Implementation of Decision Tree Algorithm to Classify Knowledge Quality in a Knowledge Intensive System

Casper Kaun1*, N.Z Jhanjhi2, Wei Wei Goh3, Sanath Sukumaran4

1 2 3 School of Computing and IT, Taylor’s University, Subang Jaya, Selangor, Malaysia
4 Agile Management Consultancy, Malaysia

Abstract. Knowledge is an important asset for an organisation as it facilitates organisational growth. To facilitate knowledge creation and sharing, this is where a knowledge-intensive system is required. One key area that hinders the effective use of knowledge-intensive systems in an organisation is the lack of knowledge quality. This causes the system to be underutilised, and as a result, knowledge will not be captured or shared effectively. Recent KM findings identified that machine learning could be beneficial to knowledge management. A literature review was conducted to identify knowledge of quality attributes and machine learning algorithms. From the findings, it was identified that the decision tree algorithm has a strong potential at classifying knowledge quality. An experiment was then devised to identify the training model required and measure its effectiveness using a pilot test. This involved using a knowledge-intensive system and mapping its variables to the respective knowledge quality attributes. From the experimentation result, the training model is then devised before implemented in a pilot test. The pilot test involved collecting knowledge using the same knowledge-intensive system before running the training model. From the results, it was identified that the decision tree could classify knowledge quality though the results yielded four different outputs at classifying knowledge quality. It was concluded that machine learning is beneficial in the area of knowledge management.

1. Introduction

Knowledge management is becoming relevant and important for an organisation to survive or thrive in the world today. Knowledge is an important asset as it helps organisational growth and boost innovation [1]. The current issue today is that the lack of understanding of knowledge quality attributes has led to the development of ineffective knowledge-intensive systems. Without an effective knowledge-intensive system, knowledge could not be shared effectively [2-5]. Furthermore, knowledge quality itself is integral in the success of a knowledge-intensive system [6-12]. The usage of knowledge-intensive systems is affected by the quality of knowledge present in the system [13-15]. One area of knowledge management could benefit from is machine learning. It was identified that machine learning could be beneficial to knowledge management, and its benefits must be looked at [16, 17]. Machine learning algorithms can be used to further enhance a knowledge-intensive system [18]. This is because machine learning is applied in a multitude of tasks [19-23].

1 Corresponding author: caspergihessimon@sd.taylors.edu.my

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).
Hence, the identified research gap is implementing machine learning algorithms to classify knowledge quality. We devise three research questions which are:

- What are the key attributes required to promote knowledge quality?
- How to improve the quality and process flow in the knowledge creation process?
- How to validate the performance of the decision tree algorithm?

The research objectives are first, to identify key attributes to promote knowledge quality, which is mapped to the first research question. Secondly, to improve quality and process flow in knowledge creation by developing a specific machine learning model that is mapped to the second research question. Lastly, to validate the machine learning model for the key attributes in contextual knowledge mapped to the third research question.

The research significance as follows:

- Implementation of decision tree algorithm to classify knowledge quality in a knowledge-intensive system
- The impact of machine learning on knowledge-intensive systems.
- Potential usages/implementation of machine learning algorithms in knowledge management.

The scope of the article focuses on knowledge management and decision tree algorithm. This article aims to show how a decision tree can be used to classify knowledge quality but more importantly, the benefits of machine learning in knowledge management both from a theoretical and practical perspective. The rest of the articles are as follows. Section 2 provides a background in knowledge management, knowledge quality, and machine learning. Section 3 discusses the proposed machine learning algorithm. Section 4 discusses the methodology used to implement machine learning in knowledge management. Section 5 mentions the research results and a discussion of the results, including theoretical and practical implications. The article is concluded in section 6 with research limitations and future research.

2. Literature Review

This section provides an overview of knowledge management and machine learning, including the implementation of machine learning into knowledge management. The aim is to provide a detailed but brief explanation on knowledge management and machine learning.

2.1. Knowledge management

Knowledge management focuses at managing knowledge effectively within an organisation and treating knowledge as an organisational asset [24-31]. Knowledge management is broken down to: systematic analysis, planning, acquisition, creation, development, storage and use of knowledge [32]. Knowledge Management comprises of people, process, technology and governance [33]. A knowledge intensive system is a form of information system that facilitates the implementation of KM within an organisation [9, 29, 34-36]. Knowledge itself is a key component in a knowledge intensive system [30, 37-40, 58]. The implementation of knowledge intensive systems plays an integral part in an organisation success [2, 3, 41, 42, 43, 58]. In essence, an effective knowledge intensive system which utilises the organisation’s knowledge capital is crucial for organisation’s to stay competitive in current market [1, 2, 5, 44, 45, 58].

2.2. Knowledge quality

Knowledge quality is defined as created knowledge which is relevant to knowledge workers and is valuable in content [8, 38]. Knowledge quality is a significant factor when it comes to the success of knowledge intensive systems [6-12]. Knowledge quality has its own attributes and are as follows; intrinsic knowledge quality, contextual knowledge quality, actionable knowledge quality and accessibility knowledge quality [8, 38, 46].
2.2. Knowledge quality

Knowledge quality is defined as created knowledge which is relevant to knowledge workers and is crucial for an organisation to stay competitive in the current market. An effective knowledge intensive system which utilises the organisation's knowledge capital is integral to the success of an organisation. Knowledge management focuses on managing knowledge effectively within an organisation and treating knowledge as an organisational asset. Knowledge management is broken down to: systematic analysis, planning, acquisition, creation, development, storage and use of knowledge.

2.3. Machine Learning

Machine learning is defined as identifying patterns using learned data when interpreting unknown input. Machine learning is divided into supervised and unsupervised learning. Supervised learning focuses at finding or predicting patterns in a dataset and the algorithms are categorised as either classification or regression. Unsupervised learning focuses on identifying patterns in a dataset without known experience or samples. Common supervised learning algorithms are Artificial Neural Network, Decision Tree, Linear Regression, Logistic Regression, K-Nearest Neighbour, Naïve Bayes, Random Forest and Support Vector Machine. Common unsupervised learning algorithms are Apriori, Equivalence Class Transformation, Expectation Maximisation, Frequent Pattern-Growth, Hierarchical Clustering, K-Means Clustering, Mean Shift and Spectral Clustering.

2.4. Identified Research Gap

From the literature, it was identified that the lack of understanding of knowledge quality attributes has led to the development of ineffective knowledge-intensive systems. We identified that knowledge quality is integral to the success of a knowledge intensive system. The literature mentioned that knowledge intensive system usage is affected by quality of knowledge. Knowledge residing in an organisation must be preserved to ensure knowledge quality. Lastly, knowledge quality is important as it was identified that it is a catalyst for innovation within an organisation.

3. Proposed Solution

Based on our analysis from the literature, it was determined that the decision tree to be an ideal candidate of the algorithm. A study done showed that decision tree achieves a near 100% classification accuracy once the sample size is more than 20. Furthermore, the study also noted that the mean accuracy remained constant below 0.05 when sample size is more than or equal to 20. For sample size, a study by Beleites highlighted that 58 test samples are required to achieve >95% upper confidence interval. Furthermore, 106 test samples would yield an 88% observed sensitivity while 140 test samples would yield a 90% observed sensitivity. Though it is advisable that sample size of >100 would be better as the observed sensitivity was 88% for 106 samples and 90% for 140 test samples. Hence for this study, our sample size would be more than 100. To implement our proposed solution, we have come up with a flow diagram described in Figure 1.

<table>
<thead>
<tr>
<th>Knowledge Quality Attributes</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility</td>
<td>The degree of flexibility, ease of use, and ease of access [46]</td>
</tr>
<tr>
<td>Intrinsic Knowledge Quality</td>
<td>How knowledge has quality in its own right and is associated with the following, accuracy, reliability, and timeliness of the knowledge [8, 38]</td>
</tr>
<tr>
<td>Contextual Knowledge Quality</td>
<td>How knowledge is considered within the context of the task [8, 38]</td>
</tr>
<tr>
<td>Actionable Knowledge Quality</td>
<td>How knowledge is expandable, adaptable, or easily applied to tasks [8, 38]</td>
</tr>
</tbody>
</table>

Table 1. knowledge quality attributes
Figure 1 describes the flow of the proposed solution. A specific machine learning algorithm will be used, this case the decision tree algorithm with the training model at classifying knowledge quality in the system. The algorithm will then identify whether the knowledge is high, medium or low quality based on whether the knowledge quality attributes are met. This is set by the training model itself and should be identified when creating the datasets for the training model. The knowledge will then be run through by the algorithm. Once the knowledge is given its classification, the algorithm will then proceed to the next knowledge until all knowledge has been given its knowledge quality classification.
4. Methodology

This section discusses on the methods used to achieve the research objectives. The aim is to provide a detailed step-by-step method on how the It provides a detailed account of what was done to arrive at the findings and the ensuing discussion in later chapters. The research methodology was carried out in the following stages:

Phase 1
- Identify knowledge quality attributes from the literature

Phase 2
- Identify and analyse machine learning algorithms in the literature.
- Map machine learning algorithms to knowledge management areas based on algorithms suitability.
- Identify the machine learning algorithms to be used.

Phase 3
- Develop a framework to implement machine learning in knowledge management
- Develop a knowledge quality model based on knowledge quality attributes and current KM success models.

Phase 4
- Identify variables in a knowledge intensive system and map it to identified knowledge quality attributes.
- Test multiple variable combination using by measuring its accuracy index.
- Create a training model for the algorithm based on the findings from the previous experimentation.

Phase 5
- Collect knowledge using a knowledge intensive system.
- Use the algorithm to classify knowledge acquired from the system.
- From the results, identify the effectiveness of the Algorithm.

Fig 2. Research Methodology

In Phase 1, to ensure knowledge remains accurate, reliable and useful, it was identified that the following attributes must be imbued. The attributes are intrinsic, contextual and actionable. Accessible KQA was not selected as it ensures ease of use of knowledge and availability. This is already covered by actionable KQA. In Phase 2, we mapped and identified machine learning algorithms in the areas of knowledge management. The mapping is listed as follows:

A. **Systematic analysis**, it was identified that Artificial neural network, Eclat, FP-Growth, Linear regression, Logistic regression, K-Nearest Neighbour, Naïve Bayes, Decision Tree and Random Forest are suitable for this area. Artificial neural network, Linear regression, Logistic regression, K-Nearest Neighbour, Naïve Bayes, Decision Tree and Random Forest are suitable at predicting outcomes and performing regression analysis. Eclat, FP-Growth and Random Forest are suitable at data mining within the system. This means for systematic analysis, supervised and unsupervised machine learning algorithm is suitable to fulfil the assigned task.

B. **Planning**, it was identified that K-Means clustering and Expectation Maximisation were suitable for this area. These algorithms are suitable at forecasting or planning decisions within the system. This means for planning; only unsupervised machine learning algorithm can be implemented to fulfil the assigned task.

C. **Knowledge Acquisition**, it was identified that Decision Tree, Logistic Regression, Naïve Bayes, Random Forest and Support Vector Machine are suitable for this area. These algorithms are suitable at classifying knowledge types (tacit of explicit knowledge) and knowledge classification during
knowledge creation. This means for knowledge acquisition; only supervised machine learning algorithm is suitable to fulfil the assigned task.

D. **Knowledge Development**, it was identified that Artificial Neural Network, Mean Shift, Naïve Bayes and Spectral Clustering are suitable for this area. These algorithms are suitable at identifying text, images, video and audio that is present in a knowledge. This means for knowledge development, both supervised and unsupervised machine learning algorithm are suitable to fulfil the assigned task.

E. **Knowledge Creation**, it was identified that Decision Tree, Random Forest and Support Vector Machine are suitable for this area. These algorithms are suitable at determining whether knowledge quality is present in a knowledge. Knowledge quality is important as it ensures that the knowledge is useful, accurate and reliable. This means for knowledge creation; only supervised machine learning algorithm can be implemented to fulfil the assigned task.

F. **Knowledge Storage**, it was identified that Expectation Maximisation, Hierarchical Clustering and K-Means Clustering were suitable for this area. These algorithms are suitable at identifying data distribution within the system and to cluster the knowledge based on its characteristics. This means for knowledge storage; only unsupervised machine learning algorithm is suitable to fulfil the following task.

G. **Use of Knowledge**, it was identified that Apriori, Eclat and FP-Growth are suitable for this area. The algorithm is suitable at recommending knowledge to knowledge workers based on the common itemsets. These recommendations could be either whether the knowledge is suitable based on the knowledge workers expertise or based on the knowledge workers history. This means for use of knowledge; only supervised machine learning algorithm is suitable to fulfil the assigned task.

In Phase 3, to implement the decision tree algorithm at classifying knowledge quality, this is where we apply the Knowledge Management – Machine Learning Framework and the Knowledge Quality model. As the name suggests, the knowledge management – machine learning framework focuses on implementing identified machine learning algorithms to the different domains of knowledge management. The model ensures that the right machine learning model and data is identified. To successfully classify knowledge quality in a knowledge intensive system, this is where a knowledge quality model is implemented. The knowledge quality model is divided into three areas which are knowledge process, knowledge context and knowledge source. Knowledge process ensures that knowledge remains usable and preserved. Knowledge context touches on the content itself as the knowledge must be intrinsic, contextual and actionable as it must adhere to the knowledge quality attributes. Knowledge source focuses on identifying the right person.

In Phase 4, we identified the datasets required to train and test the algorithm. The purpose of which is to train the data with near perfect data conditions and measure whether it can classify the data accurately. By executing multiple experiments on these proposed datasets, it allows us to identify what are the required variables to be used for the training model. In total, eight datasets were proposed and tested in different iterations. The datasets are as follows:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Variables</th>
<th>Measured By</th>
<th>Knowledge Attributes</th>
<th>Quality Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Knowledge Usefulness</td>
<td>Measured by linking knowledge to organisation goal and the ensuing KM strategy</td>
<td>Actionable Quality</td>
<td>Knowledge Quality</td>
</tr>
<tr>
<td></td>
<td>Knowledge Reliability</td>
<td>Measured by the amount of views a knowledge has. Views is given by knowledge workers of the system</td>
<td>Contextual Quality</td>
<td>Knowledge Quality</td>
</tr>
<tr>
<td></td>
<td>Knowledge Approval</td>
<td>Measured by whether the knowledge is approved or not.</td>
<td>Intrinsic Quality</td>
<td>Knowledge Quality</td>
</tr>
<tr>
<td></td>
<td>Knowledge Rating</td>
<td>Measured by the amount of ratings a knowledge has. Ratings is given by knowledge workers of the system</td>
<td>Linked to Actionable, Contextual and Intrinsic Knowledge Quality</td>
<td>Knowledge Quality</td>
</tr>
</tbody>
</table>

Table 2. Dataset Combination for Experimentation
For Phase 5, to test the algorithm, we liaised with a client from the industry. A knowledge intensive system was then used to capture the knowledge. A small sample size is chosen as to adequately simulate the usage of the system in an organisation and a small sample size is easier to monitor and control during the pilot test. To ensure an 88% - 90% observed sensitivity, we opted to collect up to more than 120 sample data for this study. The data is then pumped into the training model and the results are then measured.

5. Results and Discussion

This section discusses on the results and discussion on the findings. The aim is to provide the results from the experimentation and pilot test stage. Also, in the following section, the section provides a discussion on the findings which includes theoretical and practical implications.
5.1. Dataset experimentation

In total, four different experimentations were carried out. The suitability is determined by the confusion matrix and accuracy index results. Each dataset was given its own criteria and label at classifying high quality, medium quality and low quality. This may differ from different datasets which explains the results in the following figure and table.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy Index</th>
<th>Suitability</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset A</td>
<td>0.89</td>
<td>Not Suitable</td>
<td>Accuracy rating is less than 1.0</td>
</tr>
<tr>
<td>Dataset B</td>
<td>1</td>
<td>Suitable</td>
<td>Accuracy rating is 1.0</td>
</tr>
<tr>
<td>Dataset C</td>
<td>1</td>
<td>Suitable</td>
<td>Accuracy rating is 1.0</td>
</tr>
<tr>
<td>Dataset D</td>
<td>0.92</td>
<td>Not Suitable</td>
<td>Accuracy rating is less than 1.0</td>
</tr>
<tr>
<td>Dataset E</td>
<td>0.88</td>
<td>Not Suitable</td>
<td>Accuracy rating is less than 1.0</td>
</tr>
<tr>
<td>Dataset F</td>
<td>1</td>
<td>Suitable</td>
<td>Accuracy rating is 1.0</td>
</tr>
<tr>
<td>Dataset G</td>
<td>0.82</td>
<td>Not Suitable</td>
<td>Accuracy rating is less than 1.0</td>
</tr>
<tr>
<td>Dataset H</td>
<td>0.66</td>
<td>Not Suitable</td>
<td>Accuracy rating is less than 1.0</td>
</tr>
</tbody>
</table>

Based on the results from the table above, the finalised datasets were conceived based on the variables in dataset B, dataset C and dataset F. Furthermore, it was determined that the identified machine learning algorithms were indeed suitable in classifying knowledge quality. This is due to the variable combination yielded a 1.0 accuracy rate while the rest yielded an accuracy rate below 1.0. The finalised datasets are then tested to measure its accuracy rating before the experimentation phase could be concluded. This is to ensure that later on, the algorithm is able to classify the knowledges accurately.

5.2. Pilot Test

Using the DecisionTreeClassifier in Python, both Gini and Entropy criteria were used to analyse the pilot test results. The figures below show the classification results of both Gini and Entropy criterion. From the results, it was noticed there were 4 different types of classification for both Gini and Entropy. The results differ in terms of how the model labeled each knowledge and the decision tree structure. Both Gini and Entropy showed similar results, although what differentiates these two criterions are how the decision tree is structured. From a machine learning perspective, we can see how the model reacts with the pilot test data with its minimal class separation and predictor connection. From a knowledge management perspective, we can clearly see the benefit of applying machine learning. The result clearly
shows that we can clearly identify which knowledge is labelled as low, medium or high quality. The application of machine learning ensures knowledge quality remains present in a knowledge intensive system thus ensuring knowledge and the system itself to be beneficial to not only knowledge workers and the organisation as well.

5.3. Discussion

Based on the experimentation results, it shows that machine learning has a strong potential when implemented in a knowledge intensive system. This also shows from a practical perspective that the machine learning is indeed beneficial. Machine learning does have strong theoretical and practical implementation. It not only enhances a knowledge intensive system but also opens up a new area of exploration. From a theoretical perspective, machine learning algorithms is beneficial to knowledge intensive system. From a practical perspective, machine learning has the potential to further enhances knowledge intensive system. The findings also open up potential research areas for machine learning in knowledge management.

5.3.1 Theoretical Implementation

The results from the experimentation and implementation phase does show that machine learning algorithm has a strong benefit to knowledge management and knowledge intensive systems in general. From the results we can see how machine learning can have a positive impact in knowledge management. Machine learning has a strong potential in knowledge management and further algorithms can be applied. The experiment also shows that by implementing machine learning, knowledge quality can be measured effectively given the right variable combination is found. It was clearly shown that knowledge quality could be ascertained, allowing knowledge to be labelled or graded accordingly. Machine learning applications are wide and from the literature, each algorithm has its own specific functionality. The literature also identified the potential usage of other machine learning algorithms to the different areas of knowledge management. The findings also open up potential research areas such applying recommender engine, regression analysis or clustering in knowledge management.

5.3.2 Practical Implementation

As we can see, from a knowledge management perspective, implementing machine learning in does benefit to not only knowledge workers and the organisation as well. The results show that knowledge quality level can be ascertained and classified accordingly. The algorithm manage to classify the data from the pilot test using the training model that was given. This ultimately shows that machine learning can be used to enhance knowledge intensive systems further. This in turn would generate more usage of the system and knowledge workers are much more willing to contribute knowledge in, knowing that high-quality knowledge is present. As more knowledge is added, the system can see an increase of knowledge sharing and knowledge reuse. Knowledge itself in an organisation will be captured effectively and knowledge loss can be mitigated.

6. Conclusion

From the research, it is shown that machine learning algorithm does have both theoretical and practical implication in knowledge management. Machine learning does have strong benefits to knowledge management and from the implementation results, it can further improve the performance of knowledge intensive systems further. From the pilot test result, the decision tree algorithm can be used to classify knowledge within a knowledge intensive system. It also shows that machine learning is indeed capable at enhancing knowledge intensive system further. The experimentation also has opened to potential areas to implement machine learning in knowledge management. The results also show the algorithm is indeed capable to determining whether a knowledge is high quality, medium quality, low quality based on the attributes that it has. The algorithm also benefits knowledge workers alike as based on the classification given; knowledge workers can determine the perceived value of the knowledge present. The limitation of the algorithm is that firstly, it is limited to the variables that are found in the knowledge-intensive system. Secondly, it is unable to analyse the content of each knowledge found and is highly dependent on inputs given by knowledge workers. We conclude that a high-quality knowledge is beneficial to
knowledge workers as it has not only a high perceived value but is usable, contextual and relevant. This is important for knowledge reuse and knowledge sharing and, ultimately, the usage of the knowledge-intensive system.

7. Future Work

In this study, we only applied decision tree algorithm to classify knowledge quality in a knowledge intensive system. For future work, the study proposes on implementing other machine learning algorithms in other areas of knowledge management based on the mapping given.

References