. This is essential for organization, browsing, retrieval, etc. (Xie & Xing, 2013). With huge amounts of data contained within the documents, achievement of better retrieval

efficiency is highly challenging. Document clustering can be applied to maximize the efficiency of information access within the documents (Negm et al., 2013).

Document clustering aims on determining the likelihood of a set in patterns, documents, points or items. Further, from organize intelligible clusters, searching and understanding the information become less complex. Besides, massive amounts of unordered text documents if it clustered would benefits user in accessing and ordering information (Karol & Mangat, 2013; Nanda & Panda, 2014). Like other clustering methods, document clustering also requires an efficient algorithm, which can solve the clustering problem. The process of information retrieval including search engines is an important example of document clustering (Xie & Xing, 2013). Figure 2 illustrates a basic document clustering process.

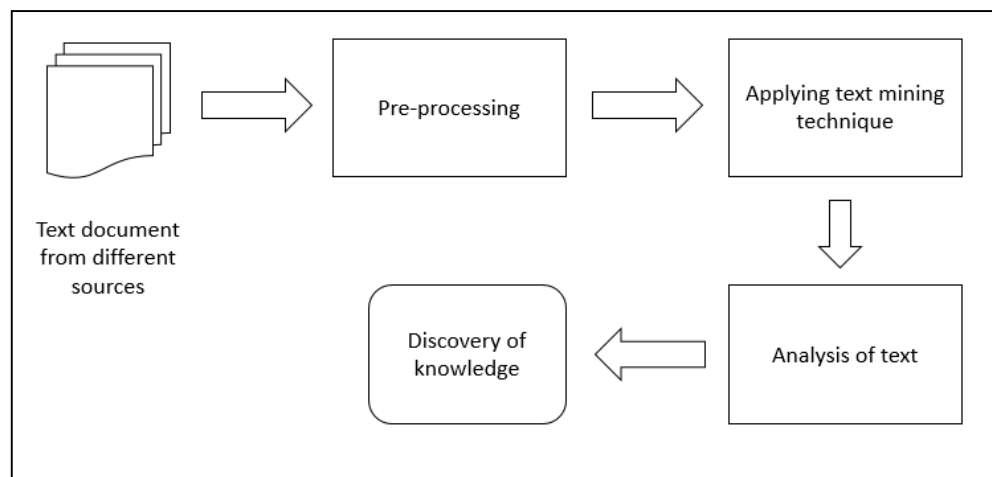


Figure 2. Basic steps in document clustering (Jensi & Jiji, 2014)

Pre-processing is essential in document clustering, in order to remove unsolicited data such as punctuation, extraneous markup and capitalization. Words that do not carry semantic meaning e.g; stop words (such as "the") are removed in this phase. In addition, stop words will be removed to reduce the amount of noise in the text collection (Latif & Wood, 2008). Tyagi et al. (2010) presented an in-depth discussion about the need and methodology of preprocessing. Encoding the text document is the next step after pre-processing. The document will be transferred into a document term matrix (DTM), a mathematical matrix defining the frequency of word occurrences in the document. It is followed by the process of dimension reduction, whereby varied number of techniques can be applied for this process e.g. feature selection and feature extraction. The objective of the dimension reduction, is to allow less dimension for a wider comparisons of the thoughts contained in the text group. Therefore, this technique helps to remove noisy, irrelevant and redundant attributes in the document (Jensi & Jiji, 2014).

The next step involves clustering, by dividing and grouping similar text into a same cluster. Several algorithms that can be applied for this phase such as:

a. K-means clustering

K-means clustering is a type of unsupervised learning that is mostly used in unlabeled data (Negm et al., 2013). Data with undefined clusters or groupings is represented in a group with a variable k , assigning each data point to a proper cluster based on maximum similarity. According to the

features that are provided using k , the algorithm will iteratively cluster each data point to one of the groups. Figure 3 illustrates an example of k -Means clustering

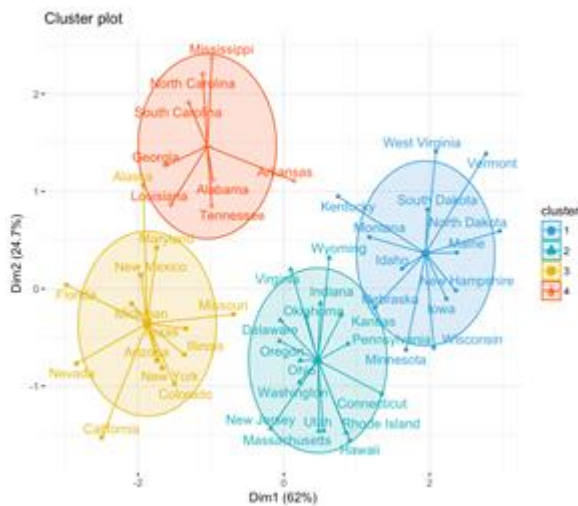


Figure 3. K-Means clustering (Dendrogram) adopted from (Kassambara, 2017)

a. Hierarchical clustering

Hierarchical clustering provides a view of data in a different level of granularity. According to Zhao and Karypis (2002), the algorithm is ideal to visualize and interactively explore a collection of large documents. Figure 4 is an example of hierarchical clustering. As illustrated, each level of hierarchy separates a set of clusters. Each of the bottom-most hierarchy has their unique cluster. However, as the hierarchy goes up, all groups belong to the same cluster.

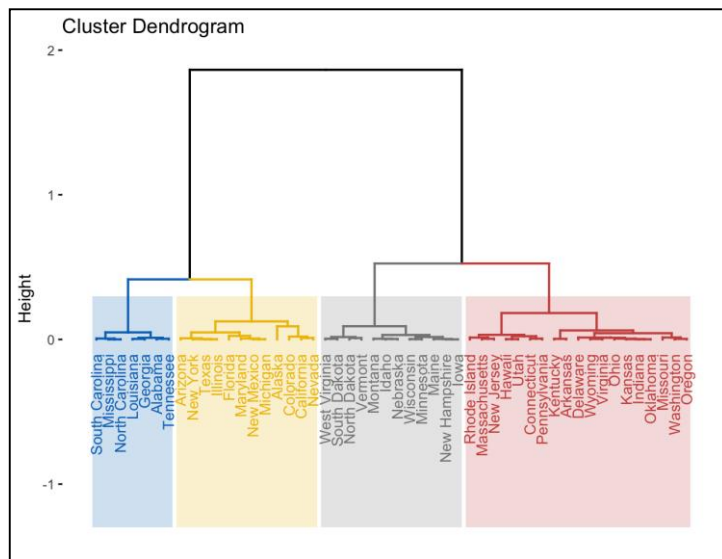


Figure 4: Hierarchical clustering (Dendrogram) adopted from (Kassambara, 2017)

Methodology

Zainol et al. (2017) proposed a framework for Text Analytics of Unstructured Data (TAUD) for discovering significant text patterns, existing in the military text documents. Figure 5 presents the

framework for TAUD, consisting of three main modules: (a) Collection of Text Documents; (b) Keyword Extraction in Military Peacekeeping Documents and (c) Analyzing and Visualizing Selection of Terms. The input to this framework is a collection of text documents acquired from UN Peacekeeping Training Manual. Also, the background of on unstructured data, text analytics and military explicit knowledge in peacekeeping operations were discussed in this work.

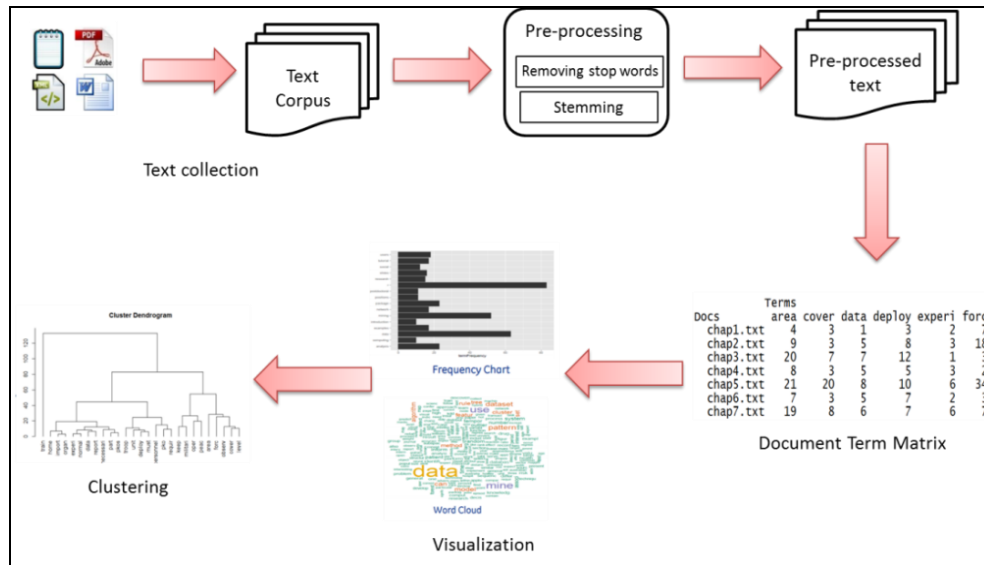


Figure 5: The Framework for Text Analytics in Military Peacekeeping Document (TAUD) adopted from (Zainol et al., 2017).

a. Collection of Text Documents

The first module of TAUD framework deals with the process of document collection from the UN Peacekeeping Training Manual. In this module, the dataset is collected from the United Nations (UN) Peacekeeping Training Manual, downloaded from the UN website ("United Nations Peacekeeping," 2017). The training manual is a fundamental reference for any UN peacekeeping operation. It provides essential knowledge and guidelines on the range of skills required for the troops in the peacekeeping operations. The training guidelines are designed for assisting the military commanders and training officers to prepare and conduct actual military troop's in-country training programs before they can be assigned to a peacekeeping operation. Generally, the manual consists of seven main chapters: (a) introduction - background, aim, restriction, limitation and use, etc. (b) the background of UN peacekeeping, (c) the weapon training's and familiarity with theatre weapons, vehicles and equipment (d) general military training areas (e) training in UN operating techniques, (f) safety measures and precautions, and finally (g) specialized training areas. In this module, the sample of dataset from the PDF file is converted into plain text format using an online converter. This is followed by the categorization of the chapters into seven plain text documents.

b. Keyword Extraction in Military Peacekeeping Documents

This module consists of two main sub-modules: (a) documents pre-processing and (b) keyword extraction (see Figure 6). These documents i.e. seven plain text documents need to be prepared

in an understandable format before they can be used for further text analysis. Technically, the seven chapters are related to the seven documents.

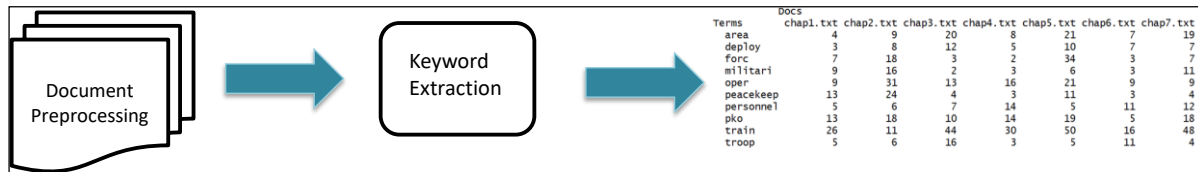


Figure 6. Sub modules in the Keyword Extraction for Military Peacekeeping Documents

The document preprocessing is the most important step in text analysis. This is because the raw dataset is often incomplete, inconsistent, contains many errors, etc. Poor data quality will affect the accuracy of text mining results. The pre-processing helps to remove errors and inconsistencies from the dataset. This process can improve the quality of data and also, the accuracy and effectiveness of document clustering.

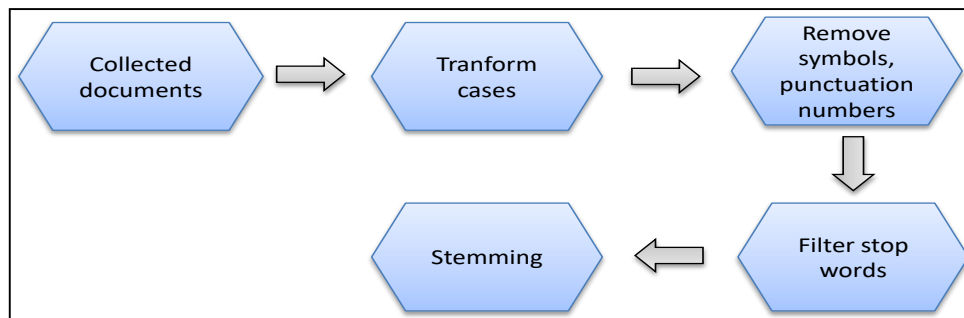


Figure 7. The preprocessing steps of text documents

Figure 7 illustrates the pre-processing steps for text documents, starting from the collection of documents to stemming. As mentioned above, the collected documents need to be prepared for further processing. In the first step of text pre-processing, all text documents (see Figure 7) are converted into lowercase for standardization. The purpose of converting the text documents into a standard case (lower case) is, to ensure that multiple forms of keywords such as ‘PeaceKeeping’, ‘peacekeeping’ and ‘PEACEKEEPING’, are identified as similar. The next step is to remove symbols, punctuation marks, and numbers from the text documents. After that, all keywords will be pre-processed by removing stop words, and extraneous whitespaces. Stop words are common words that are often used in the text document. These include words such as conjunctions (and, but, or, etc.), common verbs (is, was, were, has, etc.), articles (a, the, an, etc.), and qualifiers (yet, however, perhaps, maybe, etc.), etc. After removing the stop words, the remaining words are considered as keywords. The next step is to apply stemming algorithm (e.g., SnowballC) to reduce such related forms of a word into their common base form (Nohuddin et al., 2015). For example, the keywords ‘training’, ‘trained’, and ‘trains’ may result to stem keyword ‘train’. All the cleaned text documents are then compiled and ready for keyword extraction.

After the text documents are pre-processed, all keywords are transformed into term document matrix (TDM). TDM is a matrix that consists of the occurrences of keywords or terms in documents. During the process of keyword extraction, the frequency of each keyword in every document is recorded. For example, the keyword “area” appears 4 times in chapter document 1,

9 times in chapter document 2, 20 times in chapter document 3, etc. The result of keyword extraction from the seven text documents is represented in the form of 2D matrix. The current experiment consists of 1502 keywords in seven (7) documents with 72% sparsity. In other words, it is a 7 x 1502 dimension keyword matrix in which 72% of the rows are recorded with zero. This shows that a large number of keywords will only appear in a few text documents. R provides a support function to remove the sparse terms in tm package. For example, in this experiment, we set 10% of empty space using the following command line: 'tdms <- removeSparseTerms(tdm, 0.1)', where the result is stored in variable 'tdms'. Figure 8 shows the results of removing sparse in term document matrix with 20 keywords in seven (7) text documents with 0% of sparsity.

```
<<TermDocumentMatrix (terms: 20, documents: 7)>>
Non-/sparse entries: 140/0
Sparsity : 0%
Maximal term length: 9
Weighting : term frequency (tf)
Sample :
      Docs
Terms  chap1.txt chap2.txt chap3.txt chap4.txt chap5.txt chap6.txt chap7.txt
area   4          9          20         8          21         7          19
deploy 3          8          12         5          10         7          7
forc   7          18         3          2          34         3          7
militari 9         16         2          3          6          3          11
oper   9          31         13         16         21         9          9
peacekeep 13       24         4          3          11         3          4
personnel 5          6          7          14         5          11         12
pko    13         18         10         14         19         5          18
train  26         11         44         30         50         16         48
troop  5          6          16         3          5          11         4
```

Figure 8. Term document Matrix (TDM) after removing sparse terms (keywords)

As shown in Figure 8, the matrix n x m consists of list of different keywords (terms) as the row and UN Peacekeeping training manual chapters as the column. The matrix dimension is reduced to 20 x 7. In this study, only 20 keywords (terms) are selected from the set of seven (7) documents. Based on TDM, it is very difficult to analyze the relationship or connection between the keywords and chapters. Therefore, document clustering techniques such as K-Means and hierarchical clustering were applied to group similar content of UN Peacekeeping training manual. The result of these techniques will be further discussed in the section of Result and Analysis.

c. Analyzing and Visualizing the Military Peacekeeping Document

The final stage of TAUD framework is the visualization of UN Peacekeeping training manual. In this phase, all cleaned text data is ready for text analytics and visualization into various forms such as frequency graph, word cloud and document clustering. The next step is to calculate the cumulative keyword frequency across seven (7) documents.

```
> freq <- colSums(as.matrix(dtms))
> freq
  area  cover  data  deploy  experi  forc  import  level  militari  observ  oper  organ  peacekeep  personnel
88    47    37    52    23    74    16    39    50    49    108    13    62    60
pko  report  train  troop  unit  use
97    27    225    50    49    25
```

Figure 9. A screenshot for the frequent keywords

As shown in Figure 9, the top five (5) keywords in seven (7) documents are "train" (225) followed by "oper" (108), "pko" (97), "area" (88) and "forc" (74). The keywords count can be visualized in

the form of a graph. Figure 10 presents the visualization of top 20 keywords frequency for seven (7) chapters in the form of histogram chart.

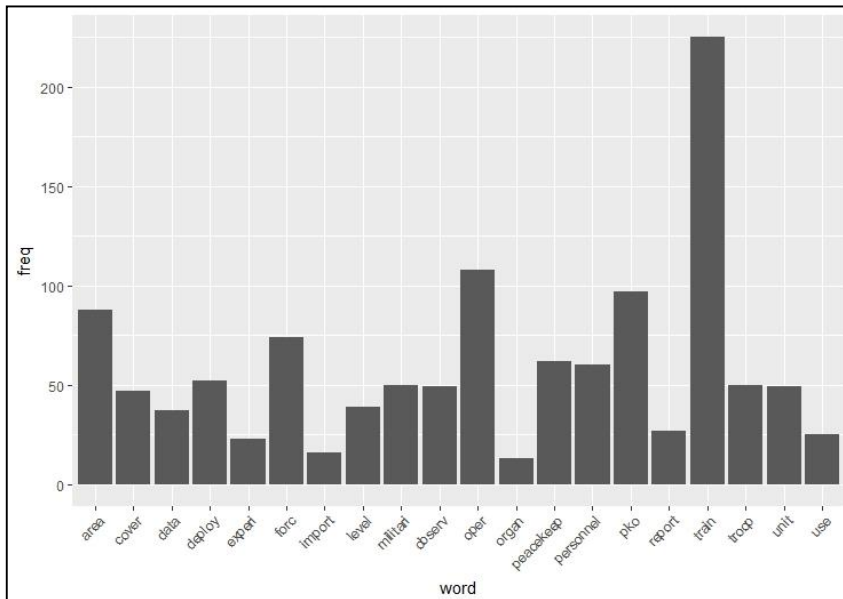


Figure 10. Word frequency histogram with 20 keywords

As shown in Figure 11, the keywords “train”, “oper”, “pko”, “forc” and “area” are among the top 5 frequently used words in text documents which validates that, the seven (7) documents present information that is related to peacekeeping operation and training. In this experiment, we customized the word cloud by setting the maximum number of keyword plotting to hundred (100). It indicates that the size of keywords is corresponding to the frequency of the keyword.



Figure 11: Word cloud with 100 most frequently occurring

Results and Discussion

In this section, the experimental results for the proposed framework of TAUD are presented and discussed. The pre-processed dataset consists of a TDM of seven documents and 20 terms, being

used as input dataset for analyzing the military peacekeeping documents. Figure 12 presents a cluster dendrogram or tree diagram for 20 terms (keywords) in seven documents using hierarchical cluster analysis.

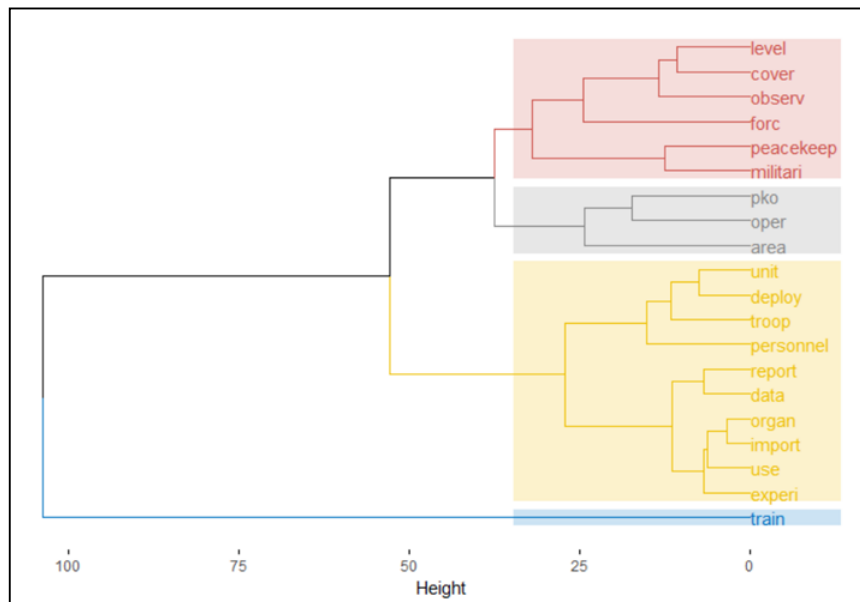


Figure 12. Hierarchical clustering for 20 keywords in 4 main clusters

The distances between clusters are represented in the horizontal axis whereas the keywords are represented in the vertical axis. The distance between each pair of grouped clusters is represented by the height of the horizontal branches. From the bottom of diagram, we can see that the keyword “train” is placed in the first cluster. The keywords “unit”, “deploy”, “troop”, “personnel”, “report”, “data”, “organ”, “import”, “use” and “experi” are grouped together in the second cluster. This shows that there are multiple chapters explaining the importance of organization and unit, experience personnel and troops before they can be deployed in a peacekeeping operation. The third cluster consists of the keywords “pko”, “oper”, and “area”. They are clustered together into one group of chapters focusing on the operation’s area for the PKO (Peacekeeping Operation). The fourth cluster consists of keywords “peacekeep”, “militari”, “forc”, “observ”, “cover” and “level”. This shows that there are chapters describing the military peacekeeping observers and forces cover at all levels. This result can also be visualized in the of tree structure (see Figures 13 and 14).

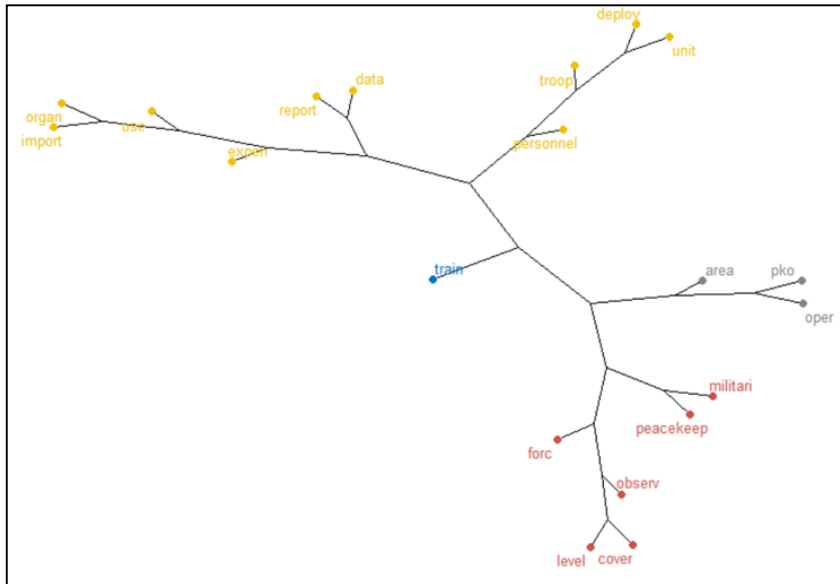


Figure 13. A sample of phylogenetic tree dendrogram for 20 keywords in 4 main clusters

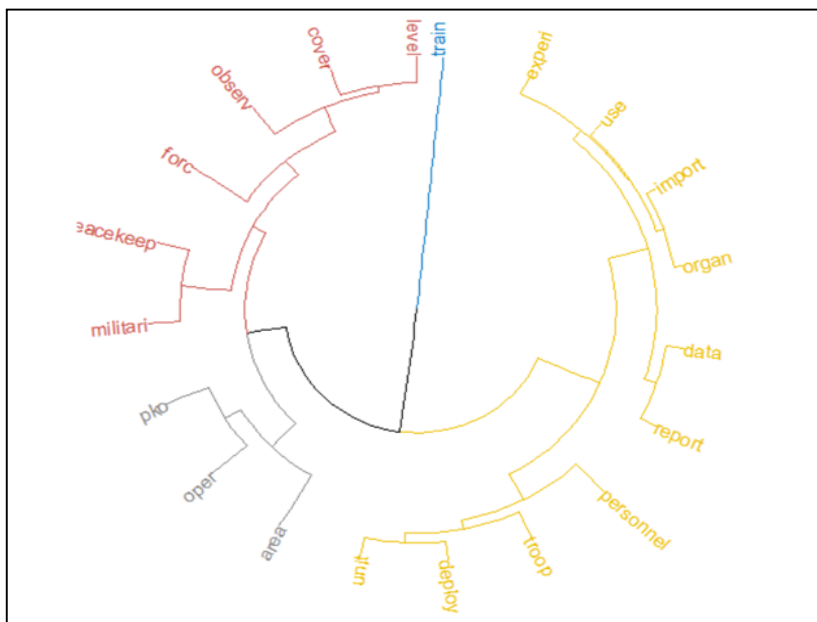


Figure 14. A sample of circular or "fan" dendrogram for 20 keywords in 4 main clusters

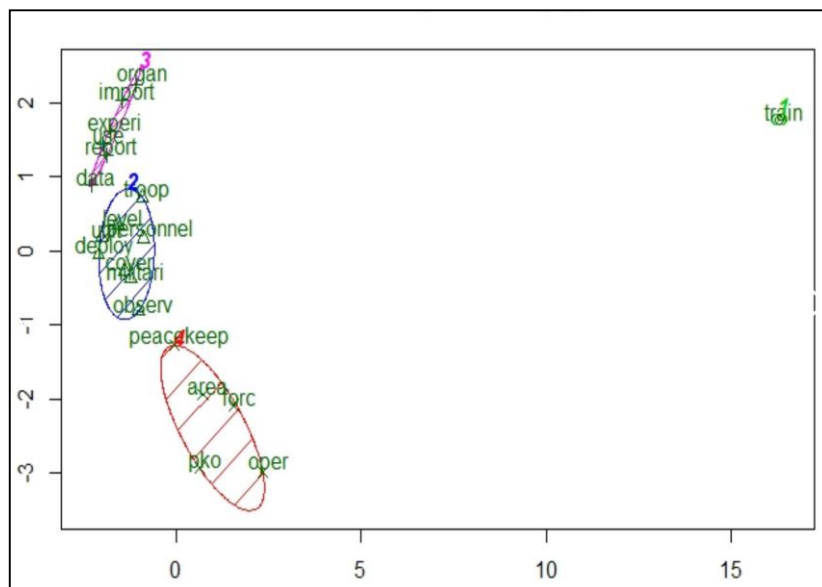


Figure 15. K-Means clustering for 20 keywords with 4 clusters

One of the simplest unsupervised learning algorithms is K-Means. It can be applied to unlabeled data. In this experiment, the number of cluster K is set to 4. Figure 15 presents the K-means clustering for 20 terms in seven (7) documents. On the right hand side of the diagram, we can clearly see the keyword “train” is placed in the first cluster. However, the clusters 2,3,4 are placed on the left hand side of the diagram. The second cluster consists of the keywords “troop”, “personnel”, “deploy”, “cover”, “observ”, “level”, “unit” and “militari”. The third cluster comprises of keywords “data”, “report”, “experi”, “import”, “home” and “organ”. This relates to the chapters on military peacekeeping observers and forces cover at all levels. Finally, the fourth cluster consists of keywords “oper”, “pko”, “forc”, “area” and “peacekeep”.

The results highlight the most frequently words and important terms in the seven (7) documents related to peacekeeping operation and training. The findings of this study (i.e. visualization from frequency histogram, word cloud, hierarchical clustering and K-means clustering) allow the military personnel to understand and access the military knowledge which they had learnt and gathered during the training programs before they can be deployed into any peacekeeping mission.

Conclusion

The paper presents discovery and visualization of essential information and significant text patterns contained within UN peacekeeping manuals, etc. This was achieved using Text Analytics of Unstructured Data (TAUD) framework using the hierarchical cluster analysis and K-Means Algorithm. The documents were obtained from official UN peacekeeping sources. The findings highlight the most frequently used words and terms in the documents related to peacekeeping missions. The results are indicative that the methodology has been successful in extracting and focusing on areas that are essential for the troops before and while being deployed on a peacekeeping mission. The discovery and ease of availability of such information can be a critical factor for the peacekeeping forces to cat efficiently during their deployment, especially in times of crisis situations.

Acknowledgement

The authors would like to thank Universiti Pertahanan Nasional Malaysia (UPNM) and Ministry of Higher Education under NRGs/2013/UPNM/PK/P3 for sponsoring this publication.

Corresponding Author

Zuraini Zainol^{1*}

Department of Computer Science,
Faculty of Defence and Science Technology,
National Defence University of Malaysia,
Sungai Besi Camp,
57000 Kuala Lumpur,
Malaysia

References

- Alkhred, F., Nohuddin, P. N. E., & Zainol, Z. (2018). Sharing Explicit Knowledge: Designing a Peacekeeping Operation Databank. *Advanced Science Letters*, 24(2), 956-960.
- Amin, M. N. M., Nohuddin, P. N., & Zainol, Z. (2014). *Trend cluster analysis using self organizing maps*. Paper presented at the Information and Communication Technologies (WICT), 2014 Fourth World Congress on.
- Chakraborty, G., Pagolu, M., & Garla, S. (2014). *Text mining and analysis: practical methods, examples, and case studies using SAS: SAS Institute*.
- Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know: Harvard Business Press*.
- Desai, A. (2015). A Review on Knowledge Discovery using Text Classification Techniques in Text Mining. *International Journal of Computer Applications*, 111(6).
- Esmine, A. A. A., Coelho, R. A., & Matwin, S. (2015). A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. *Artificial Intelligence Review*, 44(1), 23-45. doi:10.1007/s10462-013-9400-4
- Feldman, R., & Dagan, I. (1995). *Knowledge Discovery in Textual Databases (KDT)*. Paper presented at the KDD, Canada.
- Galbraith, J. R. (2014). Organizational design challenges resulting from big data.
- Godbole, S., Bhattacharya, I., Gupta, A., & Verma, A. (2010). *Building re-usable dictionary repositories for real-world text mining*. Paper presented at the Proceedings of the 19th ACM international conference on Information and knowledge management, Toronto, ON, Canada.
- Hatamlou, A. (2013). Black hole: A new heuristic optimization approach for data clustering. *Information Sciences*, 222(Supplement C), 175-184. doi:<https://doi.org/10.1016/j.ins.2012.08.023>
- Hultman, L., Kathman, J., & Shannon, M. (2013). United Nations peacekeeping and civilian protection in civil war. *American Journal of Political Science*, 57(4), 875-891.
- Jensi, R., & Jiji, D. G. W. (2014). A survey on optimization approaches to text document clustering. *arXiv preprint arXiv:1401.2229*.
- Karol, S., & Mangat, V. (2013). Evaluation of text document clustering approach based on particle swarm optimization. *Open Computer Science*, 3(2), 69-90.
- Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning* (Vol. 1): STHDA.

- Latif, S., & Wood, M. M. (2008). *Text Pre-processing for Document Clustering*. Paper presented at the 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008, London, UK.
- Lau, E. K. W. (2016). *Deploying the Measurement of Knowledge Management using the Balanced Scorecard*. Paper presented at the Proceedings of the The 11th International Knowledge Management in Organizations Conference on The changing face of Knowledge Management Impacting Society, Hagen, Germany.
- Massey, G. (2015). In Context: Extracting Relevance from Unstructured Medical Data: December.
- Mukherjee, S. K., & Bandyopadhyay, S. (2015, 9-11 July 2015). *Clustering to determine predictive model for news reports analysis and econometric modeling*. Paper presented at the 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS).
- Nanda, S. J., & Panda, G. (2014). A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm and Evolutionary computation*, 16, 1-18.
- Negm, N., Amin, M., Elkafrawy, P., & Salem, A. B. M. (2013). Investigate the Performance of Document Clustering Approach Based on Association Rules Mining. *Editorial Preface*, 4(8).
- Nohuddin, P. N. E., & Zainol, Z. (2014). *Context-based Keyword Pattern Cluster Analysis technique on tacit knowledge of military centre of expertise (COE): A preliminary study*. Paper presented at the Computational Science and Technology (ICCST), 2014 International Conference on.
- Nohuddin, P. N. E., Zainol, Z., & Amir, M. A. U. (2018). Trend Cluster Analysis of Wave Data for Renewable Energy. *Advanced Science Letters*, 24(2), 951-955. doi:10.1166/asl.2018.10665
- Nohuddin, P. N. E., Zainol, Z., Chao, F. C., James, M. T., & Nordin, A. (2015). Keyword based Clustering Technique for Collections of Hadith Chapters. *International Journal on Islamic Applications in Computer Science And Technologies – IJASAT*, 4(3), 11-18.
- Rathi, S., Shirgaonkar, S., & Dhote, C. A. (2010). *Text mining: applied for bomb blast detection*. Paper presented at the Proceedings of the International Conference and Workshop on Emerging Trends in Technology, Mumbai, Maharashtra, India.
- Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, 344(6191), 1492-1496.
- Smith, E. A. (2001). The role of tacit and explicit knowledge in the workplace. *Journal of Knowledge Management*, 5(4), 311-321.
- Sumathi, S., & Sivanandam, S. (2006). *Introduction to data mining and its applications* (Vol. 29): Springer.
- Tyagi, N. K., Solanki, A., & Tyagi, S. (2010). An algorithmic approach to data preprocessing in web usage mining. *International journal of information technology and knowledge management*, 2(2), 279-283.
- United Nation. (2017). Retrieved from <http://www.un.org/en/sections/about-un/overview/>
- United Nation Peacekeeping. (2017). Retrieved from <http://www.un.org/en/peacekeeping/>
- Xie, P., & Xing, E. P. (2013). Integrating document clustering and topic modeling. *arXiv preprint arXiv:1309.6874*.
- Yusof, W. S. E. Y. W., Zakaria, O., & Zainol, Z. (2016, 16-17 May 2016). *Establishing of knowledge based framework for situational awareness using Nonaka's and Endsley's models*. Paper presented at the 2016 International Conference on Information and Communication Technology (ICICTM).

Zainol, Z., Nohuddin, P. N. E., Mohd, T. A. T., & Zakaria, O. (2017). *Text Analytics of Unstructured Textual Data: A Study on Military Peacekeeping Document using R Text Mining Package* Paper presented at the the 6th International Conference on Computing & Informatics.

Zhao, Y., & Karypis, G. (2002). *Evaluation of hierarchical clustering algorithms for document datasets*. Paper presented at the Proceedings of the eleventh international conference on Information and knowledge management.