JIRKM | **Journal of Information Retrieval and Knowledge Management**
e-ISSN: 2289-2664
Volume 3, 2017

The Society of Information Retrieval and Knowledge Management (PECAMP) would like to express its appreciation and gratitude to the contributors and reviewers of the journal's articles.

Published by:

The Society of Information Retrieval and Knowledge Management

Content

COMPLEXITY OF KEYWORD EXTRACTION: EXPERTS VS. SLIDING WINDOW

Siaw Nyuk Hiong
Institut Pendidikan Guru
Kampus Batu Lintang
Jalan College
ftsm2006@yahoo.com

ABSTRACT

This paper studied the complexity of keyword extraction via the experts. The research findings can bridge the gap in automated keyword extraction with a human added value inputs. A set of 100 complex sentences were prepared from TimeBank1.2 corpus as the data for keyword extraction. A baseline that uses sliding window of size n=5 based on graph-based approach is used to extract keywords with the same data set. Three graph centrality measures are used as the metric to compare sliding window against expert extracted keywords. Kappa co-efficient is computed to measure inter-rater (expert) agreement. The overall results indicate that keyword extraction is a complex task as the experts have difficulty to agree on identifying and ranking the keywords.

KEYWORD: *Keyword Extraction, Graph-Based Text, Inter-Rater Agreement*

1.0 Introduction

An important task in many document processing research like information retrieval, text clustering (Han et al. 2007), classification, text summarization (Zha, 2002) is keyword extraction. This task involves the extraction of a few important words to represent a text. The quality of the extracted keywords affect the performance of document processing. Keyword extraction can be corpus dependent or independent. Corpus dependent approach needs a collection of data which is usually domain specific. A state-of-the-art measure used by corpus dependent measure for keyword indexing is *tfidf*. This measure extracts keywords that have high occurrence frequency in a document but low occurrence frequency in the corpus. A main drawback of *tfidf* measure is that it ignores the semantic structure of a text.

Generally, corpus independent approach yields better performance (Liu & Wang 2007). Currently, domain-independent keyword extraction based on graph-based ranking have been an active area of research (Mihalcea & Tarau 2004; Palshikar 2007; Liu 2008; Wan & Xiao 2008; Tsatsaronis et al. 2010; Litvak et al. 2011). The nodes for the graphs are represented using words or phrases. Edges between the nodes are represented by co-occurrence or semantic relations. A characteristic of this approach is that it does not need any language-specific resources or any manually constructed data (Mihalcea & Tarau, 2004; Cheng et al. 2013). It is highly portable to other language domain and it takes structural features of a text into account. The performance of this approach is comparable to *tfidf*.

Since many keyword extraction researches have focused on evaluating the performance of the approach used, not many have studied the complexity of the task

when it is carried out by a human. It cannot be denied that human has been involved in many text based research as in Message Understanding Conference (MUC), Automatic Content Extraction (ACE), text mining and Predicate Argument Structure (PAS). These research do mention 'labour and cost intensive' as in MUC and ACE as well as knowledge-driven text mining that required knowledge from the expertise. As for PAS, the semantically annotated lexical databases are prepared by human. Thus, it is the interest of this paper to study the complexity of keyword extraction task via human approach. Graph-based approach is used since no corpus is needed and it can be employed on a single sentence or document. A commonly used approach to extract keyword is the sliding window or window of size-n. A keyword extraction baseline using sliding window of size $n=5$ employing graph-based ranking approach is used for comparison. This research is significant as it provides findings that can lead to future research in proposing an automated keyword extraction method to augment the problem faced by human.

2.0 Graph-Based Text Representation

Different level of text, for example phonology, phonetics, morphology, syntax, semantics and pragmatis can have network structure (Liu et al. 2008). The different relationships defined for the edge of a graph will decide on the type of graph that can be constructed. Three different types of graphs can be built and are described below.

2.1 Co-occurrence Network

As the name implied, nodes of this type of graph consist of words that appear together in a fixed size window or a sentence or in the same paragraph or are connected to one another inside the network. The link that exists between two nodes represents the edge of the graph. Directed or undirected graph can be constructed. Directed graph takes into consideration the direction of word order in the window or otherwise for undirected graph. The strength of connection between words is indicated by the weight of the edge. Weight can be computed using log-likelihood measure (Dunning, 1993). Besides, Chi-square and pointwise mutual information measures can be used as well.

2.2 Dependency Network

Words that appear together based on dependency syntax are extracted to represent the node of the graph. Word co-occurrence formalism is based on dependency grammar, for example head modifier. This type of syntactic structure consists of words as the node and relationship between words is the dependencies relations. Dependency network based on this structure has words as the nodes and the dependency relation as the link between the nodes. Directed graph is constructed by taking the dependency relation that goes from the head word to its modifier. The following example of the dependency network for the sentence "Susan has a car". The three syntactic dependencies are: "has Susan", "has car" and "car a". The node "has" is the root of the dependency tree. This is a directed network

with the arrow pointing from the modifier to the head (from the child to the parent). From research, it is observed that 70 per cent of syntactic dependencies is within a distance of two words (Mihalcea & Radev 2011).

2.3 Semantic Network

From linguistic perspective, semantic network represents concepts or meanings of a sentence in the form of a network. The concepts or meanings represent the nodes and semantic relationships between the concepts or meanings represent the edge of the graph. The graph can be directed or undirected. WordNet (Miller, 1995) is one of the most widely used semantic networks. It has been applied in NLP processing and information retrieval. Sigman and Cecchi (2002) studied semantic relationship in semantic network based on WordNet: polysemy and synonym, antonym, hypernymy (*is-a* relations), hyponymy and meronymy (*part-whole* relation).

3.0 Related Research

Many researches have applied graph-based text representation for natural language text. Graphs have been applied for text summarization (Erkan & Radev, 2004; Mihalcea & Tarau, 2004), keyword/keyphrase extraction (Mihalcea & Tarau, 2004; Palshikar, 2007; Liu et al. 2008; Wan & Xiao 2008; Tsatsaronis 2010; Litvak et al. 2011; Zhao & Zeng 2013; Abilhoa & de Castro 2014), classification (Wang et al. 2005; Hassan et al. 2007; Valle & Ozturk 2011), word sense disambiguation (Mihalcea et al. 2004), answering complex sentence (Chali et al. 2011), information retrieval (Tomita et al. 2004), event detection and tracking (Sayyadi et al. 2009). Text as a graph is able to capture relations and structure in the text (Valle 2011). This section will focus review on automated keyword extraction based on graph-based text.

Mihalcea and Tarau (2004) used TextRank for keyword extraction. Window of *N* words is used to extract co-occured words. The words represent the vertices of the graph and syntactic link between words represent the edge of the graph. Syntactic filter is applied to filter for words with certain part of speech (POS), for example nouns, verbs. TextRank is then runs on the graph and the vertices are ranked. The top ranked N words are extracted as the keywords. Palshikar (2007) proposed a hybrid structural and statistical approach to extract keywords from documents. The document is preprocessed to remove numbers, stop words, infrequent words and stemming using Porter stemmer. The document is represented as undirected graph and word is used as the vertex. Co-occurrence frequency between two words is used to compute dissimilarity measure and used as label for the edge fo the graph. Semantic network has been constructed with the help of lexical database like WordNet or HowNet (Liu et al. 2008). Semantic similarity is computed base on the maximum similarity between word senses measured by WordNet or HowNet. Pagerank is used as the algorithm to run on the network. The nodes are then sorted and the top N ranked words are extracted as keywords. Wan and Xiao (2008) proposed an approach that extends a single document to include neighbour documents to improve keyphrase extraction. The single document provides local information and

neighbour documents provide global information. Similarity search technique is used to obtained neighbourhood documents. Words with specific POS, for example, noun and adjective, are selected to represent the nodes. Edge is represented by the relation that exists between two words. Weight of the edge is the word co-occurrence which also represents cohesion relationship between words. Window of words size *w* is used as the co-occurrence window where $2<w<20$. PageRank algorithm is used to compute the score for each word in the graph. Tsatsaronis, Varlamis and Nørvag (2010) proposed a graph-based ranking algorithm, SemanticRank, for keyword extraction. Semantic graph is constructed based on the semantic relatedness between words measured with reference to WordNet and Wikipedia. Word of *n-gram* up to a size of 5 is used as the window to extract semantically related words. Litvak et al., (2011) proposed a graph-based language-independent keyphrase extractor, DegExt, in their research. Document is first pre-processed by stemming and removing stopwords. Distinct words are used to represent vertices of the graph. The order-relationships between words represent the edge. Zhao and Zeng (2013) applied graph model, semantic space and location of words in their research to extract keywords from Chinese micro blog. Micro blog API of users are being pre processed by data cleaning, word segment, POS tagging and stop word removal. A word graph is then constructed and a sequence number is given to the word based on the location. Keyword is extracted based on word co-occurence and a weight calculated using Score formula. Statistical weight TFIDF is then computed for a semantic space created based on topic detection. Finally, the rank value for words are computed based on word location. Words with smaller location will rank higher. Abilhoa and de Castro (2014) proposed a three steps graph based approach to extract keywords. After pre processing, a word graph is constructed that take nearest neighbours and all neighbours into consideration. Graph centrality measure is computed to identify the keywords.

State-of-the-art method that has been applied very frequently in these researches is word co-occurrence within a fixed window size (Mihalcea & Tarau 2004; Wan & Xiao 2008; Tsatsaronis et al. 2010). A fixed size window is similar to mobile window and word co-occurrence is study based on linear word proximimity. Vertex of graph is represented by word (Mihalcea & Tarau 2004; Palshikar 2007; Liu et al. 2008; Wan & Xiao 2008; Tsatsaronis et al. 2010; Zhao & Zeng 2013; Abilhoa & de Castro 2014). Not only extracting good word representation for vertices of the graph is important, it is also important to formulate better methods to represent the edge and weights of the graph. Relationships that occur between vertices represent the edge of the graph. The type of relationship defined is depending on the needs of the research. Edge is represented by dissimilarity measure (Palshikar 2007), content semantic relation (Liu et al. 2008), semantic relation (Tsatsaronis et al. 2010; Zhao & Zeng 2013) and word order (Litvak et al. 2011). Weight of graph represents the frequency of occurrence of an edge. It is represented differently by strength of semantic relatedness (Liu et al. 2008), cohesion of relation between words (Wan & Xiao 2008), compound weight of semantic measure and co-occurrence frequency of edge (Tsatsaronis et al. 2010; Zhao & Zeng 2013; Abilhoa & de Castro 2014).

All the researches reviewed in this section have applied graph centrality measures as a way to rank vertices, term weighting or as a method of evaluation with other proposed methods. The more commonly used standard technique to rank vertices is PageRank and its equivalent Random Walk (Mihalcea & Tarau 2004; Liu et al. 2008;

Wan & Xiao 2008; Tsatsaronis et al. 2010). Better results and performance are reported for graph-based representation of text for keyword extraction for all the researches reviewed in this section. Table 1 summarizes some important aspects of the graphs from these researches and Table 2 summarized the related research for graph edge and weight representation.

Table 1. Summary of Graph-Based Keyword Extraction Research

Research	Graph-based Text Representation	Key Word Extraction	Approach Linguisti/ Statistica/ Hybrid	Centrality Measures	Application
TextRank: Bringing order into texts. (Mihalcea & Tarau 2004)	Vertex=word Edge=co-occur relation	Window of *N* size	Statistical	Pagerank	Keyword extraction
Keyword Extraction from a Single Document Using Centrality Measures (Palshikar 2007)	Undirected graph. Vertex=word Edge=dissimilarity measure between words	Frequency of co-occurrence	Statistical	Eccentricity Betweeness Closeness Proximity	Keyword Extraction
A Text Network Representation Model (Liu et al. 2008)	Vertex=content words Edge = semantic relation between content Weight = strength of semantic relatedness	Semantic relatedness using WordNet/How Net	Hybrid	Pagerank	keyword extraction
Single Document Keyphrase Extraction Using Neighborhood Knowledge (Wan & Xiao 2008)	Vertex=POS filtered word Edge=relation between word Weight=cohesion of relationship between words	Window of size w ($2<w<20$)	Statistical	PageRank	Keyword extraction

SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs (Tsatsaronis 2010)	Node=word Edge=semantic relation Weight= semantic measure between words * word co-occurrence frequency	Word of *n-gram* (n≤5)	Hybrid	PageRank HITS	Keyword extraction
DegExt – A Language-Independent Graph-Based Keyphrase Extractor (Litvak 2011)	Vertex=distinct word Edge=word order-relationships	Distinct words	statistical	HITS	Keyphrase extraction
Micro-blog keyword extraction method based on graph model and semantic space (Zhao & Zeng 2013)	Vertex= word Edge= word co-occurrence	Keyword based on co-occurrence Weight by *TFIDF*	Statistical	-	Keyword extraction
A keyword extraction method from twitter messages represented as graphs (Abilhoa & de Castro 2014)	Vertex= word Edge= word co-occurrence	Keyword based on co-occurrence	Statistical	Closeness Proximity	Keyword extraction

Table 2. Summarization of Graph Edge and Weight Representation

Graph Edge & Weight Representation	Related Research
Edge	dissimilarity measure (Palshikar, 2007), content semantic relation (Liu et al. 2008), semantic relation (Tsatsaronis 2010) and word order (Litvak et al. 2011)
Weight	strength of semantic relatedness (Liu et al. 2008), cohesion of relation between words (Wan & Xiao 2008), compound weight of semantic measure and co-occurrence frequency of edge (Tsatsaronis et al. 2010; Zhao & Zeng 2013; Abilhoa & de Castro 2014)
Centrality Measure	as a way to rank vertices, term weighting or as a method of evaluation with other proposed methods
	more commonly used standard technique to rank vertices is PageRank and its equivalent Random Walk (Mihalcea & Tarau 2004; Liu et al. 2008; Wan & Xiao 2008; Tsatsaronis et al. 2010)

4.0 Discussion

The review on automated keyword extraction has shown that it is not a simple task as many different resources and techniques are involved. An approach has to be selected to decide how word should be extracted. Based on review in this section, word co-occurrence within a fixed window size (Mihalcea & Tarau 2004; Wan & Xiao 2008; Tsatsaronis et al. 2010) has been applied. Since graph is used for text representation, the type of relation for edge representation and ways to compute edge weight require knowledge from the domain of graph theory. Natural language processing tasks are needed as a written text has to be pre-processed (POS tagging, stemming, tokenization, etc.) before relevant word for vertex of a graph can be extracted. Statistical knowledge is also needed for the computation of graph centrality measures. It can be concluded that for an automated keyword extraction to be successfully implemented, expertise from various fields has to work together. It is the interest of this paper to study whether keyword extraction is also an equally complex task for human experts since all the knowledge needed for keyword extraction is already in the minds of the experts.

5.0 Graph Centrality Measures

In order to evaluate the effectiveness of the constructed graphs in capturing the informative content of a sentence, various centrality measures based on node neighbourhood can be employed. According to Valle (2011), centrality measures can be

used to compute the relative importance of nodes in a graph. In this research, each sentence is a graph network consisting of link words which represent the vertices. The link words resemble more to a neighbourhood network. Thus, three neighbourhood based centrality measures are used to compute centrality scores in this research.

5.1 Eigenvector

Eigenvector centrality emphasizes on the number of connection a node has to other nodes as well as how well the neighbours of these nodes are connected.

5.2 Pagerank

Pagerank study link between pages. The link from one page to another is the votes among pages. A page normally looks for a good page to link and pages with a lot of inlink should have higer rates. Voting of a link is normalized by the total outgoing links for a page. Pages having the same Pagerank will have the same voting and pages with higher votes is considered as more important.

5.3 Degree

This is the simplest centrality measure presented by Freeman (1978). Degree refers to the number of edges a node has with other nodes in a graph network.

6.0 Method

6.1 Data

The shallow grammatical information provided by POS can help to identify the informative content of a text (Lioma & Blanco 2009). They also indicated that POS *n-gram* for verb-noun should give more information compared to prepositions and adverbs *n-gram*. Jespersen's Rank Theory (Jespersen 1920) actually gives degree ranking in terms of content-bearing for POS as below (Table 3):

Table 3. Jespersen's Rank Theory for POS

Degree of ranking (content-bearing)	POS
1st	nouns
2nd	adjectives, verbs and participles
3rd	adverbs
4th	remaining POS

Stanford POS tagger is used to filter out funtional words to extract only verb and noun. A total of 100 complex sentences are prepared from [1]TimeBank1.2 corpus for this research. First, the sentences are parsed with the tagger. All identified verbs and nouns which function as the content words are underlined. Examples of the data are shown in Figure 1.

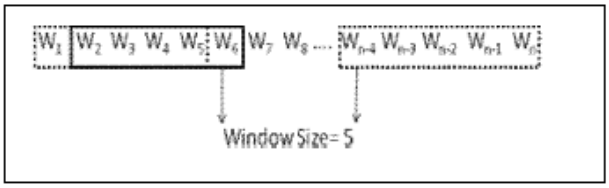


Figure 1. Sliding Window of Size 5

6.2 Sliding Window

Sliding window is a common statistical collocation extraction technique introduced by Church and Hanks (1990). This technique moves a sliding window of size *T* over a text as shown in Figure 1.

The 100 sentences tagged with Stanford POS tagger are also used to extract collocate word based on sliding window of size 5. Window of size *n=5* is selected as a large *n* word window actually lower the precision value (Mihalcea & Tarau 2004). A word *w* with its collocate w_i that appears within the same window will be extracted to construct the vertex of a graph. *w* and w_i have the following definition:

$$w, w_i \in \{verb, noun\}$$

A relation that exists between *w* and w_i constitutes the edge of the graph. Frequency of co-occurrence constitutes the weight. Three graph centrality measures are computed for each sentence: eigenvector, Pagerank and Degree. Table 4 shows an example of a sliding window with the collocate verb-noun and noun-noun extraction.

Table 4. An Example of a Sliding Window Used to Extract Collocate Verb-Noun and Noun-Noun

<table>
<tr><td>Raw Text</td><td colspan="6">The thrift holding company said it expects to obtain regulatory approval and complete the transaction by year-end.</td></tr>
<tr><td>POS Tagging</td><td colspan="6">The/DT thrift/NN holding/VBG company/NN said/VBD it/PRP expects/VBZ to/TO obtain/VB regulatory/JJ approval/NN and/CC complete/VB the/DT transaction/NN by/IN year-end/NN ./.</td></tr>
<tr><td rowspan="3">Sliding Window</td><td></td><td>w_1</td><td>w_2</td><td>w_3</td><td>w_4</td><td>w_5</td></tr>
<tr><td>window 1</td><td>The</td><td>thrift</td><td>holding</td><td>company</td><td>said</td></tr>
<tr><td>window 2</td><td>thrift</td><td>holding</td><td>company</td><td>said</td><td>it</td></tr>
</table>

	window 3	holding	company	said	it	expects
	window 4	company	said	it	expects	to
	window 5	said	it	expects	to	obtain
	window 6	it	expects	to	obtain	regulatory
	window 7	expects	to	obtain	regulatory	approval
	window 8	to	obtain	regulatory	approval	and
	window 9	obtain	regulatory	approval	and	complete
	window 10	regulatory	approval	and	complete	the
	window 11	approval	and	complete	the	transaction
	window 12	and	complete	the	transaction	by
	window 13	complete	the	transaction	by	year-end.

Frequency of verb-noun, noun-noun collocation

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
w_1	0	3	3	1	0	0	0	4	2	0
w_2	3	0	1	0	0	0	0	3	2	0
w_3	3	1	0	3	0	0	1	2	0	0
w_4	1	0	3	0	1	0	3	0	0	0
w_5	0	0	0	1	0	3	3	0	0	1
w_6	0	0	0	0	3	0	1	0	0	1
w_7	0	0	1	3	3	1	0	0	0	0
w_8	4	3	2	0	0	0	0	0	2	0
w_9	2	2	0	0	0	0	0	2	0	0
w_{10}	0	0	0	0	1	1	0	0	0	0

w_1 = said, w_2=holding, w_3=expects, w_4=obtain, w_5=complete, w_6=transaction, w_7=approval, w_8=company, w_9=thrift, w_{10}=year-end

6.3 Keyword Identification by Experts

Three experts are asked to rank the top five content words according to the order of importance for the 100 sentences (Figure 2). The experts are the raters in this research. Based on Cambridge Advanced Learner‘s Dictionary 2013, an expert is "a person with a high level of knowledge or skill relating to a particular subject or activity". The 100 sentences are related to the English subject area, thus two of the experts involved with the data annotation are from the field of English Language. Since the data is also related to

the activity of natural language processing, the third expert is from the field of Computational Linguistic who is also very well verse in the English language. Table 5 shows the information for the three human experts and Figure 3 shows the human experts data annotation information.

Table 5. Information on the Three Experts for Data Annotation

Expert	Expertise and Qualification
1	• Senior Language Teacher attached to Center for Language Studies in a local university • Bachelor degree in TESL, Master degree in EnglishLiterature
2	• Primary school English teacher with more than 10 years of teaching experience in English. • Bachelor Degree in Education
3	• Assoc. Professor (Dr.) attached to Faculty of Computer Science and Information Technology in a local university • PhD in Natural Language Processing

This data is prepared to collect language expert annotation of keywords in a sentence. It will be used for a PhD research in the field of computational linguistic. The main objective is to compare an expert annotated data with a computer annotated data. All annotated data will be treated with confidentiality.

Instruction: Select and rank the underlined words in each sentence according to the top five order of importance.

ID	Sentence	Ranking (more important ← → less important)				
		1	2	3	4	5
Example	We *noted*(a) how some electronic techniques(b) *developed*(c) for the defense(d) effort(e) have eventually been *used*(f) in commerce(g) and industry(h).	c	d	b	f	g
1	The thrift(a) *holding*(b) company(c) *said*(d) it *expects*(e) to *obtain*(f) regulatory approval(g) and *complete*(h) the transaction(i) by year-end (j).					
2	Organizers(a) *state*(b) the two days(c) of music(d), dancing, and speeches(e) is *expected*(f) to *draw*(g) some two million people(h).					
3	Computer experts(a) familiar with the flaws(b), found in Intel's 80486 chip(c), *say*(d) the defects(e) don't *affect*(f) the average user(g) and are likely to be cleared up before most computers(h) *using*(i) the chip(j) as their "brains(k)" *appear*(l) on the market(m) sometime next year(n) .					
4	Integra-A Hotel amp Restaurant Co. (a) *said*(b) its *planned*(c) rights(d) offering to raise about $9 million was *declared*(e) effective and the company(f) will *begin*(g) *mailing*(h) materials(i) to shareholders(j) at the end(k) of this week(l).					

Figure 2. Examples of Data for Expert Annotation

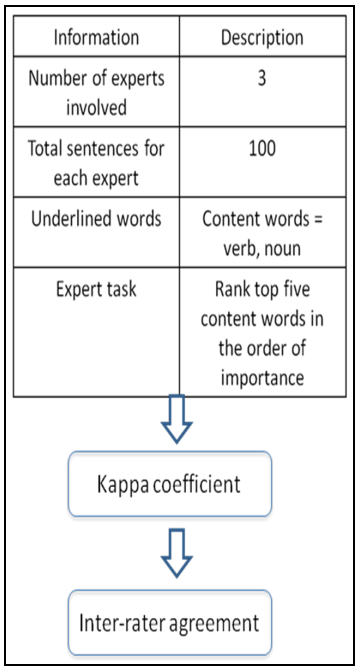


Figure 3. Human Expert Data Annotation Information

6.4 Inter-Rater Agreement Evaluation

The 100 data annotated by the three experts are each evaluated against the top five ranking words obtained from the graph centrality measures. The top five ranking words are computed using the three graph centrality measures base on adjacency matrix prepared from sliding window. In order to compare with the baseline sliding window method, three computation results with 100 top five ranking words each is obtained as the output. Thus, a total of three 100 top five ranking words sets are computed for sliding window method. These results are computed from the three graph centrality measures based on frequency matrix only. Each expert data are then evaluated against the 300 top five ranking words. Thus, each expert has three 100 sets of data comparison output with the graph centrality measures for Sliding Window.

Based on the comparison of expert data and graph centrality measures output, inter-rater Kappa coefficient is computed. Each graph centrality measure is used in turn as the reference method to compute inter-rater Kappa coefficient. For example, eigenvector centrality measure is used as the reference method to evaluate inter-rater agreement between expert1 (*A1*) and expert2 (A2). Table 6 shows the information for the computed output. The Embedded Semantic Window (E-SR Window) is a linguistically injected model (Linguistic Window Model) to interpret multiple events in a complex sentence in a research carried out by Siaw (2015). The model takes thematic hierarchy (Jackendoff, 1972) into consideration. Table 7 shows the sets of expert data comparison with graph centrality measures.

Table 6. Graph Centrality Measures Output Information

Method	***E-SR* Window**					**Sliding Window**				
Graph Centrality Measures	Eigenvector		Pagerank		Degree	Eigenvector		Pagerank		Degree
Type of Matrix	*F*	*F*T*	*F*	*F*T*	-	*F*		*F*		-
Top Five Ranking Words (Total Number of Sets)	100	100	100	100	100	100		100		100

F = frequency, F*T = frequency * thematic hierarchy

Table 7. Expert Data Evaluated Against Top Five Ranking Words for Each Graph Centrality Measure

Sliding Window												
Graph Centrality Measure/ Reference Method	1	2	3	Graph Centrality Measure/ Reference Method	1	2	3	Graph Centrality Measure/ Reference Method	1	2	3
A1		1		A2		1		A3			
		1							1		
	1	1			1	1	1		1		
		1	1		1	1	1		1	1	1
		1	1		1	1	1				1
		1				1				1	1
		1	1		1				1		
	1	1	1		1	1	1		1	1	
… … … … … .. …	.	.	.	… … … … … .. …	.	.	.	… … … … … .. … … …	.	.	.
	.	.	.		.	.	.		.	.	.
	.	.	.		.	.	.		.	.	.
	.	.	.		.	.	.		.	.	.
	.	.	.		.	.	.		.	.	.
	.	.	.		.	.	.		.	.	.
	.	.	.		.	.	.		.	.	.

A1, A2, A3 = expert 1, 2, 3 respectively
1= Eigenvector (F), 2= Pagerank (F), 3= Degree

According to Warrens (2013), inter-rater agreement can be used to indicate the reliability of nominal categories items. Inter-rater agreement approach is used since it complies with the definition of nominal categories item as ranking is similar to classification task and at least two raters are involved to perform the rating tasks for inter-rater agreement.

Based on the top five ranked content words, a one to one evaluation for each sentence is carried out to compare the ranking by two experts. Any three out of the top five ranked content words which are similar between the two experts is considered as a positive case, otherwise it is a negative case. Cohen's Kappa for two categories complies with the classical definition of reliability (Kraemer 1979). This coefficient is the most popular for indicating nominal scale inter-rater agreement (Hsu & Field 2003; Zwick 1988). The following terminology and notations are used for inter-rater agreement calculation.

- $\{i \mid i \in I\}$: a set of items with cardinality i.
- $\{c \mid c \in C\}$: a set of categories with cardinality c.
- $\{r \mid r \in R\}$: a set of raters with cardinality r.

The category c for inter-rater agreement used in this research is given by the following definition. Since top five ranking words are being compared, a precision at n P@n of ≥ 0.6 is used as the cut-off point to return matched cases.

$$c = \begin{cases} 1, & \text{if } n(A \cap B) \geq 3 \\ & A \in \text{top five content words of } r_A \\ & B \in \text{top five content words of } r_B \\ 0, & \text{otherwise} \end{cases}$$

Let π_{ij} for i,j =1, ……, c denote the proportion of items classified into category i by the first rater and into category j by the second rater. The square table $\{\pi_{ij}\}$ is an agreement table. The marginal totals of $\{\pi_{ij}\}$ are denoted by

$$\pi_{i+} = \sum_{j=1}^{c} \pi_{ij} \quad \text{and} \quad \pi_{+i} = \sum_{j=1}^{c} \pi_{ji}$$

The main diagonal of $\{\pi_{ij}\}$ is the cell probabilities π_{ii} that shows the total sentences that are put into the same categories by both raters. Cohen's Kappa is defined by:

$$k = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i}}$$

where $\sum \pi_{ii}$ is the proportions of observed agreement and $\sum \pi_{i+}\pi_{+i}$ = expected agreement.

Table 8 is an example of an agreement table in this research. The proportions of observed agreement for two raters r_{A2} and r_{A3} are 34 and 40 which represent the matched (+case) and non-matched (-case) category respectively.

Table 8: An Example of Agreement Table

		r_{A3}	r_{A3}
		+case	-case
r_{A2}	+case	34	11
r_{A2}	-case	10	40

The total percentage of observed agreement $\sum \pi_{ii}$ is given by:

$$\sum \pi_{ii} = \frac{34+40}{95} = 0.778$$

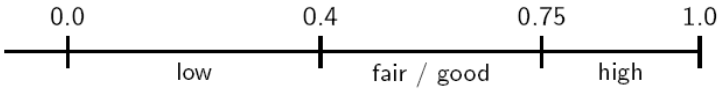
where the total items= 34+40+11+10=95. The total percentage of expected agreement is given by:

$$\sum \pi_{i+} \pi_{+i} = \left[\left(\frac{(34+10)}{95}\right) * \left(\frac{(34+11)}{95}\right)\right] + \left[\left(\frac{(11+40)}{95}\right) * \left(\frac{(10+40)}{95}\right)\right] = 0.501$$

Cohen's Kappa, *k*, is given by:

$$k = \frac{0.778 - 0.501}{1 - 0.501} = 0.556$$

Kappa coefficient and strength of agreement given by Green (1997) have the following interpretation:



Based on this intepretation, the Kappa coefficient of 0.501 shown in the above example indicates that the inter-rater agreement between r_{A2} and r_{A3} is fairly good in ranking the top give content words in this research.

7.0 Results and Discussion

Analysis of inter-rater agreement for sliding window shows that both *A1-A2* and *A2-A3* have similar inter-rater agreement with a Kappa coefficient (κ) of 0.40 and 0.41 respectively where Pagerank is the reference method. A slightly higher Kappa coefficient of 0.30 is obtained for *A1-A2* inter-rater agreement where degree is the reference method.

An equally low Kappa coefficient of 0.29, 0.26 and 0.24 are obtained for *A1-A3, A2-A3* and *A1-A2* inter-rater agreement respectively where eigenvector is the reference method. Even though most $\kappa \leq 0.4$, all the co-efficient are significant at $p<0.05$. The 95% confidence interval for all κ indicate that an upper bound of $\kappa \geq 0.40$ can be reached to make all κ fall within the range of fair/good agreement. The overall results indicate that keyword extraction is a difficult task since the experts have difficulty to agree on ranking the content words for the complex sentences. Table 9 summarizes the overall results of the Kappa coefficient for inter-rater agreement..

Many possible human factors can account for the low inter-rater kappa coefficient. A few inferences on how the experts carry out the keyword indentification task can be postulated for the explaination.

(a) each expert uses a different mental model to analyze and comprehend text,
(b) each expert uses a different strategy to identify keywords,
(c) each expert uses a different context to interprete keywords.

These are variables that need furhter research to investigate their correlation with keyword identification tasks.

The analysis of the results also showed that:

(a) experts emphasize more on nouns as the top ranked words;
(b) experts view context of top ranked words from the perspective of nouns;
(c) proper nouns are key event players for the experts.

These findings could provide useful inputs to propose an automated method that take into consideration human experts' views to induce content words from complex sentences. It would be of interest to gain some insights from these researches which could shed lights that guide automated keyword extraction.

One drawback of sliding window is that the linear word proximity approach actually has ignored the linguistic criteria of the words (Seretan & Wehrli 2007). Sliding window methods can be enhanced by integrating it with linguistic analysis, for example dependency grammar to detect word co-occurrence in research by Valle and Ozturk (2011). Not only extracting good word representation for vertices of the graph is important, it is also important to formulate better methods to represent the edge and weights of the graph. Another direction of future research from this paper is to investigate alternative approaches other than sliding window to improve the performance of inducing keywords for ranking.

Table 9. Overall Kappa Coeffiecient for Inter-Rater Agreement

Sliding Window				
Findings	**Kappa Coefficient**	**Reference Method**	**Standard Error (SE)**	**Confidence Interval (95% CI)**
A1 vs A2	0.2412	Eigenvector	0.099	0.047-0.44
	0.4048	Pagerank	0.093	0.22-0.59
	0.3018	Degree	0.1	0.1-0.5
A1 vs A3	0.2946	Eigenvector	0.099	0.1-0.49
	0.2677	Pagerank	0.096	0.08-0.45
	0.2201	Degree	0.1	0.021-0.42
A2 vs A3	0.2630	Eigenvector	0.099	0.07-0.46
	0.4143	Pagerank	0.1	0.21-0.61
	0.2559	Degree	0.099	0.061-0.45

A1, A2, A3 = expert 1, 2, 3 respectively

8.0 Conclusion

It has been reviewed earlier on that automated keyword extraction is a complex task in terms of the various resources, techniques and domain knowledge that are involved. This paper demonstrated that keyword extraction is equally a complex task when it is carried out by human even the experts. A total of 100 complex sentences are prepared as the data for three experts to carry out the task. The experts' keyword identification output is compared to a baseline which uses sliding window of size $n=5$ employing on graph-based approach. The overall results indicate that keyword extraction is a difficult task since the experts have difficulty to agree on ranking the content words for the complex sentences. The disadvantages of this approach are that it takes time to prepare the data and experts need time to annotate the data. Due to these factors, the amount of data that can be prepared is limited. This might affect the accuracy of results obtained in this research. As conclusion, this study has provided empirical information that open up more possibilities for future research on better automated keyword extraction methods. This can help to overcome the shortcoming of current state-of-the-art keyword extraction approach as well as the complexities faced by human in doing such task.

References

Chali, Y., Hasan, S. A. & Imam, K. (2011). Using semantic information to answer complex questions. *Canadian AI 2011, LNAI 6657,* 68-73.

Cheng, K., Li, Y. and Wang, X. (2013). Single document summarization based on triangle analysis of dependency graphs. International Conference on Network-Based Information Systems, 38-43.

Church, K. & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics, 16* ,22-29.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics.* , 61-74.

Erkan, G. & Radev, D. R. (2004). LexRank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* , 457-479.

Freeman, L. (1978). Centrality in social networks conceptual clarification. *Social Networks* , 215–239.

Han, J., Kim, T. & Choi, J. (2007). Web document clustering by using automatic keyphrase extraction. In Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, 56-59.

Hassan, S., Mihalcea, R. & Banea, C. (2007). Random-walk termweighting for improved text classification. *Semantic Computing, ICSC 2007,* 242-249.

Hsu, L. M. & Field, R. (2003). Interrater agreement measures: Comments on kappa, Cohen's kappa, Scott's and Aickin's . *Understanding Statistics* , 205-219.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge: MIT Press.

Jespersen, O. (1920). *The Philosophy of Grammar.* Allen and Unwin,.

Kraemer, H. C. (1979). Ramifications of a population model k for as a coefficient of reliability. *Psychometrika* , 461-472.

Lioma, C. & Blanco, R. (2009). Part of speech based term weighting for information retrieval. ECIR '09 Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, 412 - 423. Springer-Verlag Berlin, Heidelberg.

Litvak, M., Last, M., Aizenman, H., Gobits, I. & Kandel, A. (2011). DegExt – A Language-Independent Graph-Based Keyphrase Extractor. *Advances in Intelligent and Soft Computing* , 121-130.

Liu, J. & Wang, J. (2007). Keyword extraction using language network. International Conference on Natural Language Processing and Knowledge Engineering, 2007. NLP-KE 2007, 129-134.

Liu, J., Wang, J. & Wang, C. (2008). A text network representation model. Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 150-154. IEEE Computer Society.

Mihalcea, R. & Radev, D. (2011). *Graph-based Natural Language Processing Information Retrieval.* New York: Cambridge University Press.

Mihalcea, R. & Tarau, P. (2004). TextRank: Bringing order into texts. Proceedings of EMNLP, 404-411.

Mihalcea, R., Tarau, P. & Figa, E. (2004). Pagerank on semantic networks, with application to word sense disambiguation. The 20th International Conference on Computational Linguistics (COLING 2004), 1126-1132. Geneva, Switzerland.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM. 38(11)*, 39-41.

Palshikar, G. (2007). Keyword Extraction from a Single Document Using Centrality Measures. In A. Ghosh, R. De, & S. Pal, *PReMI 2007, LNCS 4815,* 503–510. Springer-Verlag Berlin Heidelberg.

Sayyadi, H., Hurst, M. & Maykov, A. (2009). Event detection and tracking in social streams. Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2009). *AAAI*, 311-314.

Seretan, V. & Wehrli, E. (2007). Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation, 43(1),* 71-85.

Siaw, N.H. (2015). Semantic Event Extraction In Unstructured Text Based On Prominence And Discourse-Level Dependencies, PhD. Thesis. Universiti Malaysia Sarawak, Malaysia.

Sigman, M. & Cecchi, G. (2002). Global organization of the wordnet lexicon. Proceedings of the National Academy of Sciences of the United States of America, 1742-1747.

Tomita, J., Nakawatase, H. & Ishii, M. (2004). Graph-based text database for knowledge discovery. 13th International World Wide Web Conference (WWW2004), 454-455.

Tsatsaronis, G., Varlamis, I. & Nørvag, K. (2010). An experimental study on unsupervised graph-based word sense disambiguation. *Computational Linguistics and Intelligent Text Processing,* 184-198.

Valle, K. (2011). *Graph-Based Representations for Textual Case-Based Reasoning.* Master Thesis: Norwegian University of Science and Technology.

Valle, K. & Ozturk, P. (2011). Graph-based representations for text classification. *India-Norway Workshop on Web Concepts and Technologies, Trondheim, Norway* .

Wan, X. & Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. The Twenty-Third AAAI Conference on Artificial Intelligence, 855-860.

Wang, W., Do, D. B. & Lin, X. (2005). Term graph model for text classification. *ADMA 2005, LNAI 3584,* 19-30.

Warrens, M. J. (2013). A Comparios of Cohen's Kappa and agreement coefficients by Corrado Gini. *IJRRAS*, 345-351.

Zha, H. (2002). Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In Proceedings of SIGIR 2002, 113-120.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin,* 374-378.

EVALUATION TECHNIQUES FOR CONTEXT-AWARE RECOMMENDER SYSTEMS: A SYSTEMATIC MAPPING

Khalid Haruna[1,2] and Maizatul Akmar Ismail[1*]
[1]Department of Information Systems
Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

[2]Department of Computer Science
Faculty of Computer Science and Information Technology, Bayero University, Kano, Nigeria

*Corresponding Author Email: *maizatul@um.edu.my*, Phone number: +60379676375

ABSTRACT

The purpose of this research is to perform a systematic mapping on the evaluation techniques and metrics for context-aware recommender systems (CARS) in order to assist novice and new researchers willing to enter the field for the first time. Published journals and conference proceedings papers from 2010 to 2016 are reviewed based on the methodology proposed by Kitchenham in performing a systematic mapping. The review identifies the merits and demerits of each evaluation technique and presented the various evaluation metrics for context-aware recommender systems. The review reveals that user studies as an evaluation technique have been utilized more than any other technique and Precision and Recall are the most dominantly used evaluation metrics. It is intended for researchers to use this review as a starting point for further advancement, as well as an exploration of other approaches that received little or no attention.

KEYWORD: *Context-aware, Recommender System, Evaluation Techniques, Evaluation Metrics, Systematic Mapping.*

1.0 Introduction

Recommender system (RS) is a software tool and technique that provides the best suggestions for items and services to users, typically from a large information space (Haruna & Ismail 2016). The suggestions are mainly based on the user's interests and preferences among different alternatives, and then present those items and services to the user in a suitable manner. RS is becoming an appropriate tool for facilitating and accelerating the process of information seeking especially with the dramatic increase in big data (Haruna et al. 2017; Haruna et al. 2016). At its emergence in the 1990s, classical RS were predominantly used to predict users' interests (Adomavicius & Tuzhilin 2001). These approaches utilized the users and items as the set of entities to project the ratings that are either implicitly inferred by the system (Zeng et al. 2012) or are explicitly provided by the users (Codina et al. 2015).

$$R: User \times Item \rightarrow Rating$$

However, in the early 2000s, RS researchers extend the research in RS to leverage contexts in addition to the classical two-dimensional process in order to provide a better-personalized recommendation (Haruna et al. 2017; Panniello et al. 2014).

$$R: User \times Item \times Context \rightarrow Rating$$

Context is an all-round concept (Adomavicius & Tuzhilin 2011), which has been studied across various disciplines, including philosophy, linguistics, and cognitive and organizational science. In the late 1980s, computer science as a discipline embraced the concept of context mainly in ubiquitous computing and artificial intelligence (Hong et al. 2009).

Context has been studied across various disciplines and each discipline tends to provide its idiosyncratic view. 150 different views of context from different fields of studies have been presented and examined in (Bazire & Brézillon 2005). The authors concluded that it will be very much difficult to find a single definition of context that is unanimously satisfying all research disciplines. However, the most reported view of the context in the field of computer science is that of (Dey 2001), which viewed context as: "Any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves".

There have been many algorithmic approaches to context-aware recommender systems (CARS), each claiming to be superior for some purpose. However, identifying the best algorithm for a given purpose has proven challenging (Herlocker et al. 2004). This is due to the disagreement among researchers on which attributes should be measured, and on which metrics should be used for each attribute (Herlocker et al. 2004). Notwithstanding, three evaluation strategies are used in practice for the performance evaluation of recommender systems to prove the effectiveness of any proposed approach (Haruna et al. 2004). These are; online, offline and users study.

In this paper, we aim to explore the coverage of the different evaluation methods and metrics used for context-aware recommender systems. This will help specifically novice and new researchers in identifying the merit and demerit of each and also knowing when which is appropriate. The review is based on the guidelines presented in (Keele 2007; Kitchenham 2004; Okoli & Schabram 2010).

The rest of the paper is organized as follows: Our adopted methodology is presented in Section 2. Section 3 presents the results of the review, and finally, the conclusion is laid out in Section 4.

2.0 Methodology

As defined by (Keele 2007; Kitchenham 2004; Okoli & Schabram 2010), a systematic mapping is a three-phase process (see Figure 1) that aims at critically making an appraisal on the previous contributions based on certain research questions in order to draw attention to research gaps that need to be addressed.

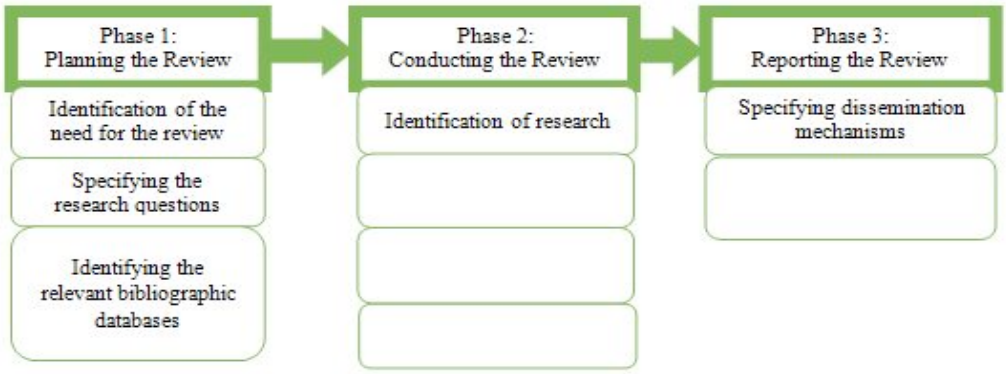


Figure 1. Phases and Activities of a Systematic Mapping

The following sub-sections described our methodology in conducting the systematic mapping as summarised by Figure 1.

2.1 Phase 1: Planning the Review

The first phase of a systematic mapping is planning the review in order to specify its objectives, identify the need and how the review will be conducted, it also justifies why the review is important. During the planning phase, the following activities have been conducted.

2.1.1 Identification of the Need for the Review

The aim of this paper is to undergo a systematic mapping on the different evaluation methods and metrics for CARS in order to assist new researchers to have background knowledge on what is needed and what previous researchers undergo in each process. Based on our knowledge, this paper is the first attempt to handle such a task, and hence the need for this review.

2.1.2 Specifying the Research Questions

Research questions derive the entire methodological process in any systematic mapping. In this paper, we aim to answer the question; *how has the context-aware recommender systems being evaluated?*

2.1.3 Identifying the Relevant Bibliographic Databases

In an attempt to perform an exhaustive search to provide a reasonable answer to the aforementioned question, the major bibliographical databases for computer science field have been identified. These are ScienceDirect, ACM, SpringerLink, IEEE and Web of

Science. As stated in (Stapić et al. 2012), setting a starting and closing date for a review is a rule and therefore, January 2010 and January 2016 was chosen to be the starting and closing dates, respectively for this review. The study covers all of the published journals and conference proceedings papers that are published within the stated period.

2.2 Phase 2: Conducting the Review

2.2.1 Identification and Selection of Primary Studies

The searching process of the research papers on CARS is based on the Boolean search criteria as depicted in Table 1. The two authors reviewed all of the selected papers based on title, abstract and conclusion, and included all papers that satisfy the following:

(a) A journal article that mentioned context and recommender in both the title and abstract.
(b) A conference proceeding paper that mentioned context and recommender in both the title and abstract, which had citations.
We also exclude any:

(a) Paper that is not written in the English language
(b) Master's and doctoral dissertations, textbooks, book chapters and news articles.
(c) Conference proceeding papers without citations.

2.2.2 Data Extraction and Synthesis

Table 1. Data Extraction Strategy

1st step	Retrieving papers using keyword Title (Context) AND (Recommender OR Recommendation)	1558 papers
2nd step	Reviewing papers based on title and abstract	187 papers
3rd step	Excluding surveys and reviews	183 papers
4th step	Excluding conference proceeding papers without citations	65 papers
5th step	Observing the references of the selected papers to include the missing ones that satisfy our inclusion criteria.	65 papers
6th step	Searching the selected papers in Google Scholar to retrieve more papers	68 papers

As shown in Table 1, a total of 68 papers satisfied for the final review, and each paper was prudently reviewed to identify the evaluation techniques and the metrics used.

2.2.3 Validity Control

To retrieve a maximum number of publications, all from the list of references of the 65 selected papers from step 4 were reviewed and compared with the local databases in order to find any missing paper that may satisfy the inclusion requirement, but which are not included. Fortunately, not a single paper was found that is not included. Each of the selected papers has also been searched in Google Scholar and had a brief overview over the title of all the papers in 'cited by', 'related articles' and 'all versions'. Three papers were found and added making a total of 68 selected papers.

2.3 Phase 3: Reporting the Review

Phase 3 is the final phase of the systematic mapping, it involves reporting the outcome of the review in a presentable manner for dissemination channels and target audiences. The following section presents the results of the systematic mapping to respond to the question posed in section 2.1.2.

3.0 Results and Discussions

Researchers in the field of recommender system had taken the challenge raised by some other researchers such as (Adomavicius & Tuzhilin 2001) that the current situation of a user seeking for recommendation needs to be taken into account at the spot of recommendations. This can easily be seen from the trend of publications depicted by Figure 2, that the publications related to context-aware recommender systems are increasing in an arithmetical progression. This shows that the field is active and is attracting the interest of many researchers and practitioners. Henceforth, doing a review that will serve as a guideline especially to novice and new researchers in the field is important.

Figure 2. Distribution by Year of Publication

Zero (0) papers are retrieved in 2016 because the review range ends in January 2016. Nevertheless, as can be seen, and asserted from the publication trend, many papers may be published at the end of the year.

As have been pointed out earlier and from the results of the review, three evaluation strategies are used for evaluating context-aware recommender systems. These are users study, offline and online evaluation techniques. Each of the techniques utilized several metrics for measuring the performance and the effectiveness of a proposed method. Figure 3 presents the results of the systematic mapping and in the following section, the merits and demerits of each technique are detailed.

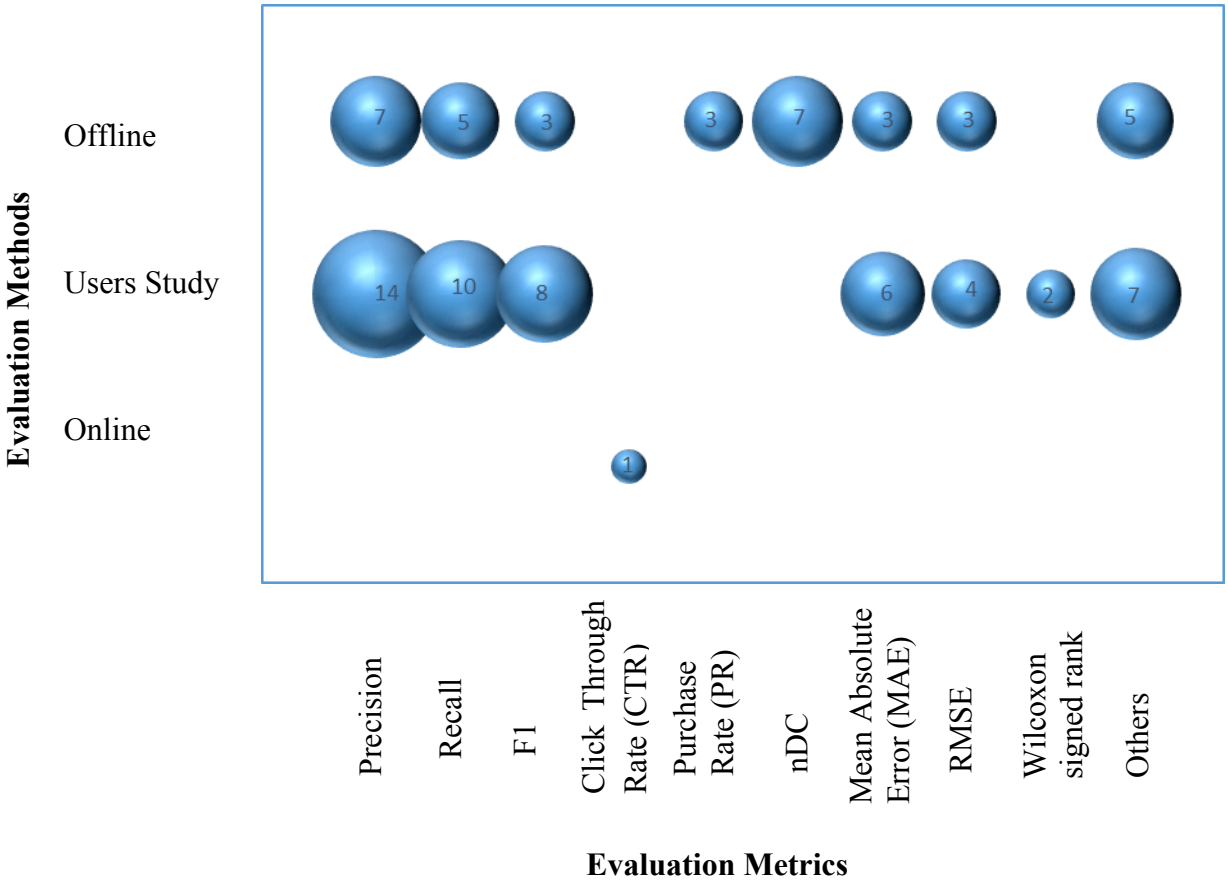


Figure 3. Bubble Map of Evaluation Techniques for CARS

3.1 Users Study

Users study is an evaluation technique that measures the users' satisfaction through explicit ratings. In users study, the generated recommendations are rated by the users, and the approach with the highest average rating is considered most effective (Kapoor et al. 2007).

As can be seen from Figure 3, users study has been widely used based on the reviewed literature. This is because researchers consider users study to be the most optimal evaluation method (Said et al. 2012). However, the results often depend on the questions asked. The authors of (Cremonesi et al. 2012) observed results discrepancies when subjects are asked for the perceived relevance or the global satisfaction of the recommendation. One major drawback of users study is the need for a large number of participants, which makes it relatively expensive to conduct.

3.2 Online Evaluations

Online evaluation is an evaluation technique that measures the recommendations acceptance rates in the real world. Online evaluation is considered the best method for evaluating context-aware recommender systems (Beel & Langer 2014). However, conducting an online evaluation requires a significant amount of time. In addition, it is the most expensive technique and can only be conducted by researchers who have access to the real-world recommender systems. It can be noted that only one research based on the reviewed literature evaluated their developed system online.

The performance of online evaluation is usually measured by click-through rates (CTR). However, the authors of (Zheng et al. 2010) argued that relevance and CTR do not always correlate and concluded that "CTR may not be the optimal metric for online evaluation of recommender systems" and "CTR should be used with precaution"

3.3 Offline Evaluations

Offline evaluation technique, on the other hand, is the easiest and the most convenient method of evaluation, as it requires no interaction with the real users (Beel & Langer 2014). It measures the recommendation's accuracy based on the ground truth. Offline evaluation is aimed to identify some promising recommendation approaches (Dong et al. 2009; Kapoor et al. 2007), which will then be evaluated in details with online or user studies. However, the use of offline evaluation technique has been criticized by many researchers (Cremonesi et al. 2011; Cremonesi et al. 2012; McNee et al. 2002). This is because offline evaluation technique does not put into consideration the human factors that advertently affect the overall users' satisfaction. Notwithstanding, as can be seen from the results of the review, offline evaluation is among the predominant evaluation methods in the recommender community.

3.4 Evaluation Metrics

Depending on the evaluation approach being used, certain metrics determine the effectiveness of a recommender system. Table 2 summarizes the common metrics used and the condition and the merits for each. Precision and Recall are the most used evaluation metrics.

Table 2. Evaluation Metrics

Evaluation Metric	Condition	Merits	References
Precision	To measure the proportion of retrieved documents that are relevant.	It measures the system's ability to reject any non-relevant document in the retrieved set.	(Alhamid, Rawashdeh, Al Osman, Hossain, & El Saddik, 2015; Alhamid, Rawashdeh, Dong, et al., 2015; Bagci & Karagoz, 2015; Chan, Chiu, & Yu, 2011; Codina et al., 2015; Colombo-Mendoza, Valencia-García, Rodríguez-González, Alor-Hernández, & Samper-Zapater, 2015; Gantner, Rendle, & Schmidt-Thieme, 2010; Gedikli & Jannach, 2013; Han, Schmidtke, Xie, & Woo, 2014; Hariri, Mobasher, & Burke, 2012; L. Hong et al., 2015; Hussein, Linder, Gaulke, & Ziegler, 2014; Panniello et al., 2014; Yan, Guo, & Cheng, 2011; Yang, Long, Smola, Zha, & Zheng, 2011; Yuan, Cong, Zhao, Ma, & Sun, 2015; Y. Zheng, Mobasher, & Burke, 2014)
Recall	To measure the proportion of relevant documents that are retrieved.	It measures the system's ability to find all the relevant documents.	(Alhamid, Rawashdeh, Dong, et al., 2015; Bagci & Karagoz, 2015; Colombo-Mendoza et al., 2015; Han et al., 2014; Hariri et al., 2012; Hidasi, 2015; Hidasi & Tikk, 2015; Panniello et al., 2014; Yan et al., 2011; Yang et al., 2011; Yuan et al., 2015; Y. Zheng et al., 2014)
F-measure	To measure the harmonic mean of Recall and Precision	It measures the balance between precision and recall.	(Bagci & Karagoz, 2015; Colombo-Mendoza et al., 2015; Panniello et al., 2014; Zeng et al., 2012)
Click Through Rate (CTR)	To measure the number of recommendations	It provides a broad view of how well an item or a	(Shi, Ghedira, & Marini, 2015)

	end up being clicked	campaign attracts users.	
Purchase Rate (PR)	To measure the number of recommendations end up being purchased	It measures the number of times an item being purchased.	(Bouneffouf, Bouzeghoub, & Gancarski, 2012; Shi et al., 2015)
Discounted Cumulative Gain (nDCG)	To measure the top-k recommendation quality	It ignores recall and focuses on the precisions at the top of the ranks	(Biancalana, Gasparetti, Micarelli, & Sansonetti, 2013; Codina et al., 2015; L. Hong et al., 2015; Ruiz-Iniesta, Jimenez-Diaz, & Gómez-Albarrán, 2014; Yang et al., 2011)
Mean Absolute Error (MAE)	To measure the accuracy of rating predictions	It measures the deviation of recommendations from the user-specified ratings.	(Codina et al., 2015; Gedikli & Jannach, 2013; L. Hong et al., 2015; Karatzoglou, Amatriain, Baltrunas, & Oliver, 2010; Liu & Aberer, 2013)
Root Mean Squared Error (RMSE)	To measure the accuracy of rating predictions	It gives a sense of how close a predicted value is from the actual data.	(Gedikli & Jannach, 2013; L. Hong et al., 2015; Liu & Aberer, 2013)

4.0 Conclusion

This paper presented a systematic mapping on the evaluation techniques and metrics for context-aware recommender systems from 2010 to 2016. The review revealed that three evaluation approaches are used for measuring the performance and the effectiveness of a context-aware recommender system. These are users study, offline and online evaluations. The paper has identified the merits and demerits of each approach and presented in the form of a table the various evaluation metrics used in which Precision and Recall are dominantly used. Users study as an evaluation approach has also been utilized the most than any other evaluation method.

References

Adomavicius, G., & Tuzhilin, A. (2001). *Extending recommender systems: A multidimensional approach.* Paper presented at the Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-01), Workshop on Intelligent Techniques for Web Personalization (ITWP2001), Seattle, Washington, August.

Adomavicius, G., & Tuzhilin, A. (2011). Context-aware recommender systems *Recommender systems handbook* (pp. 217-253): Springer.

Alhamid, M. F., Rawashdeh, M., Al Osman, H., Hossain, M. S., & El Saddik, A. (2015). Towards context-sensitive collaborative media recommender system. *Multimedia Tools and Applications, 74*(24), 11399-11428.

Alhamid, M. F., Rawashdeh, M., Dong, H., Hossain, M. A., Alelaiwi, A., & El Saddik, A. (2015). RecAm: a collaborative context-aware framework for multimedia recommendations in an ambient intelligence environment. *Multimedia Systems*, 1-15.

Bagci, H., & Karagoz, P. (2015). Context-aware location recommendation by using a random walk-based approach. *Knowledge and Information Systems*, 1-20.

Bazire, M., & Brézillon, P. (2005). Understanding context before using it *Modeling and using context* (pp. 29-40): Springer.

Beel, J., & Langer, S. (2014). Research Paper Recommender Systems: A Literature Survey. *Under Review. Pre-print available at http://www. docear. org/publications*.

Biancalana, C., Gasparetti, F., Micarelli, A., & Sansonetti, G. (2013). An approach to social recommendation for context-aware mobile services. *ACM Transactions on Intelligent Systems and Technology (TIST), 4*(1), 10.

Bouneffouf, D., Bouzeghoub, A., & Gancarski, A. L. (2012). *Following the user's interests in mobile context-aware recommender systems: The hybrid-E-greedy algorithm.* Paper presented at the Advanced Information Networking and Applications Workshops (WAINA), 2012 26th International Conference on.

Chan, W. K., Chiu, Y. Y., & Yu, Y. T. (2011). A web search-centric approach to recommender systems with URLs as minimal user contexts. *Journal of Systems and Software, 84*(6), 930-941.

Codina, V., Ricci, F., & Ceccaroni, L. (2015). Distributional semantic pre-filtering in context-aware recommender systems. *User Modeling and User-Adapted Interaction*, 1-32.

Colombo-Mendoza, L. O., Valencia-García, R., Rodríguez-González, A., Alor-Hernández, G., & Samper-Zapater, J. J. (2015). RecomMetz: A context-aware knowledge-based mobile recommender system for movie showtimes. *Expert Systems with Applications, 42*(3), 1202-1222.

Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A. V., & Turrin, R. (2011). *Looking for "good" recommendations: A comparative evaluation of recommender systems.* Paper presented at the IFIP Conference on Human-Computer Interaction.

Cremonesi, P., Garzotto, F., & Turrin, R. (2012). Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. *ACM Transactions on Interactive Intelligent Systems (TiiS), 2*(2), 11.

Dey, A. K. (2001). Understanding and using context. *Personal and ubiquitous computing, 5*(1), 4-7.

Dong, R., Tokarchuk, L., & Ma, A. (2009). *Digging friendship: paper recommendation in social network.* Paper presented at the Proceedings of Networking & Electronic Commerce Research Conference (NAEC 2009).

Gantner, Z., Rendle, S., & Schmidt-Thieme, L. (2010). *Factorization models for context-/time-aware movie recommendations.* Paper presented at the Proceedings of the Workshop on Context-Aware Movie Recommendation.

Gedikli, F., & Jannach, D. (2013). Improving recommendation accuracy based on item-specific tag preferences. *ACM Transactions on Intelligent Systems and Technology (TIST), 4*(1), 11.

Han, J., Schmidtke, H. R., Xie, X., & Woo, W. (2014). Adaptive content recommendation for mobile users: Ordering recommendations using a hierarchical context model with granularity. *Pervasive and Mobile Computing, 13*, 85-98.

Hariri, N., Mobasher, B., & Burke, R. (2012). *Context-aware music recommendation based on latenttopic sequential patterns.* Paper presented at the Proceedings of the sixth ACM conference on Recommender systems.

Haruna, K., Akmar Ismail, M., Damiasih, D., Sutopo, J., & Herawan, T. (2017). A collaborative approach for research paper recommender system. *PloS one, 12*(10), e0184516.

Haruna, K., & Ismail, M. A. (2016). An Ontological Framework for Research Paper Recommendation. *International Journal of Soft Computing, 11*(2), 96-99.

Haruna, K., Ismail, M. A., Bichi, A. B., Danjuma, S., Kakudi, H. A., & Herawan, T. (2017). An Overview on the State-Of-The-Art Comparison of the Three-Contextualization Paradigms. *The 8th International Conference on Information Technology (ICIT 2017), May 17-18, 2017 Amman, Jordan*.

Haruna, K., Ismail, M. A., & Shuhidan, S. M. (2016). Domain of Application in Context-Aware Recommender Systems: A Review. *Knowledge Management International Conference (KMICe) 2016, 29 –30 August 2016, Chiang Mai, Thailand*.

Haruna, K., Ismail, M. A., Suhendroyono, S., Damiasih, D., Pierewan, A. C., Chiroma, H., & Herawan, T. (2017). Context-Aware Recommender System: A Review of Recent Developmental Process and Future Research Direction. *Applied Sciences*.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS), 22*(1), 5-53.

Hidasi, B. (2015). *Context-aware Preference Modeling with Factorization.* Paper presented at the Proceedings of the 9th ACM Conference on Recommender Systems.

Hidasi, B., & Tikk, D. (2015). Speeding up ALS learning via approximate methods for context-aware recommendations. *Knowledge and Information Systems*, 1-25.

Hong, J.-y., Suh, E.-h., & Kim, S.-J. (2009). Context-aware systems: A literature review and classification. *Expert Systems with Applications, 36*(4), 8509-8522.

Hong, L., Zou, L., Zeng, C., Zhang, L., Wang, J., & Tian, J. (2015). Context-Aware Recommendation Using Role-Based Trust Network. *ACM Transactions on Knowledge Discovery from Data (TKDD), 10*(2), 13.

Hussein, T., Linder, T., Gaulke, W., & Ziegler, J. (2014). Hybreed: A software framework for developing context-aware hybrid recommender systems. *User Modeling and User-Adapted Interaction, 24*(1-2), 121-174.

Kapoor, N., Chen, J., Butler, J. T., Fouty, G. C., Stemper, J. A., Riedl, J., & Konstan, J. A. (2007). *Techlens: a researcher's desktop.* Paper presented at the Proceedings of the 2007 ACM conference on Recommender systems.

Karatzoglou, A., Amatriain, X., Baltrunas, L., & Oliver, N. (2010). *Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering.* Paper presented at the Proceedings of the fourth ACM conference on Recommender systems.

Keele, S. (2007). Guidelines for performing systematic literature reviews in software engineering *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University, 33*(2004), 1-26.

Liu, X., & Aberer, K. (2013). *SoCo: a social network aided context-aware recommender system.* Paper presented at the Proceedings of the 22nd international conference on World Wide Web.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., . . . Riedl, J. (2002). *On the recommending of citations for research papers.* Paper presented at the Proceedings of the 2002 ACM conference on Computer supported cooperative work.

Okoli, C., & Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research. *Available at SSRN 1954824*.

Panniello, U., Tuzhilin, A., & Gorgoglione, M. (2014). Comparing context-aware recommender systems in terms of accuracy and diversity. *User Modeling and User-Adapted Interaction, 24*(1-2), 35-65.

Ruiz-Iniesta, A., Jimenez-Diaz, G., & Gómez-Albarrán, M. (2014). A Semantically Enriched Context-Aware OER Recommendation Strategy and Its Application to a Computer Science OER Repository. *IEEE Transactions on Education, 57*(4), 255-260.

Said, A., Tikk, D., Shi, Y., Larson, M., Stumpf, K., & Cremonesi, P. (2012). *Recommender systems evaluation: A 3D benchmark.* Paper presented at the ACM RecSys 2012 workshop on Recommendation utility evaluation: beyond RMSE, Dublin, Ireland.

Shi, F., Ghedira, C., & Marini, J.-L. (2015). Context Adaptation for Smart Recommender Systems. *IT Professional, 17*(6), 18-26.

Stapić, Z., López, E. G., Cabot, A. G., de Marcos Ortega, L., & Strahonja, V. (2012). *Performing systematic literature review in software engineering.* Paper presented at the Proceedings of the Central European Conference on Information and Intelligent Systems.

Yan, X., Guo, J., & Cheng, X. (2011). *Context-aware query recommendation by learning high-order relation in query logs.* Paper presented at the Proceedings of the 20th ACM international conference on Information and knowledge management.

Yang, S.-H., Long, B., Smola, A. J., Zha, H., & Zheng, Z. (2011). *Collaborative competitive filtering: learning recommender using context of user choice.* Paper presented at the Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.

Yuan, Q., Cong, G., Zhao, K., Ma, Z., & Sun, A. (2015). Who, Where, When, and What: A Nonparametric Bayesian Approach to Context-aware Recommendation and Search for Twitter Users. *ACM Transactions on Information Systems (TOIS), 33*(1), 2.

Zeng, C., Jia, D., Wang, J., Hong, L., Nie, W., Li, Z., & Tian, J. (2012). *Context-aware social media recommendation based on potential group.* Paper presented at the Proceedings of the 1st International Workshop on Context Discovery and Data Mining.

Zheng, H., Wang, D., Zhang, Q., Li, H., & Yang, T. (2010). *Do clicks measure recommendation relevancy?: an empirical user study.* Paper presented at the Proceedings of the fourth ACM conference on Recommender systems.

Zheng, Y., Mobasher, B., & Burke, R. (2014). *Context recommendation using multi-label classification.* Paper presented at the Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on.

SOFTWARE ENGINEERING APPROACH FOR DOMAIN ONTOLOGY DEVELOPMENT: A CASE STUDY OF ISLAMIC BANKING PRODUCT

Ahmad Shaharudin Abdul Latiff[1], Haryani Haron[2] and Muthukkaruppan Annamalai[3]
Universiti Malaysia Kelantan
Kota Bharu, Kelantan
[1]shaharudin@umk.edu.my

Universiti Teknologi MARA
Shah Alam, Selangor
[2]haryani@tmsk.uitm.edu.my, [3]mk@tmsk.uitm.edu.my

ABSTRACT

In order to be effective, information systems that cater for the need of a particular domain ought to deploy suitable ontology of the industry. The ontology will be useful for their general operations, decision making and also to facilitate the financial engineering. The software engineering approach develop ontologies that are more reliable, long lived, and continually adapted. It also enables the ontology to adapt to the context of its purpose. The Software Engineering Approach for Domain Ontology Development adapts the software engineering methodology to attend to domain problems. It conducts root cause analysis to identify the characteristics and ontology development criteria for specified domains. In the context of this paper which is designing the ontology for Islamic banking product; the peculiarities of the Islamic banking businesses demand several crucial characteristics to be taken into account in the development of the ontology. The characteristics relate to the business model, diverse roles, distinct concepts, representation rules, reuse of other ontologies and varieties of disciplines related to the industry. The software engineering approach guides the conduct of root cause analysis to understand the issues faced by the domain to be put up as requirement specification of the ontology.

KEYWORD: *Ontology, Ontology Development, Software Engineering, Islamic Banking*

1.0 Introduction

Domain ontologies are related to particular domains, and consist of concepts used to describe the domain knowledge and the relationships between those concepts (Hadzic, Wongthongtham, Dillon & Chang 2009; Yu 2011). Development of domain ontologies is not an easy and straightforward task; the team of knowledge engineers and domain experts assigned to develop to develop the ontologies must ensure that the concepts in the domain of discourse, their properties and relations are accurate. As complexities of the world increases, there is need to study the particular aspects that are peculiar to the specified domains. Hence, the approach undertaken for the development of the ontology should be emphasised.

There are many methods of developing ontologies; they could be built using the traditional Ontology Development Life Cycle (ODLC) such as TOVE (Gruninger & Fox 1995), ENTERPRISE (Uschold & King 1995) METHONTOLOGY (Fernandez-Lopez et al. 1997), and DILIGENT (Pinto et al. 2004) or by adopting a software engineering approach. Software engineering was not initially meant to cater for software artefacts such as the ontology. Yet, to develop domain ontologies that are reliable, long lived, and continually adapted and improved; the software engineering approach would be

recommended (Annamalai & Rosli 2012). There are several methods of software engineering that has been adapted for the development of ontologies.

During this period of economic volatility, it is important for domain ontologies to meet the specific needs of the businesses of the domain and attend to the prevailing issues and problems faced by them. The domain ontologies should be specifically built to attend to prevailing issues faced by the specified domains. Apparently there were grievances against the Islamic banks that relates to the customers' understanding of the Islamic banking products and processes (Latiff et al. 2015), and the use of appropriate ontology by information systems is deemed as an avenue for improving knowledge dissemination of the domain. The Sofware Engineering Approach for Domain Ontology Development was adopted to develop the ontology for the Islamic banking domain.

This paper looks at the Software Engineering Approach for Domain Ontology Development. In section 2, several methods of software engineering approach for ontology development are discussed to explain the need for a software engineering approach for domain ontology development. Section 3 provides the methodology which describes the phases of the Software Engineering Approach for Domain Ontology Development. Section 4 discusses the ontology produced by the approach and the final section summarises the outcome of ontology development method.

2.0 Related Works

Software engineering put emphasis on systematic processes and standards, integrative tools, and established management and organisational methods in software development (Annamalai & Rosli 2012). It is important to take note that software is not confined to computer programs, but includes all associated documentation and configuration data that is required for the programs to operate correctly. Ontology is not excluded as it defines the representational terms of the data (Gruber 1993b). Being software artefacts, ontologies are best to be developed using the software engineering approach like any other software components in order for the ontologies to be reliable, long lived, and continually adapted (Annamalai & Rosli 2012). Nevertheless, there is need to adapt the software engineering approach for ontology development to take into account the peculiarities of the characteristics and criteria of the domain involved.

There have been several approaches that have been undertaken by ontology developers to adapt software engineering approaches for ontology development; among them are as follows:

a) Unified Process for ONtology (UPON)

The UPON is an approach for large-scale ontology development that adopts the Unified Process (UP) and the Unified Modeling Language (UML) in its development (De Nicola et al. 2009). It is not meant for generic domain ontologies, but is use-case driven and aims at developing well-defined application areas. Following the UP approach, the UPON consists of cycles, phases, iterations and workflows. Every cycles are made of inception, elaboration, construction and transition phases. The phases are subdivided into iterations and each iterations comprise of requirements,

analysis, design, implementation and test workflows. The workflows for the UPON methodology are as below:

i) *The requirements workflow* - which specifies the semantic needs and user view of the knowledge to be encoded in the ontology. It involves the determination of the domain of interest and the scope of the ontology, defining it's business purpose, writing one or more storyboards, creating application lexicons, identifying competency questions and modeling use cases.

ii) *The analysis workflow* - which refines and structures the ontology requirements. This workflow acquires domain resources and builds domain lexicon, builds the reference lexicon, models the application scenario using UML diagrams, and builds the reference glossary.

iii) *The design workflow* - which gives an ontological structure to the set of glossary entries gathered in the reference glossary. It involves the tasks of modeling concepts into primary categories (business actor, business object and business process) and also complementary categories, and then model concept hierarchies and domain-specific relationships.

iv) *The implementation workflow* – which encodes the ontology into a rigorous formal language. In selecting the formal language, it is necessary to consider it's expressive power, computational complexity and level of acceptance.

v) *The test workflow* – which tests the ontology to verify the semantic and pragmatic quality of the ontology. The semantic quality concerns the absence of contradictory concepts, which the pragmatic quality refers to the ontology content and it's usefulness to the users.

b) Enterprise Strength Ontology Engineering (EsOE)

The Enterprise Strength Ontology Engineering (EsOE) by Annamalai and Rosli (2012) adopts the value-added activities of Rational Unified Process (RUP) model, Agile model, Capability Maturity Model Integration (CMMI) staged model and IEEE 1074-1995 standard (IEEE) to produce a project-oriented software engineering approach for ontology development. The methodology is structured into three levels of realization as below:

i) *Level I Engineering process* - which focuses on the development processes. The processes involved at this level are problem exploration, infrastructure allocation, requirements management, design, development, testing, deployment, maintenance and retirement of the ontology.

ii) *Level II Project-focus process* - which concerns the management of the Engineering. The processes involved are project management, risk management,

supplier agreement management, configuration management, product and process quality assurance, and documentation.

iii) *Level III Organisation-focus process* - which establishes the capabilities of the performed activities. This level comprises of the processes of training, organisational process definition, process measurement and analysis, and also organisational process improvement.

c) Simplified Agile Methodology for Ontology Development (SAMOD)

Like the software engineering domain, agile methodologies were also proposed for ontology engineering. Peroni (2016) introduced the Simplified Agile Methodology for Ontology Development (SAMOD), which is partially inspired by the Test-Driven Development process in Software Engineering. The methodology that targets the development of ontologies with limited amount of ontological entities is organised by the following iterative steps:

i) *Step 1: define a new test case*. Ontology engineers and domain experts work together to collect information and write motivating scenarios of the specified domain. With the motivating scenarios, sets of informal competency questions will be identified. Modelets will be developed according to the scenarios, informal competency questions and the glossary of terms.

ii) *Step 2: merge the current model with the modelet*. The modelets will be merged by adding all the axioms and then collapsing semantically-identical entities. The models will then be test queried according to their formal requirements.

iii) *Step 3: refactor the current model*. Finally the model is refactored and tested. In doing so, the ontology engineers should reuse existing knowledge, document the ontology and take advantages from technologies.

d) Lightweight Methodology for Rapid Ontology Engineering (UPON Lite)

The Lightweight Methodology for Rapid Ontology Engineering (UPON Lite) by De Nicola and Missikoff (2016) derived from the Unified Process for ONtology (UPON). With UPON Lite, ontologies are constructed by domain experts and ontology engineers only intervene to deliver final ontology formalisation. The domain content is elicited, organised and validated by domain experts. The methodology is organised as below:

i) *Step 1: Domain terminology*. This step involves the creation of domain-specific terminology to produce a domain lexicon. Resources such as textual documents, directories, dictionaries, taxonomies, standards and also ontologies could be use to develop the lexicon.
ii) *Step 2: Domain glossary*. Upon the production of the lexicon, this next step is to enrich it with textual description of each entry. An ontology structuring method

called OPAL (Object, Process, Actor modeling Language) is adopted to group the concepts into three main categories (i.e., object, process and actor) and also three auxiliary categories (i.e., complex, atomic and reference properties). Synonyms are also identified to pinpoint the preferred term and label the rests as other terms.

iii) *Step 3: Taxonomy*. The lexicon and glossary of the previous steps were used to build the taxonomy.

iv) *Step 4: Predication*. This step concentrates on the properties to characterise the entities of the domain.

v) *Step 5: Parthood (meronymy)*. This step looks at the architectural structure of business entities or parts of composite entities to create hierarches based on "part of" relationship.

vi) *Step 6: Ontology*. From the previous five steps, the ontology engineers could finally produced the ontology.

As compared to the ontology development methods above, the Software Engineering Approach for Domain Ontology Development is targeted for specified domains and focuses on the problems of the domain that requires the development of the ontology as a solution. It is aimed to attends to the prevailing problems of the specified domains.

3.0 Methodology

The Sofware Engineering Approach for Domain Ontology Development was adapted from the Level I (Engineering process) of the Enterprise Strength Ontology Engineering (EsOE). The phases of the domain ontology development adapted from the software engineering approach are as follows:

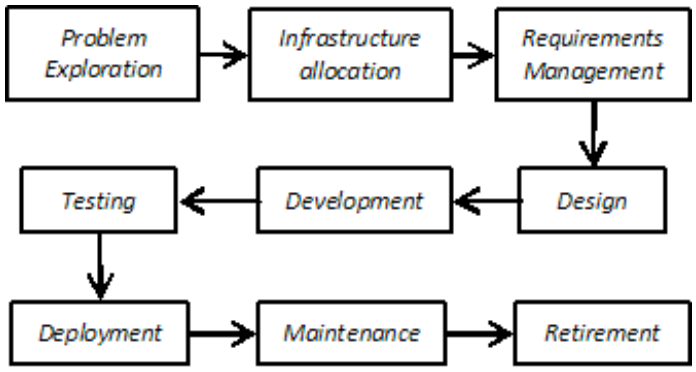


Figure 1. Ontology Development Based On Software Engineering Approach

In meeting the needs of the domain, it is important that the ontology developed be implicitly embedded with the purpose and reasons why it was designed. Hence, the activities in the problem exploration and requirements management phases are designed to identify the root causes of the problem and meet the purposes of the ontology. In order to do so, the problem exploration stage includes root cause analysis of the problem that the ontology intended to attend. The outcome of the root cause analysis shall be studied in order to produce the characteristics and development criteria of the ontology requirement specification. The ontology developed according to the defined specification would be deemed as fit to attend to the prevailing issues of the specified domain. In the case of the ontology development for the Islamic banking industry, the ontology was supposed to attend to the customers' grievances against the Islamic banks. The root cause analysis conducted produced the characteristics and development criteria of the Islamic banking ontology (Latiff et al. 2016)

3.1 Problem Exploration

The problem exploration which is equivalent to software specification or also referred to as requirements engineering is a process whereby the details of the services required from the software is being understood and defined (Sommerville 2011). The activities primarily transform the stakeholders' needs into ontology requirements; which include the process of requirements elicitation and analysis. During this process, the requirements of the ontology should be derived through observation of the problems of the domain that ignited the goals and purposes for the ontology development. The purpose and the intended use of the ontology should be made clear before the design and development of the ontology could take place (Uschold & Gruninger 1996). In order to do so, the intended users of the ontology ought to be identified beforehand and their characteristics be ascertained in order to clarify the purpose and scope of the ontology. It could be done through brainstorming, interviews, questionnaires, text analysis and inductive techniques (Pinto & Martins 2004). Besides identifying the users, the root causes of the problem should also be identified to clearly understand the services that the ontology would cater for and other pertinent matters related to it.

Although it is being unusual for root cause analysis to be conducted upfront during specification stages of software engineering as it was commonly meant to prevent software failures and attend to defects (IEEE 2006; Team 2010), it would be undeniably valuable to understand the problems and needs of the domain. As such, it would be advantageous that root cause analysis be conducted in building the specification for ontologies. Clear understanding of the root cause is crucial to comprehend the domain's problems and the result of root-cause analysis will in turn help to enhance the specification requirements of the ontology. It is only with clear identification of the users of the ontology and understanding of the problem that will make the requirements clearly elicited. A proper elicitation of requirements will facilitate analysis of them. Eventually, the analysis will enable the ontology to meet the user needs, which will in turn improve the quality of the software program and application which utilise the ontology.

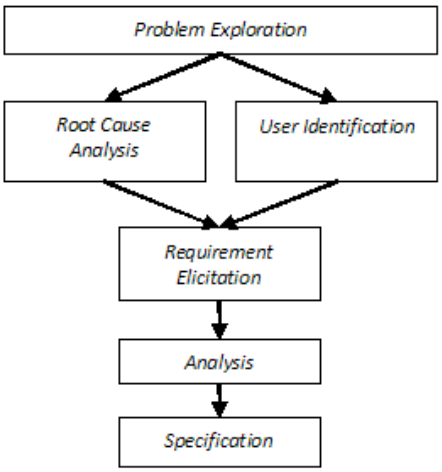


Figure 2. Problem Exploration Stage

The proposal for the development of the ontologies should also describe the motivating scenarios, together with the set of intended solutions to the scenario problems as recommended by Gruninger & Fox (1995). From the given scenario, a set of queries that need to be answered by the ontology could be raised. They become the informal competency questions that need to be expressed in the formal language of the ontology later.

3.2 Infrastructure Allocation

During this phase, the work environment for the ontology development is established. The project team is familiarised with the development environment, which include the processes and tools for knowledge engineering. For the development of the ontology for the Islamic banks, the processes and tools for knowledge engineering of the research, have been predetermined to be using open-source ontology editor, *Protégé* and the enterprise data unification platform *Stardog* respectively.

3.3 Requirements Management

The requirements management include the activities of managing the technical and non-technical requirements of the ontology, and also to identify and resolve the inconsistencies among the requirements (Annamalai & Rosli 2012). The technical requirements are properties or attributes of the ontology to be developed (Team 2010). Meanwhile, the non-technical requirements are requirements that affect the ontology development that are not properties of the product or service (Team 2010).

This stage comprises of:

a) Requirements specification, which is the activity that translates the information gathered during the analysis into a requirements document. It should be decided whether to produce either an informal, semi-formal or formal ontology specification document using intermediate representations (Fernandez-Lopez et al. 1997). The Software Requirement Specification Document for the ontology development was

slightly modified from those recommended for other software components (Suárez-Figueroa, Gómez-Pérez & Villazón-Terrazas 2009) as below:

Table 1. Ontology Requirement Specification

	Islamic Banking Product Ontology Requirement Specification Document
1.	Purpose
	The primary purpose of building the ontology is to provide information to enable the banks' customers to have sufficient understanding for their decision in selecting a product. The ontology could also useful to bring explicit information on Shariah concepts and assist bank staff in conceptualising a proposed new product and testing the features of the proposed product during product developments.
2.	Scope
	The ontology focuses on the domain of Islamic banking products of financings and deposits, especially for the personal/ consumer target market. It includes business premise financings, but excludes large corporate financing, trade financing, negotiable certificate of deposits, and other corporate-targeted products. As the ontology of the Islamic banking products is intended merely to assist customers in their decision-making, it need not be a formal ontology, although there should be some minimal rules and axioms to cater for the complexities of the Islamic banking knowledge.
3.	Implementation Language
	No specified formal representative language, but the Web Ontology Language (OWL) would be used for validation purposes.
4.	Intended End-Users
	i) Customers and prospective customers of the Islamic banks, ii) Staff of the bank who are involved in new product development.
5.	Intended Uses
	i) Publishing of Product Disclosure Sheet for disseminating of information on Islamic banking products, ii) Displaying product comparison in assisting the customers to select the Islamic banking products, iii) Displaying the features and testing new Islamic banking products during new product development.
6.	Ontology Requirements
	a. Non-Functional Requirements
	Characteristics of the Ontology The **general characteristics** of the ontology for Islamic banking are as follows: i) The concepts to be dealt with in the Islamic banking industry are distinct ii) The rules in the representation of the concepts in the Islamic banking system

	The **specific characteristics** of the ontology for Islamic banking are as follows: i) The business model or Shariah concept adopted by the Islamic banking product ii) The diverse roles played by the banks and their customers iii) The differences of opinion of different juristic schools iv) The borrowings of Arabic terminologies <u>*Development Criteria to be Fulfilled by the Ontology*</u> *i) Clarity* The clarity of the contents should be made just adequate for the customers to make effective comparison among the different products of the Islamic banks. *ii) Concise* The ontology should emphasise on brevity and avoids redundant definitions and explanations that could be counter-productive and might end up confusing the customers. *iii) Coherence* *a) Individual Conceptual Integrity* The individual conceptual integrity need to be verified to ensure that there is no contradiction in the interpretation of a concrete concept pertaining to the entities it represents. *b) Collective Consistency* The collective consistency should be analysed to verify that the relationships that bind the concepts reflect the dependencies between their corresponding entities. *iv) Extensibility and Expansibility* The designed ontology need to be ensured as much as possible that the hierarchies in the ontology is diversified in order to increase the power provided by multiple inheritance mechanisms. *v) Minimal Encoding Bias* The ontology should be close to a natural textual form and independent of certain symbol-level encoding.
	b. Functional Requirements
	The ontology should be able to answer at least the following competency questions: 1. What are the Shariah concepts adopted in the products? 2. What is the Shariah source referred to by the products? 3. What is the scholarly opinion for the product? 4. What is the translated meaning of certain verse in the source of reference that relate to the Islamic banking? 5. What is the market segment of the product?

	6. What is the minimum asset value for the product? 7. What is the maximum asset value for the product? 8. What is the maximum income eligible to apply for the product? 9. What is the minimum age for the customer to apply for the product? 10. What is the maximum age for the customer to apply for the product? 11. What is the customer's role in the contract? 12. What is the bank's role in the contract? 13. What is the type of commodity dealt with by the product? 14. What is the profit rate of the product? 15. What is the margin of financing of the product? 16. What is the maximum tenure for the contract? 17. What is the minimum tenure for the contract? 18. Does the product require any collateral? 19. Does the product require processing fee? 20. What is the maximum annual fee to be incurred on the product? 21. What is the minimum annual fee to be incurred on the product? 22. What is the takaful requirement of the product? 23. Does the product require a security? 24. What is the ta'widh (compensation) charge for overdue account? 25. What is the lock-in duration of the product? 26. Does the contract incorporate clauses of the bank's right to set-off with other accounts? 27. What is the form of return that the deposit product provides? 28. Does the product provide grace period for payment? 29. Does the product provide payment holiday? 30. What is the Arabic word for a specific term? What is the defined meaning used in the Islamic banking?

For the purpose of ontology development, the non-functional ontology requirements refer to the development criteria, qualities or general aspects that are not related to the ontology content (Suárez-Figueroa et al. 2009). On the other hand, the functional ontology requirements are the content specific requirements, which refer to the particular knowledge to be represented by the ontology (Suárez-Figueroa et al. 2009). The functional ontology requirement should include the characteristics and criteria related or required in the ontology contents, which derived from requirements analysis in the problem exploration stage. The characteristics of the ontology refers to the distinctive nature of the concepts handled by the ontology. They are divided into general characteristics which should also be present in other domains and the specific characteristics that are peculiar for the particular domain of concern. Meanwhile, the development criteria of the ontology are the generally accepted quality of the resulting ontology (Annamalai 2005; Gruber 1993a).

The characteristics of the Islamic banking products include the distinct concepts dealt by the Islamic banks, the rules in the representation of the Islamic banking concepts, the diverse Shariah concepts adopted by the Islamic banking products, the diverse roles played by the banks and their customers, the differences

of opinions among the Islamic scholars, and the borrowings of Arabic terminologies by the Islamic banking concepts. Meanwhile, the criteria of the ontology involve the clarity, conciseness, coherence, extensibility and expansibility, minimal encoding bias, epistemological completeness and also the competency of the ontology.

b) Requirements validation checks the requirements for realism, consistency and completeness. The inconsistencies among the requirements were identified and resolved (Annamalai & Rosli 2012). Errors in the requirements documents were checked and be corrected (Sommerville 2011). The activity of the requirements validation need not be carried out in strict sequence after the requirements specification, but could be interleaved (Sommerville 2011).

3.4 Design

This stage involves the designing of the model for the selected domain. The description of the ontology will be built in a conceptual model in order to meet the specification defined. The conceptual model consists of concepts in the domain and relationships among those concepts are made using of conceptual models, such as from informal sketchy models, binary relations diagram or concept dictionary. With the terminology of the ontology defined, the informal competency questions could then be defined formally as an entailment or consistency problem pertaining to the axioms of the ontology (Gruninger & Fox 1995).

The design of the ontology is made with the requirement specifications in hand. It involves translating the characteristics of the ontology that has taken into account any complexities of the domain and also the criteria of the ontology. The intention for the designing activity is to structure the domain knowledge in a conceptual model that describe the problem and its solution in terms of the domain vocabulary identified during the ontology specification activity (Fernandez-Lopez et al. 1997). It involves identification of the key concepts and the relationships in the domain of interest, production of precise unambiguous text definitions for such concepts and relationships, and identification of terms to refer to such concepts and relationships as highlighted by Uschold & King 1995.

The steps of the design involve conceptualisation activity of the ontology, which are as follows:

a) Building a Glossary of Terms (GT)
The terms include concepts, instances, verbs and properties pertaining to the domain are gathered and listed. The terms could be identified from documents such as procedural manuals, guidelines, text books and articles related to the domain. The lists were expanded as the ontology development progresses. For the Islamic banking product ontology, the terms were gathered from the Ontology Requirement Specification Document, the Bank Negara Malaysia Guidelines on Product Transparency and Disclosure, and also from text books and articles on Islamic banking.

b) Grouping the terms as concepts and verbs

The terms are then grouped into concepts and verbs. Each set of concepts will be group in accordance to the sub-disciplines of the domain so that the concepts that are closely related to each other. Similarly, the verbs will also put in sets that are related to each other.

The ontologies for a particular domain could also reuse other common ontologies. Such reuse will involve fusion and integration of existing ontologies.

3.5 Development

The development stage involves the transformation the design model into components, and integrates the components into a complete ontology. The conceptual descriptions are transformed into formal models. Concepts are defined through axioms in order to restrict the interpretations of those concepts. As recommended by Fernandez-Lopez et al. (1997) and also Uschold & King (1995), definitions that have already been built into other ontologies will also reused.

The ontology is then encoded in an appropriate representation language. The coding will involve committing to the basic terms that will be used to specify the ontology (e.g. class, entity, relation), choosing a representation language, and writing the code. The Islamic Banking Product Ontology was encoded into the Web Ontology Language (OWL). OWL is a standard for coding of ontology and Semantic Web documents which were developed by the World Wide Web Consortium (W3C). It enables the coding of classes, properties, individual, and data values, and can be used with information written in RDF (Resource Description Framework).

3.6 Testing

The ontology need to be tested to ensure that its appropriateness and the definition of the terms should be semantically coherent (Fernandez-Lopez et al. 1997). Axioms could be used to define the semantics or meaning of the terms (Gruninger & Fox 1995). They shall specify the definitions of the terms in the ontology and the constraints on their interpretation, guided by the formal competency questions. The ontology will then be validated to ensure that it meets the specified requirements. It should conform to its specification and meets the expectations of its users (Sommerville 2011).

According to the Ontology Requirement Specification in Table 1, the ontology for the Islamic banks need to be tested for its non-functional and also functional requirements.

i) **Testing of Non-Functional Requirements**

The ontology need to be ensured that it will meet their characteristics and the development criteria.

a) Characteristics of the ontology

The **general characteristics** of the ontology were ensured to be as follows:

- The ontology need to include the distinct concepts of the Islamic banking industry, especially the Shariah concepts and also other distinct Islamic banking concepts such as *rabbul-mal*, *mudharib*, *khiyar*, *ta'widh*, *muqasah* and many more.

- The ontology of the Islamic banking products are meant to facilitate understanding on the banking products and processes by the customers so that they could make informed decisions. They are not targeted for use in automating the banking transaction or processes, but meant for facilitating their selection of banks and product. As such, full axiomatization of the concepts and relations in the ontology may not be a necessity, except for some concepts that need to be axiomatised to technically clarify their definitions. Hence, the rules in the representation of the concepts in the Islamic banking system were included in the ontology.

The ontology also need to be ensured that it include the **specific characteristics** meant for the ontology for Islamic banking which are as follows:

- The ontology need to be ensured to include the business models or Shariah concept commonly adopted by Islamic banking products.

- The ontology was ensured to include the diverse roles played by the banks and their customers. Visualisations made using the graphic tool of the ontology editor, Protégé shown that the ontology was able to cater to the different contractual roles based on their respective Shariah concepts.

- The ontology must take into account the differences of opinion of the different juristic schools.

- The ontology should also take into account the borrowings of Arabic terminologies by the Islamic banks.

b) Development Criteria

The development criteria to be fulfilled by the ontology

- *Clarity*

  The clarity of the contents should be made adequate for the customers to make effective comparison among the different products of the Islamic banks. As described by Gruber (1993), definitions should be provided in natural languages in order to provide clarity.

- *Conciseness*

  The ontology need to emphasise on brevity and avoids redundant definitions and explanations that could be counter-productive and might end up confusing the customers. According to Bezerra, Costa and Okada (2009), the ontology is considered as concise when it does not store unnecessary definitions, when there is no redundancy between terms definitions, and redundancies are not inferred by other definitions.

- *Coherence*

  The ontology in its natural language documentation should be defined coherently; whether individual or collectively. For its individual conceptual integrity, the definitions of the ontology were ensured were based from a single, yet reliable definition. In the development of the Islamic banking ontology, the definitions were based on documentation by ISRA (2012) so that there is no contradiction in the interpretation of the concrete concept pertaining to the entities represented. The collective definitions should also be made consistent to reflect the dependencies between their corresponding entities.

- *Extensibility and Expansibility*

  The ontology need to be ensured that the hierarchy in the ontology was diversified in order to increase the power provided by multiple inheritance mechanisms. The concepts in the ontology, such as the class: Business below, were diversified in order to allow to be extendable and expandable in the future.

- *Minimal Encoding Bias*

  In order to ensure that the ontology be close to a natural textual form and independent of certain symbol-level encoding, it was built in textual natural language. Its codification into .owl was only made for the purpose of testing and evaluation.

- *Standardisation of Names*

  Names of concepts or relations in the ontology were standardised whenever possible and as much as it could to adopt the terms commonly used in the Islamic banking industry.

- *Ontological Completeness*

  The ontology is required to produce the information needed by customers for their decision making purposes. As such, the ontology need to be tested for completeness based on the requirements of the Product Disclosure Sheet of the Bank Negara Malaysia and other pertinent information on the products as listed in the competency questions in Table 1.

- *Competency*

  The ontology need to be tested for its competency to at least answer or provide basic information to fill up the Product Disclosure Sheets of the Bank Negara Malaysia as listed in Table 1.

ii) Testing of Non-Functional Requirements

The testing of non-functional requirements need to be conducted with the use of competency questions. This was to ensure that the ontology meets the purpose of its design. The testing could be done by querying the ontology with the competency questions using the SPARQL Protocol and RDF Query Language (SPARQL) query.

3.7 Deployment

During this stage, the ontology is integrated with the application software in order to realise its planned use.

3.8 Maintenance

The maintenance involve the management of solution of faults and failures in the ontology-based application. It include updates and corrections of the ontology when needed due to the necessities of the existing use and also reuse in related applications.

3.9 Retirement

This is the phase when an existing ontology is removed from its active environment.

4.0 Results and Discussion

Using the Software Engineering Approach for Domain Ontology Development, the Islamic banking ontology produced comprises of 404 classes, 52 object properties and 66 data properties; thus producing 2,323 axioms.

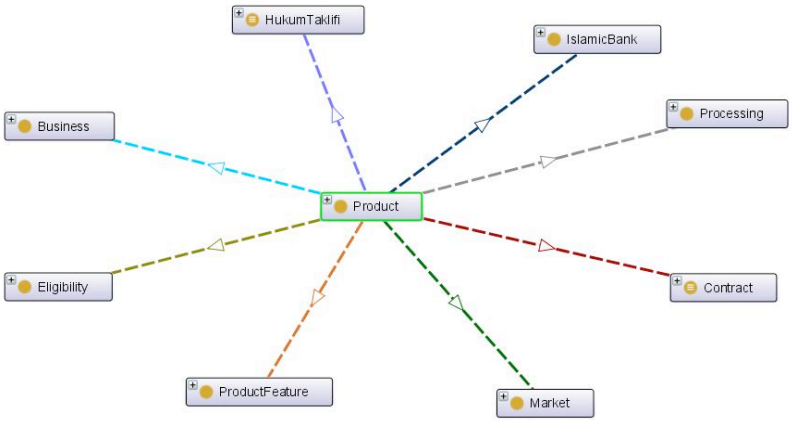


Figure 3: Ontology for Islamic Banking

Resulting from the outputs of the competency questions as listed in Table 1, the ontology is able to be queried to disclose the products' compliance with the Shariah according to the different juristic school and also to produce detailed information on the banking products. This will eventually facilitate the customers to understand the products and enable informed decision making.

The Software Engineering Approach for Domain Ontology Development was adapted from the Engineering Process (Level I) of EsOE. Yet, EsOE does not conduct root cause analysis as a method to clearly identify domain problems. As compared to UPON, SAMOD and UPON Lite where the ontology specifications were derived from use-cases or test-cases, the Software Engineering Approach for Domain Ontology Development identifies its specifications from root cause analysis. The analysis had enabled the identification of the characteristics and development criteria of the ontology which constitutes the non-functional requirements of the ontology.

5.0 Conclusion

In order to provide a reliable, long lived, and continually adapted ontology, it is best that domain ontologies be developed using the software engineering approach. However, the ontologies are dependent on the characteristics of the domain and criteria needed to attend to their problems. The Software Engineering Approach for Domain Ontology Development is able to adapt to the peculiarities of particular domains. With the conduct of the root cause analysis, ontology engineers will clearly understand the issues faced by the domain which will then be put up as requirement specification of the ontology.

References

Annamalai, M. (2005). Formative evaluation of ontologies for information agents. In Conference in Computer Science, Technology and Networking, Shah Alam.

Annamalai, M., & Rosli, M. M. (2012). Software Engineering: A Pathway to Enterprise-Strength Ontology Engineering. *International Journal of Digital Content Technology and its Applications,* 6(22), 98.

Annamalai, M., & Sterling, L. (2003, July). Guidelines for Constructing Reusable Domain Ontologies. In OAS (pp. 71-74).

De Nicola, A., Missikoff, M., & Navigli, R. (2009). A software engineering approach to ontology building. Information systems, 34(2), 258-275.

De Nicola, A., & Missikoff, M. (2016). A lightweight methodology for rapid ontology engineering. *Communications of the ACM*, 59(3), 79-86.

Fernandez-Lopez, M., Gomez-Perez, A., Juristo, N. (1997) METHONTOLOGY: from ontological art towards ontological engineering. In: Proceedings of AAAI'97 spring symposium series, workshop on ontological engineering, Stanford, CA, page no. 33-40.

Gomez-Perez, A., Corcho, O., Fernandez-Lopez, M. (2003) *Ontological engineering: with examples from the areas of knowledge management.* E-commerce and the Semantic Web, Springer-Verlag, UK. 2003, pp. 125-140.

Gruber, T.R. (1993a). "Toward Principles for the Design of Ontologies Used for Knowledge Sharing (Revision: August 23, 1993)", *In International Journal Human-Computer Studies* 43, p. 907-928.

Gruber, T.R. (1993b), A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 5, 199-220, 1993.

Gruninger, M., & Fox, M. S. (1995). Methodology for the Design and Evaluation of Ontologies.

Guarino, N. (1998). Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy (Vol. 46). IOS Press.

Hadzic, M., Wongthongtham, P., Dillon, T., Chang, E. (2009) Introduction to Ontology, Studies in Computational Intelligence, Volume 219, *Ontology-Based Multi-Agent Systems*, 2009, Pages 37-60.

IEEE (2006). IEEE Standard for Developing a Software Project Life Cycle Process. IEEE-SA Standards Board, The Institute of Electrical and Electronics Engineers, Inc. (IEEE), New York, 2006.

Latiff, A. S. A., Haron, H., & Anamalai, M. (2015). Grievances on Islamic Banks: A Survey. In *Islamic perspectives relating to business, arts, culture and communication* (pp. 295-306). Springer Singapore.

Latiff, A. S. A., Haron, H., & Annamalai, M. (2016, August). Characteristics and development criteria for Islamic banking ontology. In Information Retrieval and Knowledge Management (CAMP), 2016 Third International Conference on (pp. 136-142). IEEE.

Peroni, S. (2016, November). A simplified agile methodology for ontology development. In *International Experiences and Directions Workshop on OWL* (pp. 55-69). Springer, Cham.

Pinto, H.S. & Martins, J.P. (2004). Ontologies: How can They be Built? *Knowledge and Information Systems* (2004) 6: 441-464. Springer-Verlag London Ltd.

Pinto,H.S., Staab, S., Tempich, C. (2004). Diligent: Towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies. In ECAI, pages 393–397, 2004.

Sommerville, I. (2011). Software Engineering, 9th Edition. Pearson Addison-Wesley, Boston, 2011.

Studer, R., Benjamins, V.R., Fensel, D. (1998). Knowledge Engineering: Principles and methods, Data & Knowledge Engineering 25 (1998) 161-197.

Suárez-Figueroa, M.C., Gómez-Pérez, A., Villazón-Terrazas, B. (2009) How to write and use the Ontology Requirements Specification Document. In Proceedings of ODBASE 2009. On the OTM 2009. LNCS 5871. Pp: 966-982. Vilamoura, Algarve-Portugal. November 2009.

Team, C.P. (2010), CMMI for Development, v.1.3., Technical Report, CMU/SEI-2010-TR-033, Carnegie Mellon Software Engineering Institute.

Uschold, M., Gruninger, M. (1996). Ontologies: Principles, methods and applications. *Knowledge engineering review*, 11(2), 93-136.

Uschold, M., & King, M. (1995). *Towards a methodology for building ontologies* (pp. 15-30). Edinburgh: Artificial Intelligence Applications Institute, University of Edinburgh.

Yu, L. (2011). *A developer's guide to the semantic Web*. Springer Science & Business Media.

A REVIEW ON TRADE FLOW ANALYTICS

Kalongo Hamusonde
Asia Pacific University, Malaysia
kalongoh@gmail.com

ABSTRACT

This paper seeks to examine studies that have been conducted mainly in the period of 2015 to 2017 with regards to the analysis of trade flows in understanding their relationship with economic growth. Most studies examined causality relationships and employed co-integration tests and the Augmented Dickney-Fuller (ADF) model. The studies provided evidence that a relationship existed between trade flows and economic growth and this relationship was either bidirectional or unidirectional and existed both in the short-run and long-run depending on other factors other than trade flows. Other studies sought to compare statistical methods to determine which model produced statistically significant results in explaining and predicting trade flows with regards to economic growth.

KEYWORD: *Trade, Co-intergration, Machine Learning, Exports, Imports*

1.0 Introduction

Of the many indicators needed for a country to achieve potent economic growth, trade flows are one of those indicators that cannot be ignored. Trade flows refer to the buying and selling of goods from one country to another. Trade flows measure the balance of trade which is simply the exports minus the imports of goods and services. Trade flows play a major role in in influencing the growth of a country. If suppose a country is a major exporter of certain goods, the demand for that country's currency hence increasing the currency's value and in turn a rise in economic growth. With the importance of trade flows been put to light, this paper seeks to investigate studies that have been conducted on the analysis of trade flows i.e. exports and import. Quite a substantial amount of studies has been done with regards to analytics on trade flows in the region and most of these studies have used traditional statistical methods in their analysis. These studies have analysed exports or imports separately and sometimes collectively. The sole purpose of these analysis has been to determine the relationship between trade flows and at times Gross Domestic Product (GDP) with regards economic growth.

2.0 Discussion

With trade being one of the important indicators of economic growth as discussed above, many studies have devoted time to determining the relationship that exists between trade flows and economic growth. Tapsin (2015) conducted a study to ascertain what kind of relationship existed between foreign trade and economic growth in Turkey. To achieve this objective, the Augmented Dickney-Fuller (ADF) and unit root test were performed. The results showed that export and GDP had a bidirectional causality relationship while imports and GDP had a unidirectional causality relationship. The study finally concluded that imports and exports are important factors to consider when one seeks to measure

economic growth in Turkey because the variables share a positive and significant relationship. Similarly, Ucan et al. (2016) conducted a study to determine the relationship between exports and economic growth in Turkey. The results from this study were similar to the ones found by Tapsin (2015) though the only notable difference was relationship between exports and GDP had a unidirectional causality relationship. The two results were not consistent because Tapsin (2015) had not considered other factors in the model like exchange rates which could affect trade flows but were considered in the study by Ucan et al. (2016).

Saaed & Hussain (2015) main hypothesis was to test for causality and co-integration between GDP, exports and imports in Tunisia. To test this hypothesis, Tunis annual data from the year 1977 to 2012 with GDP, imports and exports as the attributes was used. The Vector Error Correction Model (VECM) framework which employs the Granger-causality test was used. The study results showed that imports, exports and GDP have a unidirectional causality with imports being the major source of economic growth in Tunisia. The results further showed that imports increased economic growth in the long-run whilst exports did not. The models or techniques used for analysis in this study were traditional methods which have their own drawbacks. Like many traditional statistical analytics, the methods used in this study employed a lot of assumptions. The ADF procedure which was used to test for integration assumes that the error term of Autoregressive(p) process to be white noise which is a less strong assumption than Gaussian noise stationarity. Running the procedure under this assumption is erroneous and leads to less reliable results. Another study conducted in India by Mehta (2015) also primarily focused on testing for causality and co-integration among exports, imports and GDP. Here, time series data from the year 1976 to 2014 was used and the ADF procedure was implemented to test for causality while the VECM was used to test for co-integration. The results showed that for the ADF to be implemented, the data had to be differenced at least once to attain stationarity. Once the data was stationary, further tests were conducted and it was found that long term GDP lead to increased exports but exports did not lead to a better GDP. It was also found that no causality exists between GDP and imports which implied that GDP does not lead to imports and imports do not lead to a greater GDP.

Similarly, Albiman & Suleiman (2016) conducted a study to determine if a relationship existed among exports, imports and domestic investment. This study used annual time series data from the year 1967 to 2010 for Malaysia and the VECM framework was used to test for causality and co-integration. Unlike the two studies discussed above, Saeed et al. (2015) & Mehta (2015), this study introduced capital formation into the analysis and the results differed in such a way that capital formation affected exports in the long-run while it affected imports in the short-run in determining the levels of domestic growth in Malaysia. Another study conducted by Altaee et al. (2016) to determine the effects of trade flows on economic growth in the Kingdom of Saudi Arabia showed that a fixed capital formation affected imports and exports in both the short-run and long-run. On the other hand, the study results also showed that financial development affected economic growth negatively in the short-run while turning out to have a positive impact in the long-run.

Bakari & Mabrouki (2017) sought to investigate the relationship between exports, imports and economic growth in Panama. Data ranging from 1980 to 2015 was tested for

co-integration using the Vector Auto Regressive (VAR) Model and for causality using the Granger-Causality tests. A few adjustments were made to the analysis like the introduction of a second test called the Phillips Perron (PP) test to validate the results produced by the Granger-Causality tests. Another alteration was the introduction of the augmented function, *'GDP=f (exports, imports)'* which was the aggregate production function of exports and imports expressed in logarithms. The study results from the VAR model showed that there was an absence of co-integration among exports and economic growth in Panama. This indicated that exports had no effect on the economic growth of Panama. On the other hand, a bidirectional causality existed among the variables imports and economic growth, hence imports were seen as a source of economic growth in Panama. With the same model and techniques used in the analysis of Panama trade flows conducted by Bakari & Mabrouki (2017), this time, Bakari (2017), used Germany annual data ranging from 1985 to 2015. In this study, one alteration was made, and this was the differencing of the export and imports data by using logarithms in order to make the model stationary before testing for causality and co-integration. The results were similar to the ones found in the Panama study though the only difference was that this study found a unidirectional causality between exports and imports. These results still provided evidence that exports and imports where a source of economic growth in Germany. A unidirectional causality between exports and imports was also found in another study and the results proved that exports and imports were a major source of economic growth in Pakistan, (Raza & Ying 2017).

Bakari (2016), again conducted an empirical study using Egypt annual data ranging from 1965 to 2015 to ascertain the relationship between exports, imports and economic growth. He used the VAR to test for co-integration and the Granger-Causality tests to test for causality. In this study, the variable domestic investment was introduced in the analysis to test against each of the three variables export, import and economic growth. The results showed that the introduction of domestic investment in the analysis significantly affected each variable differently. The co-integration results provided evidence that domestic investment, import and exports had no effect on economic growth while the causality results indicated that imports and domestic investment had a significant effect on the economic growth in Egypt. With Libya being one of the major exporters of petroleum in the world, Abdulhakim & Tarek (2016) conducted a study to determine the relationship between Foreign Direct Investment (FDI) and economic growth in Libya. This study focused solely on how foreign investments affect petroleum exports in relation to economic growth. Note that imports were not considered in this study. Annual export data ranging from 1992 to 2010 was gathered and the VAR model was used in the testing of this data to achieve the desired objectives of the study. Co-integration and causality between the variables was tested in this study. The results indicated that a long-term relationship existed between FDI and petroleum exports. The other results also showed that a unidirectional causality existed between FDI and petroleum exports. Though, it should be noted that one of the results provided evidence that there was no significant relationship between FDI and economic growth. This result was different from one of the deliverables of this study for it was expected that a significant relationship would exist between FDI and economic growth. One could assume that this output was a result of the researchers not controlling or putting into

consideration other factors that could affect FDI like tax rate, wage rate, exchange rate, political situation to mention but a few.

In addition to analysis on exports, Verter & Becvarova (2016) conducted an analysis to determine the impact of agricultural exports on economic growth in Nigeria. Granger Causality, Odinary Least Square (OLS) regression, Impulse Response Function (IRF) and Variance Decomposition analysis were the statistical techniques employed in this study whilst considering annual time series data ranging from 1980 to 2012. Since the data was a time series data, ADF and unit tests were additional tests which were run to make the data stationary and this was achieved after the first difference. The results from the IRF where not constant thus showing a fluctuation of an upward and backward shock of agriculture exports in relation to economic growth. A shock to agricultural exports affecting economic growth was also seen in the results from the Variance Decomposition analysis. Both results from the OLS and the Granger Causality tests provided evidence that agricultural exports did lead to economic growth in Nigeria. Nevertheless, these results contradicted to the previous works done by Ojide et al. (2014) who conducted a study to evaluate the impact on non-oil exports on economic growth in Nigeria. By performing the Autoregressive Distributed Lag (ARDL) model and co-integration tests on annual time series data ranging from 1970 to 2011, the study found that an inverse relationship existed between non-oil exports and economic growth. This difference between the results of the two studies can be attributed to the trade deficit in agricultural products which Nigeria which has incurred for the past six years. At the time Ojide et al. (2014) conducted the study, Nigeria Imported more goods than it exported hence the value for Nigerian currency was not high and this affected its exchange rate and in return resulting in an inverse relationship between non-oil exports and economic growth.

With exports playing an important role in influencing Uganda's economic growth, Karamuriro & Karukuza (2015) conducted a study to determine if the influence on economic growth by exports was statistically significant using the Gravity Model analysis. Since Uganda is a member state in the international organisations Common Market for Eastern and Southern Africa (COMESA) and East African Community (EAC), the two organisations were added to the model as dummy variables to determine if a significant relationship existed between Uganda's exports and its affiliation to these organisations. Other variables included in the model were exchange rate and common language. Annual panel data ranging from 1980 to 2012 was used. Since the data was of time series nature, the ADF test had to be conducted first in order to test for stationarity before any other analysis could be done. The data was found to be stationary after the second difference. The results of this study showed that exports had a statistically significant effect on economic growth. The variable common language as well showed a statistically significant effect on the exports. This meant that exporting goods from Swahili speaking countries like Kenya and Tanzania had a major effect on Uganda's economy. The results further showed that Uganda's affiliation to the two international organisations had a significant positive effect on the exports because the two organisations provided intra-trade among member states which came with prized benefits such as reduced tax rates. Another study to determine Egypt's intra-trade intensity with COMESA member states was conducted by Elmorsy (2015). The study objective was to determine which variables play a major role in promoting trade in Egypt with relation to

COMESA member states. To achieve the objective, the Gravity Model analysis was employed and data from 2005 to 2011 regarding COMESA total imports, Egypt's total exports and world total exports was used. The results gave evidence that intra-trade agreement among COMESA member states contributed to the economic growth in Egypt. Nevertheless, a few obstacles such as political, social and infra-structural issues were also identified as potential threats that would affect trade flows negatively in the long-run. Alkhateeb et al. (2016) further conducted a study on Egypt's intra-trade agreements and their effects on agricultural trade flows. The study results were similar to the ones achieved by Elmorsy (2015) but added that population size and exchange rate also had a significant contribution on economic growth in Egypt.

Additionally, Abedini & Darabi (2015) conducted a study in which they examined the relationship between GDP, exports, imports, inflation and insurance and how all these variables affect economic growth. This study used annual data ranging from 2003 to 20111 and focused only on Organization of Petroleum Exporting Countries (OPEC) member states. The regression analysis was the statistical technique used in achieving the desired objective of this study. The results indicated that all the variables in the study had a positive and significant relationship with each other and thus implied that when productivity increases, insurance would increase too and that lead to a rise in economic growth. Similarly, Belke et al. (2014) sought to determine the relationship between exports and domestic demand in six European countries which are part of the European Union. A smoothing transition regression model was the main statistical technique used but other models were also employed like the sunk-cost model used in capturing non-linear hysteresis dynamics and the unit root tests to test for stationarity. The data was sourced from national statistical offices of each of the six countries in the study ranging from 1980 to 2012. The results from the smooth transition regression provided evidence that domestic demand is a necessity in the short-run to improving exports. The results further indicated that the non-linear relationship between domestic trade and exports was extreme during some stages of the business cycle. This indicated that paying sunk costs for shifting sales would give rise to the performance of the export market in the long-run and would still remain high even in an economic turmoil.

Not all trade analysis studies have been based on determining the relationship between trade flows and economic growth. Some of the studies have been based on analysing and comparing models to determine which one is better at achieving the desired objectives or testing how effective a statistical technique can perform at analysing trade flows. Milad et al. (2015) conducted a study to determine which model was better in forecasting Malaysia's imports of crude material by comparing two composite models. The first composite model was one with regression processing whilst the second model was one without regression processing. Data ranging from 1991 to 2013 was used and unit root tests were performed on the dataset to make it stationary. The results indicated that the model with regression processing outperformed the model without regression processing. The model with regression processing was able to reduce forecasting errors better than the other model. This is because the model with a regression process had a lower percentage of U-statistics which meant that it provided a best fit with highly significant p-values than the other model. Additionally, Kompas & Che (2016) conducted a study to determine if the structural and stochastic optimal model was significant in forecasting imports and exports of liquefied natural gas (LNG) in Asia-Pacific. The

results from this study indicated that LNG exports in Asia-Pacific would increase by 90% in a period of 15 years starting from 2015 onwards. The results also projected an increase in LNG demand that would in turn increase imports of LNG within Asia-Pacific. Therefore, the study provided evidence that the structural and stochastic optimal model was reliable in forecasting trade flows. Similarly, Mladenovic et al. (2016) conducted a study to determine which Artificial Neural Network (ANN) algorithm between Extreme Learning Machine (ELM) and the Back Propagation (BP) was more accurate in predicting GDP based on trade flows. Data from 28 European countries acquired from EUROSTAT was used in this study. The results showed that the ELM algorithm outperformed the BP in prediction accuracy and it had very small number of underestimated values as compared to the BP algorithm.

3.0 Conclusion

The studies have shown that trade flows can be used in determining economic growth in the short-run and long-run. When it comes to modelling, causality and stationarity tests should be the very first tests one should undertake when analysing trade flows for the data is of time series nature. This helps in the reduction of errors and provides very reliable results. It is very important to also take note that other variables like exchange rate, domestic investments, sunk costs etc. should be considered when modelling for they tend to affect trade flows differently hence omitting them would lead to erroneous results that would not be generalizable. It was also noted from some of the studies that affiliating to international organisations proves helpful for the intra-trade agreements tend to improve economic growth for affiliated countries by providing certain incentives like reduced tax rates. It can be further noted that not many works have been done where big data analytics and the use of machine learning algorithms is incorporated. Future studies would do well to venture in that direction in order to describe, predict or prescribe trade flows with a much greater accuracy and minimised errors.

References

Abdulhakim, A. A. & Tarek, Z. (2016) The effects of foreign direct investment on economic growth in Libya: A causality analysis. *Open Science Journal.* 1 (2). P. 1-15.

Abedini, R. & Darabi, R. (2015) The effect of private investment, exports, imports, inflation and GDP on per capita premium: Evidence from members of OPEC countries. *Management Science Letters.* 5 (1). P. 657-662.

Albiman, M. & Suleiman, N. (2016) The relationship among export, import, capital formation and economic growth in Malaysia. *Journal of Global Economics.* 4 (2). P. 1-6.

Alkhateeb, Y. T. T., Mahmood, H. & Maalel, N. (2016) Egyptian intra agriculture trade with Common Market for Eastern and Southern Africa trading partners: A gravity model. *International Journal of Economics and Financial Issues.* 6 (6). P. 177-182.

Altaee, A. H. H., Al-Jafari, K. M. & Khalid, A. M. (2016) Determinants of economic growth in the Kingdom of Saudi Arabia: An application of autoregressive distributed lag model. *Applied Economics and Finance*. 3 (1). P. 83-92.

Bakari, S. (2016) The relationship between export, import domestic investment and economic growth in Egypt: Empirical analysis. *Munich Personal RePEc Archive.*

Bakari, S. (2017) Trade and economic growth in Germany. *Munich Personal RePEc Archive.*

Bakari, S. & Mabrouki, M. (2017) Impact of exports and imports on economic growth: New evidence from Panama. *Journal of Smart Economic Growth.* 2 (1). P. 67-79.

Belke, A., Oeking, A. & Setzer, R. (2014) Exports and capacity constraints: A smooth transition regression model for six Euro area countries. *Working Paper Series.*

Elmorsy, S. (2015) Determinants of trade intensity of Egypt with COMESA countries. *Journal of the Global South.* 2 (5). P. 1-25.

Karamuriro, T. H. & Karukuza, W. (2015) Determinants of Uganda's export performance: A gravity model analysis. *International Journal of Business and Economic Research.* 4 (2). P. 45-54.

Kompas, T. & Che, N. T. (2016) A structural and stochastic optimal model for projections of LNG imports and exports in Asia-Pacific. *Heliyon.* 2 (6). P. 1-35.

Mehta, N. S. (2015) The dynamics of relationship between exports, imports and economic growth in India. *International Journal of Research in Humanities & Social Sciences.* [Online] 3 (7). P. 39-47.

Milad, A. H. M., Ibrahim, I. R. & Marappan, S. (2016) A comparison among two composite models (without regression processing) and (with regression processing), applied on Malaysian imports. *Applied Mathematical Sciences.* 9 (116). P. 5757-5767.

Mladenovic, S. S., Milovancevic, M., Mladenovic, I. & Alizamir, M. (2016) Economic growth forecasting by artificial neural network with extreme learning machine based on trade, import and export parameters. *Computers in Human Behaviour.* 65 (2016). P. 43-45.

Ojide, G. M., Ojide, C. K. & Ogbodo, C. J. (2014) Export-led growth hypothesis in Nigeria: Applications of ARDL model and co-integration analysis. *Global Journal of Emerging Market Economies.* 6 (1). P. 5-13.

Raza, M. & Ying, X. Z. (2017) The causal relationship between export and economic growth of Pakistan. *International Journal of Economics, Commerce and Management.* 5 (2). P. 210-231.

Saaed, A. A. & Hussain, A. M. (2015) Impact of exports and imports on economic growth: Evidence from Tunisia. *Journal of Emerging Trends in Economics and Management Sciences.* [Online] 6 (1). P. 13-21.

Tapsin, G. (2015) The relationship between foreign trade and economic growth in Turkey. *International Review of Research in Emerging Markets and the Global Economy.* 1 (3). P. 417-429.

Ucan, O., Akyildiz, A. & Maimaitimansuer, M. (2016) The relationship between export and economic growth in Turkey. *European Scientific Journal.* 1 (6). P. 61-70.

Verter, N. & Becvarova, V. (2016) The impact of agricultural exports on economic growth in Nigeria. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis.* 64 (2). P. 691-700.