Semantic association rule mining in text using domain ontology

Ibukun Afolabi*, Olaperi Sowunmi and Olawande Daramola

Department of Computer and Information Sciences,
Covenant University,
Ota, Nigeria
Email: ibukun.fatudimu@covenantuniversity.edu.ng
Email: olaperi.sowunmi@covenantuniversity.edu.ng
Email: olawande.daramola@covenantuniversity.edu.ng
*Corresponding author

Abstract: This paper reports a procedure for ontology-based association rule mining for knowledge extraction from text. Association rule mining (ARM) algorithms have the limitations of generating many non-interesting rules, huge number of discovered rules, and low algorithm performance. This research demonstrates a procedure for improving the performance of ARM in text mining by using domain ontology. A study context of Nigerian politics using news text from a Nigerian online newspaper was selected, and a methodology that combined natural language processing, ontology-based keywords extraction, and the modified Generating Association Rules based on Weighting (GARW) scheme was applied. The result revealed significant rule reduction in the number of generated rules, and produced rules, which are more semantically related to the problem context when compared to when ARM approaches that are not ontology-based is used. The study shows that domain ontology can improve the performance of ARM algorithms when dealing with unstructured textual data.

Keywords: domain ontology; text mining; political science; association rule mining; Nigeria.


Biographical notes: Ibukun Afolabi has a PhD in Computer Science and is a Lecturer in the Department of Computer and Information Sciences of Covenant University Nigeria. Her research interests include text mining, data mining intelligent systems and customer relationship management.

Olaperi Sowunmi is a PhD student of Computer Science. Her research interests lie in the semantic web, intelligent software systems and personalised information retrieval.

Olawande Daramola is an Associate Professor of Computer Science in the Department of Computer and Information Sciences of Covenant University Nigeria. His research interests include the semantic web, intelligent systems, and applied software engineering.

1 Introduction

Currently, large amounts of politics-oriented text are now available online. This includes both official documents, such as the full text of laws, the proceedings of legislative bodies, and unofficial documents, such as postings on weblogs (blogs) devoted to politics and online newspapers (Thomas et al., 2006). In addition to this, social media is also a platform to retrieve very large volume of data related to politics. Also, there are many available text mining tools that can help to manage this outbreak of textual information. Many of these tools are derived from earlier works in Information Retrieval (IR), Natural Language Processing (NLP), Statistics, Artificial intelligence (AI), Information Theory, and Data Mining (Konchady, 2006).

Text mining, which is the focus of this research, combines techniques from areas such as information retrieval, NLP, information extraction and data mining in order to process
Specifically, the aim of this work is to achieve significant improvement of ARM for discovering knowledge in textual databases (Ah-Hwee, 2006), allows creation of a technology that combines human linguistic capabilities with the speed and accuracy of a computer. Text mining employs technology to analyse more detailed information in the content of each document, and extracts interesting information that can be provided only by multiple documents viewed as whole, such as trends and significant features that may be a trigger useful actions and decision-making (Nasukawa and Nagano, 2006).

Several techniques have been proposed for text mining. These include conceptual structure, Association Rule Mining (ARM), clustering, episode rule mining, decision trees, and rule induction methods. In addition, Information Retrieval (IR) techniques have widely used the ‘bag-of-words’ model (Baeza-Yates and Ribeiro-Neto, 1999; Raymond and Un Yong, 2005) for tasks such as document matching, ranking, and clustering.

Similarly, several authors have used ARM as a text mining technique (Saggar et al., 2004; Mutter et al., 2001; Mangla and Akhare, 2015; Tank, 2014; Mary and Malarvizhi, 2012; Al-Zawaidah and Jbara, 2011). However, according to Moreno et al. (2004), the main drawbacks of association rule algorithms are obtaining non-interesting rules, huge number of discovered rules, and low algorithm performance. As a response to these observations, some efforts have been made to improve the performance of ARM algorithms by using ontologies. Notable efforts in this area include Bellandi et al. (2007), Marinica et al. (2010), Marinica et al. (2008), Liu et al. (2013), and Martin et al. (2011). Although these approaches have used ontology to improve ARM, they focussed mostly on structured data (see Table 1). In this work, we present a framework that enables ontology-based enhancement of ARM in text mining by using unstructured data. This is applied to data obtained from the Nigerian political and historical context.

To do this, we adopted a methodology that entails gathering unstructured textual data from political and historical news sources, semi-automatic creation of domain ontology, ontology refinement, and knowledge distillation from gathered data and domain ontology using ARM. This work therefore demonstrates an instance of ontology-based improvement of ARM for discovering knowledge in unstructured data in a way that addresses the main drawbacks hitherto observed in the application of ARM. Specifically, the aim of this work is to achieve significant rule reduction, and generation of relevant rules, which can contribute semantically to the problem being solved.

The rest of this paper is described as follows. Section 2 presents literature review, while Section 3 describes the methodology adopted for this work. In Section 4, we present the results and discussion, while the paper is concluded in Section 5 with a brief note and future work.

### Table 1 Summary of efforts on ontology-based improvement of ARM

<table>
<thead>
<tr>
<th>S/N</th>
<th>Author</th>
<th>Use of ontology in ARM</th>
<th>Type of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bellandi et al., 2007</td>
<td>Improving rules generated from association rule mining</td>
<td>Structured: Market-basket data stored in a database</td>
</tr>
<tr>
<td>2</td>
<td>Marinica et al., 2008; Marinica and Guillet, 2010</td>
<td>Post-processing of the association rule mining results using an ontology for the consistency checking</td>
<td>Structured: Database</td>
</tr>
<tr>
<td>3</td>
<td>Liu et al., 2013</td>
<td>Ontology and hypergraph to discover latent association rules in the data</td>
<td>Structured: Electronic health records in database transformed into RDF</td>
</tr>
<tr>
<td>4</td>
<td>Martin et al., 2011</td>
<td>Combined financial domain ontological model with association rule mining algorithm to develop Z-score model</td>
<td>Structured: Corporate financial database</td>
</tr>
</tbody>
</table>

### 2 Literature review

In this section, we have reviewed efforts in the literature that cover relevant aspects such as use of ARM in text mining, optimisation and improvement of ARM, and ontology-based improvement of ARM.

Several instances where text mining systems were based on association rule have been reported in the literature. These include the following: Feldman et al. (1998) developed a Document Explorer, which is a tool that implements text mining at the term level. Earlier works on mining association rules from text have explored the use of manually assigned keywords. The keywords were used as features for generation of association rules (Feldman and Dagan, 1995; Feldman and Hirsh, 1996). The drawbacks of approaches that use manually assigned keywords are (1) it is time-consuming to manually assign the keywords; (2) the keywords are fixed (i.e. they do not change over time or vary based on a particular user); (3) if the keywords are manually assigned, they are subject to discrepancy; and (4) the textual resources are constrained to only those that have keywords. Several other researchers (Ahonen et al., 1997; Ahonen et al., 1998; Mannila et al., 1997) applied existing data mining techniques to discover episode rules from texts, where episode rule mining is used for language analysis because it preserves the sequential structure of terms in a text document.

So many approaches that seek to improve ARM generation have been reported in the literature. Saggar et al. (2004) optimised the rules generated by ARM using Genetic Algorithms. Mutter et al. (2001) used classification approach to evaluate the quality of a set of association rules generated...
by a confidence-based miner. Mangla and Akhare (2015) proposed an approach for filtering discovered rules. This approach uses user and domain expert knowledge. It integrates user knowledge in the post-processing with ontologies connected to data. Tank (2014) proposed an algorithm to improve Apriori algorithm by decreasing pruning operations. The approach generates the candidate 2-itemsets by the \textit{apriori-gen} operation. It does this by adopting the tag-counting method to calculate support quickly. Mary and Malavizhi (2012) combined weighted Apriori and dynamic programming in order to enhances website effectiveness, increase user browsing knowledge, improve prediction accuracy and decreases computational complexities. Al-Zawaidah and Jhara (2011) presented a novel ARM approach that can efficiently discover the association rules in large databases. The approach is derived from Apriori approach with features added to improve data mining performance.

So far, some efforts aimed at enhancing ARM by using ontologies have also been reported. Bellandi et al. (2007) showed that rules generated from ARM systems can be enhanced by using ontologies. The work presented an ontology-based ARM method, which queries the ontology to filter the instances used in the ARM process. Also, in Marinica and Guillet (2010) and Marinica et al. (2008) post-processing of the ARM results by using an ontology for consistency checking was presented. Liu et al. (2013) made a proposal to apply ontology and hypergraph to discover latent association rules in the data. Martin et al. (2011) combined financial domain ontological model with ARM algorithm to develop Z-score model, which is a new business intelligence model in order to predict bankruptcy.

From the literature, we see that although ontology has been used to improve ARM in some researches, the applications have been based mostly on structured data, not many efforts have applied ontology enhancement to ARM in the context of unstructured text (see Table 1). Also, in situations where text mining has been improved with ontology (Jiang et al., 2013; Bloehdorn et al., 2005), it has not been particularly directed towards generating inferences that pertain to the political domain using association rule, as we have attempted in this study.

### 3 Methodology

This section describes the adopted methodology used for this research. It reports the data gathering process for the experiment, the ontology-based keyword extraction process, which includes the ontology creation activity, and finally the knowledge distillation process. The architecture describing the methodology is presented in Figure 2.

#### 3.1 Data gathering

This is the first step, which has to do with gathering textual corpus or document in the application area, which is Nigerian political domain. The documents gathered for this research were obtained from one of the foremost newspaper vendors in Nigeria called Vanguard newspaper online. The textual data, which is available at: http://www.vanguardngr.com/ category/columns/people-politics/, is an archive of people and politics-related news over a long period of time. Association rules were extracted from 618 recent news from the Nigerian Vanguard website. The collection of the 618 documents (corpus) is 2.75 MB in size and each document contained on average 1000 single words. After gathering the textual data, it was pre-processed using the following steps:

1. **Filtration**: This process has to do with removing the unimportant words from documents’ content. Such unimportant words include: articles, pronouns, determiners, prepositions and conjunctions, common adverbs and non-informative verbs. As a result of this process, more important or highly relevant words were singled out.

2. **Stemming**: This removes a word’s prefixes and suffixes (such as unifying both infection and infections to infection). For this experiment, the stemming algorithm referenced the WordNet Dictionary (https://wordnet.princeton.edu/wordnet/download/current-version/) in order to have a level of semantic analysis introduced into the pre-processing.

3. **Weight Assignment to Terms**: Finally, the weighting scheme TF-IDF (Term Frequency, Inverse Document Frequency) was used to assign higher weights to syntactically distinguished terms in a document (Hany et al., 2007). The filtration, stemming and weight assignment to terms were all carried out using the rapidminer toolkit (https://rapidminer.com).

#### 3.2 Ontology-based keyword extraction

This first activity of this phase of the adopted methodology is the creation of domain ontology of for Nigerian political history. The process is described as follows.

The core idea of the approach to developing the Nigerian Political History Ontology (NPHO) is based on the prescriptions in Boyce and Pahl (2007) and Noy and McGuinness (2000). Relevant keywords and key phrases were extracted from textual information using the typical information extraction process (Noy and McGuinness, 2000). The ontology was built using the following steps: (1) extract important terms from text, (2) define the concept taxonomies, relations, attributes, instances, axioms and functions, and (3) create an application interface through which the ontology can be accessed for the purpose of querying (Afolabi et al., 2014).

The concepts that are included in the Nigerian History Ontology (NPHO) are those that are considered relevant from the political, historical, and cultural perspectives. To implement the NPHO, the Protégé ontology editor was used. Apart from using ontology for keyword extraction, a Graphic User Interface (GUI) that enables a user to query the ontology from a computer desktop was developed using the Java NetBeans and Jena ontology API. The following diagram summarises the concepts contained in the NPHO.
Figure 1 is the Protégé implementation of the main concept in the NPHO. The NPHO ontology contains the following constructs:

- **rdfs:label**: Used to store the general or popularly known name used to refer to the concept or individual.
- **rdfs:DefinedBy**: Used to store the definition of the term.
- **rdfs:Comment**: Captures the further explanation on the term or individual.
- **owl:equivalentClass**: Used to equate similar concepts in meaning, for example, `Ethnic_groups` is equated to `Tribe`.
- **owl:sameAs**: Used to equate similar individuals.
- **rdfs:subClassOf**: Used to break down general concepts to the type of classes that make them up.
- **owl:disjointWith**: Used to handle complex classes.

The NPHO contains relevant Object properties, such as HasSynonym, HasAsExample, isPredecessorOf, isSuccessorOf, SpatiallyInfers, isRelatedTo, BelongsTo, FriendOf, PartOf, and Has all of which define relationships between concepts in the ontology.

- **HasSynonym**: This is used to indicate that a single concept may have one or more names. It uses a 1:N relationship.
- **HasAsExample**: This is used to indicate that the concept being linked to is an instance or example of the concept from which the link emanates.

It also manages historic changes using the following object properties, isPredecessorOf and isSuccessorOf (Iglesias-Sucasas et al., 2009). The isPredecessorOf and isSuccessorOf are symmetrically inverse properties that are used to identify which Person, ethnic group or kingdom were the predecessor/successor of another. For example, President GoodLuck Jonathan is the successor of President Musa Yaradua, and therefore the ontology specifies that: Goodluck_Jonathan isSuccessorOf Musa_Yaradua, which also connotes that Musa_Yaradua isPredecessorOf Goodluck_Jonathan.

The NPHO contains 46 concepts, ten object properties, 23 subconcept axioms, five equivalent concept axioms, six disjoint concept axioms, ten object property range axioms, ten object property domain axioms, one object property inverse axiom and 817 individuals.

An application that shortlists the extracted keywords in step (i) using the NPHO described above was developed using the Java programming language.
3.3 Knowledge distillation

ARM, based on the modified Generating Association Rules based on Weighting scheme (GARW) (Hany et al., 2007), used for text mining is one of the most important techniques in Data Mining. The problem of ARM deals with how to discover association rules that have support and confidence greater than the user-specified minimum support and minimum confidence. It is intended to capture dependency among items in the database.

The support of an item set is the fraction of transactions in the database that contain all the items in the database.

\[
\text{Support}(W_1W_2) = \frac{\text{SupportCount}(W_1W_2)}{\text{Total NumberOf Transactions}}
\]

The confidence of rule (association rule) \( W_1 \rightarrow W_2 \) can be defined as the proportion of those transactions containing \( W_1 \) that also contain \( W_2 \).

\[
\text{Confidence}(W_1W_2) = \frac{\text{Support}(W_1W_2)}{\text{Support}(W_1)}
\]

The algorithm for generating association rules based on the weighting scheme (GARW) is given in Figure 3.

The knowledge distillation phase was carried out using the rapidminer.

4 Results and discussions

The result discovered in this research provides solution to the first part of the aim, which is to provide significant rule reduction, particularly as regards generating rules, which contribute semantically to the problem being solved. This is presented in Subsection 4.1.

4.1 Ontology-based improvement

Comparing the rules generated from applying the ARM algorithm on both the ontology extracted keyword file and the non-ontology extracted keyword files, we discovered that there was a significant rule reduction in the ontology extracted keyword method. It is also of interest to note that the generated rules are not just reduced, but have been reduced to politics-related inferences. The number of rules generated at different support thresholds is presented in Table 2.

![Figure 4](image-url)
constant confidence of 80%. From Figure 4, it can be seen that there is reduction in the number of rules generated by the ontology-based extraction (167) compared to the non-ontology-based method (82634). This connotes over 80% reduction in number of generated rules, which is significant because it helps to address the problem of huge number of discovered rules in ARM.

The other part of our objective, which is to discover knowledge or trends, which are semantically related to the area of application, i.e. Nigeria politics is presented in Subsection 4.2.

Table 3 Extraction of some of the generated rules

<table>
<thead>
<tr>
<th>Ontology extracted keyword method</th>
<th>Non-ontology extracted keyword method</th>
</tr>
</thead>
<tbody>
<tr>
<td>state,federal,power</td>
<td>Read, come, state</td>
</tr>
<tr>
<td>political,leader</td>
<td>Read, Nigerian, year</td>
</tr>
<tr>
<td>power,governor</td>
<td>People, read, year</td>
</tr>
<tr>
<td>state,political,jonathan</td>
<td>State, take</td>
</tr>
<tr>
<td>state, election</td>
<td>People, Nigerian, state</td>
</tr>
<tr>
<td>nigeria, government,power</td>
<td>Nigerian, come, take</td>
</tr>
<tr>
<td>nigeria,political,general</td>
<td>Read, Nigeria, come, take</td>
</tr>
<tr>
<td>party,chief,election</td>
<td>Read, year</td>
</tr>
<tr>
<td>nigeria,general,north</td>
<td>People, read</td>
</tr>
<tr>
<td>country,north,leader</td>
<td>Come, say</td>
</tr>
<tr>
<td>state,power,federal,jonathan</td>
<td>People, read, Nigeria, state</td>
</tr>
<tr>
<td>nigeria,political,federal,military</td>
<td>Nigeria, president, year</td>
</tr>
<tr>
<td>nigeria,political,power,chief</td>
<td>Read, come, say</td>
</tr>
<tr>
<td>nigerian,government,change</td>
<td>People, Nigeria, state, say</td>
</tr>
<tr>
<td>nigeria,country,security</td>
<td>President, people</td>
</tr>
<tr>
<td>nigeria,government,south</td>
<td>Read, president</td>
</tr>
<tr>
<td>nigeria,government,good</td>
<td>People, read, president</td>
</tr>
<tr>
<td>state, leader, group</td>
<td>People, read, Nigeria, president</td>
</tr>
<tr>
<td>military,political</td>
<td>People, read, Nigeria, come</td>
</tr>
</tbody>
</table>

4.2 Semantic relatedness to the application area

Finding association rules in text documents can be useful in a number of contexts such as investigation and general understanding of events in the real world. Therefore, relations between features in the textual information gathered, as presented in Table 3, are strictly related to the application area which is Nigerian politics, thereby reducing ambiguity in the interpretation of generated knowledge. It also gives the user or reader of the news the useful inferences about Nigerian politics without having to read through the whole 618 documents. The inferences obtained from the research carried out will provide the necessary foundation and direction for further investigation on Nigerian politics. Examples of the keyword and phrases that were generated for the corpus used for our experiment include: president, Goodluck Jonathan, Nigeria, leader, power, federal government. Table 3 helps to further present some of the rules generated in the ontology-based and non-ontology-based method so as to reveal the difference in the outputs of the two methods. Though the non-ontology-based terms and rules generated contains some few political terms and rules, the disadvantage is that these rules have been hidden in the huge volume of non-political-related terms and rules. The ontology-based terms and rules generated therefore give a better result semantically, as regards politics in Nigeria, which also addresses the problem of obtaining non-interesting rules in ARM. The discovered rules (inferences) also help to validate that indeed the political atmosphere in Nigeria is very controversial, particularly revealing issues related to political leadership and corruption. The results will therefore provide political awareness that can lead to appropriate decision-making.

5 Conclusion and future work

In conclusion, this paper has presented a text mining technique for semi-automatically extracting association rules from a collection of documents based on domain ontology. Also, the inferences obtained from the research carried out will provide the necessary foundation and direction for further investigation about Nigerian politics. The reduction in rules produced by the adopted methodology confirmed that ontology-based approaches can be used to improve the efficiency of ARM text mining by ensuring that reduced numbers rules are generated, which is better for decision-makers.

For future research, there is need to introduce other methods of semantic pre-processing of text such as Latent Dirichlet Allocation (LDA), word embedding and so on. In addition, there can be practical applications of the discovered association rules and using extracted concepts and key phrases as an input to the ARM rather than the stemmed words (removed stops words). These techniques could create concepts that may be in between words and ontology thereby producing association rules that will reduce in number but increase the semantic relatedness. Finally, since the system is ontology dependent, the quality of rules depends on the quality of the ontology, therefore, for future research, the ontology could be expanded and also integrated with other existing related ontologies so as capture more concepts.

References


