VOL. 14, NO. 8, APRIL 2019 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

DETERMINATION OF TEXT RELEVANCY BASED ON KEYWORDS ASSOCIATION FOR INTERACTIVE NEWS NETWORK

S. M. F. D Syed Mustapha
Computer Science Department, College of Computers and Information Technology, Taif University, Saudi Arabia
E-Mail: smfdsm@gmail.com

ABSTRACT

News network is an initiative to allow the content of several news to be associated by the contextual information such as event, people and location. This information may use common and proper noun to describe the similar context of an object, such as "Washington D.C" and "the capital city" or "former president" and "Barrack Obama". These words co-occur in various times such that they can be associated as keywords to describe certain context on the content of the news. In the literature, many approaches and techniques on the keywords extraction have been discussed but it is argued the lacking on keywords association based on the context of the text, particularly news. Associated keywords are used to "synonymize" the words that are related by the context of news rather than merely observing syntactical or synonymical values. From these words association, news network can be built from the news corpus such that news structure is a stratification that is based on its relevancy to the set of keywords. Named entity recognizer which is a known research area plays significant role in characterizing the relationship in the news network such that the relevancy between news are understood semantically. Event contains the essence of the news that is made up of the activities, actors who are involved in the activities, location and other non-living objects that made up part of the event, called signatures. The results demonstrate the formation of associated keywords based on the context and the building of the news network.

Keywords: text processing, news retrieval, keywords association, information retrieval.

INTRODUCTION

Deployment of digital technology in news delivery has becoming popular to surpass the traditional hard printing newspaper in which many have reported to shutdown (Time Magazine/CNN, 2009). Since the advent of digital devices, internet and multimedia technologies, the news delivery methods and presentations have changed significantly in terms of its interactivity which includes the presentation formats and content delivery method. Interactivity is enhanced using Hypertext which allows user to navigate news related to the reader's interest or topics that has been organized in chronicle order by the publishers (Hüser & Weber, 1992). Interrelated topics which cover the past related news allow users to browse and aggregate the understanding quickly. The presentation of the news has changed in various modes of presentation. News web feeds (Brenna *et al,* 2007; Trampus, and Novak, 2012) is a form of syndication where the news content is made available to the subscriber through websites or desktop or mobile devices automatically without the need to search or fetch manually. News personalization has made a major advancement in the delivering news on set-top box in which the delivered news content is filtered based on the user profiles and needs; and these make news delivery appropriate for the user's preferences prescribed in the profile (Lee and Park, 2007; Morales *et al*, 2012, Montes *et al*, 2013; Wen, *et al*, 2012). The emerging of these technologies somewhat has similar objective that the delivered news are the results of some interactions that took place with the users in order to make the delivery more satisfying. The user's preferences can be pre-determined based on the profiles which are configured by the user or in the recent attempt; it is automatically recognized based on the browsing history and patterns of the users (Fan *et al*, 2014). Another approach to increase the news interactivity is when news in the repository is classified using machine learning techniques based on some similarity features. News articles are associated as galaxy of news or news association that allows pyramidal structuring, semantic zooming and panning as well as visual presentation in multiple dimension of news collection (Rennison, 1994). News interactivity is another dimension that we observe that needs more attention by the research community as it enhances the users' experiences in receiving news. Either users are fed with pre-determined news topics based on the prescribed web feeds or users receive personalized news content through intelligent recommender system, there is always a challenge to ensure that the delivered content is fully satisfied and relevant to the users' needs for few reasons. Firstly, the predefined users' preferences (either manually set or automatically determined using some machine learning techniques) may not accurately describe the content of the delivered news due to the differences of the keywords used in the users' preferences and news text. Secondly, the terms usage for the news category may be generic. For example, user who is interested with "Political Turmoil in Honduras" may not be interested with "Suicidal Bombing at Manchester Arena" even though both are categorized under world politics. Thirdly, news texts relevancy can be defined in various ways - by the event, persons who are involved or locations. For example, readers who are investigating on various road accidents may only be interested with the "accidents" as event regardless of the locations and individuals in the incident and this news are considered as relevant with respect to the specific event. Similarly, a prominent individual may be the main subject in determining the relevancy regardless of the events and activities the person is involved.

In our perspective, having the right keywords play prominent role in retrieving the relevant interactive news. In addition to that, determining the keywords must be predetermined through prior analysis of the past news in order to form a set of keywords that can be associated to describe about the context of certain news. For example, in a news set, proper nouns and common nouns are used to describe the same entity, such as "45th president of united states", "Donald Trump" and "Republican" are collection of associated words describing the current president of United States. News text may use these variations of proper nouns and common nouns in different releases and the collection of these words are called associated keywords. The usages are based on different context of the news. To determine the associated keywords, few steps are proposed: firstly, determine the associated keywords among the interactive news articles that they are considered contextually synonymous. Secondly, to search for the keywords from an article that may coexist in other articles to establish common associated keywords across the news text. Thirdly, to recognize the similarity features of the articles based on the events and named entity for the persons and locations. Subsequently the fourth step is to build interactivity among news network using the common associated keywords. The descriptive relationship between the news can be used to retrieve news that are relevant. The following sections discuss on the general architecture, demonstrates the result of the experiment and conclusion on the technique introduced in the work.

RELATED WORKS ON THE KEYWORDS EXTRACTION

Considering the importance of the having keywords for determining the relevancy of the text, there are extensive works being done on keywords extractions for various types of text and each to be discussed. Lee *et al* (2012) developed High Relevance Keyword Extraction (HRKE) as an enhancement to traditional Bayesian Classification that some the pre-selected keywords are those that have been pre-selected based on their high relevancy to the categories. This method requires the categories of the words to be pre-determined since its probability is calculated based on the probable category. Furthermore, since the experiment is applied on the full text where the number of words is sufficient for statistical calculation. Another attempt on blogs is described by Chen, *et al* (2014) where keywords are determined from the collection of blogs. Since, the blogs are written in HTML, FRKP (Full Text Keyword Retrieval Process) is applied to remove the HTML tags prior to using TFIDF as the method to determine the first 20 top keywords. Like the online news text, blogs are an open domain, hence, determination of the categories have to be decided earlier before the extraction of the keywords. Another attempt in extracting keywords from micro-blogs is described by Biswas *et al* (2018) where selected keywords are presented using graph. Each node represents the keyword in which the importance of the node is based on the position, centrality and frequency. Another method that is worth considering is the keyword expansion by deploying statistical method (Matsuo & Ishizuka, 2004), machine learning (Sebastiani, 2002) or Google similarity distance (Chen and Lin, 2010). Rather than extracting keywords from the documents, key word expansion is used to predict what are the intended keywords that users are looking for based on the ordinary words entered by the users in the search engine. Keyword expansion is important to assist users in improving the words used in searching to more accurate keywords. Romero, *et al* (2012) used Wikipedia in order to determine the category type of a word and also supporting document for some peculiar words. The choice of the words is based on the frequency of the words used for a language based on the analysis done by COCA (Davies, 2011). A recent study which is close to our work is the proposed system to capture the main keywords and to build the network of semantics keywords from the comments generated from the forum posted by the online news reader (Beck-Fernandez *et al*, 2017). Their idea of finding the associated words from a given word (as a concept) is similar to our proposed work but the conceptual word has to be pre-determined by using n-gram and the percentage of its occurrences. The difference with our work is that the keywords extraction is applied on the comments rather than the text from the news.

The related works that were discussed earlier are different from the work proposed in this work in three ways. Firstly, the keyword extractions do not depend on the pre-processed conceptual ontology, categories or semantic network since the topics of the news change and prior determination of the category or news domain may not be a useful exercise. Secondly, the identification of the keywords that are based on statistical approach such as the number of occurrences or frequencies do not consider the context of the news. Even though we use some statistical information, but the focus is on the contextual aspects of the keywords. Thirdly, the definition of associated keywords in our work is different such that the associated keywords in our context means collection of words that co-occur due to some incidents in the news such as event, person or location and not merely those words that appear together for other reasons. The following section described how associated words are determined based on the context of the news.

Determination of the associated keywords

News is journalistic report that describes a story on what and how it happens, when it happens, where is the location and who are involved. "How" describes the cause and effect and what events prelude to the other events. The information on the cause and effect may contain within the same article or across several articles. Actions from different articles may have chronological facts that can be used to determine the sequences of the articles. "What" comprises of the main activity and other sub-activities that supplements the main activity. News article may have more than one action to describe an activity and more than one activity may appear in an article. "When" is a temporal information that describe the date of when the news is reported as well as other time instances such as tomorrow, two months ago, last week etc. "Where" is the

location-based facts that includes name of a country, city, building, venue which can be proper noun and common noun. Likewise, "Who" relates to animated object that becomes the actors to the main activities and usually are referred to the proper noun.

Associated keywords extraction

Associated keywords are group of words that co-occur in more than one article due to some contextual factors. The contextual factor could be due to the person, event or location. In the example of a person, one could possess common noun besides proper noun to describe himself. For example, "President of United States", is a common noun and "Donald Trump" is a proper noun but both could appear together in few occasions. The five words, namely, "president", "united", "states", "Donald" and "trump" can be associated in the context of the name of the current President of United States. The algorithm in Figure-1 defines the process of extracting associated keywords.

In step 1, assume that there is a collection of news article, called α and d_i represents a news article in the collection. In step 2, we need to find the collection of terms that are qualified to be the keywords to represent the article. For our case, the keywords are the terms that are among the highest frequency.

1. Given a collection of articles $\alpha = \{d_1 \cdots d_n\}$ where n is the number of articles;
2. $\forall d_i$ find $t_j^{d_i}$ where $j = \{1 \cdots m\}$ and m is the number of keywords, $t \in d_i$;
3. $\forall d_i \forall d_j \exists t^{d_i} \exists t^{d_j}$ where $t_{1 \cdots m}^{d_i} \cap t_{1 \cdots k}^{d_j} \neq 0$;
4. Perform Step 3 to obtain $t(p)$ where $\exists i, j$, such that $d_i \cap d_j$ $t(p) = \{d_1, d_2 \cdots\}^M$ and $M > 0$

Figure-1. Associated keywords algorithm.

It states that $t_j^{d_i}$ is the term t that occurs in the news article d_j and m is the number of keywords are found to be among the highest. There is a need to set the threshold on the number of occurrences for a term to be selected. The potential associated keywords which co-occur together, usually will have the same value of occurrences. Step 3 is the significant one in finding the associative keywords across the news articles. For each news article, inspect whether there exist $\exists t^{d_j}$ (i.e. the collection of keywords determined in Step 2) where these keywords appear in article d_i and also in article d_j. As stated in Step 4, this process is applied repetitively to search for the collection of keywords $t(p)$ that co-occur in both d_i and also d_j in M times. The value M is an adjustable value set by the user. Associated keywords must appear together in a number of articles and that is to be determined depending on the nature of the news collection. Since the collection of words (so-called associative keywords) co-occur together, these keywords can be used interchangeably by the users when performing searching on the news corpus. For example, if the query contains "president" "united" and states", there will be articles that when they are retrieved will have "donald" and "trump". The idea of associated keywords works well for the articles that appear in series where the collection of words is used repetitively across several articles. Figure 2 illustrates the list of keywords for the respective articles and the associated keywords. The list of associated keywords is an accumulation of the keywords of all the documents that have a subset of their keywords to overlap at least a specific number of documents.

The news was in 2010 by the Malaysian online newspaper that describe about a movement to protest Datuk Sammy Vellu who was the MIC president at that time (https://www.thestar.com.my/search/?q=%22MIC+President%22&qkey=MIC+President&pgno=127&qsort=newest). Four articles (d1, d2, d3, d4) are examples of articles that content the associated keywords in which these keywords are used to build the network of articles (so-called News Network) which is explained in the following subsection.

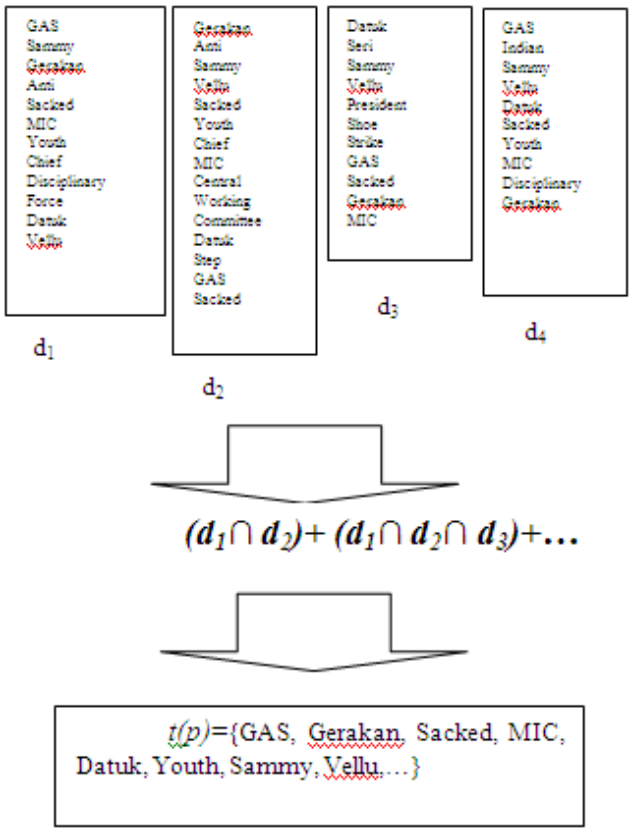

Figure-2. An illustration on associated keywords identification

News network on frequency-based keywords extraction

Like in the other works such as news classification and topic detection, keyword extraction plays the pivotal role. In both cases, the keywords selection is based on the discriminatory attributes that the keywords possess which are determined by comparing with other collection of keywords. For example, in the topic detection, the keywords in the news title have higher significant values in terms of representing the content than the other keywords that appear in the body of the text (Zhang, *et al,* 2016). In our case, the keywords extraction does not use a pre-defined controlled vocabulary of a particular domain in order to guide the selection of the right keywords based on particular topics. All keywords are treated to be equally significant and the selection of the keywords is based on those keywords that have the highest frequency in comparison to the others within the same article. Based on this method, we preserve neutrality in the keyword selection since the news network is built based on the popularity usage of the word to indicate the significance of word occurrences as determination factor for building relationship between news. Figure-3 shows the algorithm for building the news network based on keyword frequency.

1. Given a collection of news, $D = \{d_1 \cdots d_N\}$ where N is the number of news articles.
2. $\forall d_i \in D$, find the frequency for each term $t \in d_i$, where $\kappa(d_i) = \{t_i^f \cdots t_K^f\}$ where K is the total numbers of terms in document d_i
3. Given a threshold value, τ, $\forall d_i \in D$, find in $\kappa(d_i)$ where $f > \tau$
4. $\forall d_i \in D$, given $\omega(\kappa(d_i)) = \{\kappa(d_i) \cap \kappa(d_{i+1}) \neq 0\}$, find χ which is defined as $\sum_j^N \sum_i^N \chi_{d_{i+1}}^{d_j} = \{(\omega(\kappa(d_i)) \cdots\}$ provided that $i > j$

Figure-3. Algorithm for building news network.

Step 1 defines the set *D* to be a non-empty set of news article. In step 2, for each term, *t*, for all document d_i, find the frequency which means the number of occurrences within the article that the term appears. Normalization is applied where each frequency is divided by the highest frequency in the same article to ensure the frequency values falls within the range of 0 to 1. In step 3, *κ(d)* contains the list of terms with normalized frequency values and followed by delisting those which fall below

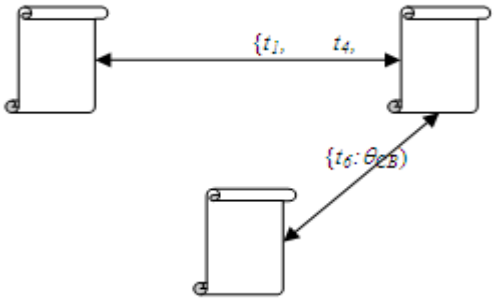

Figure-4. News network relationship with strength value, θ

the threshold, τ. Step 4 applies the condition set in step 3 to all news articles by comparing each article to every other to find at least one term which is syntactically similar. A simplified procedure to perform syntactical matching is given below which is used to compare two terms, t_a and t_b:

a) Calculate $\eta(t_a)$ and $\eta(t_b)$ which are the number of letters in each term

b) Determine $\rho(t_a$ and $t_b)$ which is total number of letters that match in ordered sequence. For example, t_a is "communications" and t_b is "communicate", then $\rho(t_a$ and $t_b)$ is 10. Since the length of the words may affect the value ρ, $\hat{\rho}$ is a normalized value between 0 and 1.

c) Set the minimum value for $\hat{\rho}$ to select the potential terms.

χ represents the news network based on syntactic words matching as illustrated in Figure-4. Article A and B are connected by three shared terms where θ shows the strength value. It is calculated based on the number of terms that match between the two articles against the total number of terms as given in equation (1):

$$\theta_{AB} = \frac{\overline{\overline{d_A \cap d_B}}}{\overline{\overline{d_A \cup d_B}}} \quad (1)$$

in which in this case, $\theta_{AB} > \theta_{CB}$ (Note: $\bar{\bar{A}}$ denotes cardinality). In calculating the cardinality of $d_A \cup d_B$, shared terms in A and B are counted once.

Event detection and named-entity news matching

Events have attributes that are recognizable and usable to determine the similarities and differences between articles. Even though the standard journalistic writing style has the common facts on who, where, when, what and how, the unstructured representation of the news text makes detecting these specific facts a challenging task. Text understanding is possible for the text that is supported with appropriate annotation to describe the content of the text passage or ontology that can be referred

to obtain the semantic meaning of the words. In the case of news, the types of events are unknown such that building ontology prior to receiving the news is impossible since ontology is usually domain specific and the text annotator that is defined using HTML or SGML scripting is not detail enough to describe the specific content within the body of news text.

The primary consideration for event detection in our approach is to determine a set of word that best describes the event from each news article. Unlike the frequency-based keywords (refer to section B), the set of words may not necessarily be the ones with highest frequency. Our belief is that some of the words have strong representative value even though they may not be the highest in frequency compared to the others. For example, some of the words appear in title have strong representation value but appear less in main body.

a. Given $E = \{A, T, L, S\}$ where A is an actor, T is an activity, L is a location and S is the signature of the event E.
b. Property A: person's name, person's title, person's position (political post, government official, social status etc)
c. Property T: an activity that consists of one or more actions, denoted $T = \{a_1, \ldots a_N\}$. Action is a collection of verbs that is performed by entity in property A.
d. Property L: location which is referred to name of country, state, city, building, park village, institution, offices etc
e. Property S: signature is the object that signifies other attributes of the events

Figure-5. Definition of property events.

To find a set of keywords that represents the event, manual subjective inspection is made across all news to determine those with similar events based on the properties defined in Figure-5. The event, E, comprises of four main properties, actor, activity, location and signature. Actors are living objects that perform the activities at certain location that may be supported by other non-living objects, called signatures. For illustration, Figure 6 illustrates the assignment of event properties to the respective words in a news text. The assigned news text with event properties is used to perform the similarity comparison with other news articles. The similarity is measured based on the number of event properties that are involved and the number of the matching hit for each category as given in the equation 2,

$$\delta_{AB}(E = \{A, T, L, S\}) = \sigma_E \frac{\sum_1^4 \overline{\overline{(\cap_A^B t_E)}}}{\sum_1^4 \overline{\overline{(\cup_A^B t_E)}}} \quad (2)$$

where δ_{AB} is similarity measure between news article A and B and σ_E is the weightage that is adjustable to emphasize the importance of certain event properties. Since the number of terms, t, in news article varies, the cardinality values ($eg. \bar{\bar{A}}$) in equation 2 is normalized.

KLANG(L): MIC members(A) and supporters(A) gate-crashed (T) the second anti-SamyVellu rally(S) here Sunday and brought it to an early end. The organizers (A) ended the gathering (T) of 1,500 at 3pm, two hours earlier than scheduled, on the advice of the Klang police chief (A).
Samy(A) called(T) for press conference on the stage and handed out a four-page memorandum (S) to the press.
The four-page memorandum (S) contained nine issues, including calling for the government (A) to lower the schooling age to five(T), reduce personal tax (T), establish kindergartens (T) within existing Tamil schools (L), rehabilitate prisoners (T) and redefine poverty (T).

Figure-6. News text with assigned event properties.

Event properties have pragmatic meaning that is affiliated to the real-world context. For example, "Lincoln Square", "Lincoln Square Cinema" and "Abraham Lincoln" share "Lincoln" as part of the name but referring to different event properties. In order to ensure the event properties are assigned correctly to the words, the named-entity word extraction is performed. There are several methods and approaches in determining the named entity and decades of research work in this area can be found in the literature (Abdul-Hamid, *et al.*, 2010; Alotaibi, *et al,* 2012; Mohit, *et al*, 2014). The approach ranges from machine learning technique (Habibi, *et al,* 2017), gazetteers dependent or gazetteers generator (Lample, 2016) and rule-based system (Gabbard, 2017). Support Vector Machine is a machine learning classification technique that work on sample of labeled data and unknown sets. The labeled data is a small set of seed list of named which are expanded using lexical pattern rules. The generated labeled data are used as the trained set to classify the unknown sets (Ekbal, 2010). Gazetteers list for location and names are good source of input to the named entity recognizer to produce high accuracy results. Some approaches consider an automated generation of gazetteers using bootstrapping technique (Kozareva, 2006; Becker, 2005). A list of annotated corpus is used as the input sources that are extracted and filtered using bigram method. In some languages such as Turkish, more complex constrained rules are needed to disambiguate the multiple meaning of a word for the purpose of POS tagging and to apply morphological matching (Oflazerand Tür, 1996).

Our current work focuses on the location and name detection. Since, news article is global in its coverage, having gazetteers for location and names can be infinitely challenging. The approach being taken is to refer to the linguistic style of the journalist and to form list of generic rules that can be applied in a controlled manner. A collection of Reuters news is analyzed subjectively through visual inspection to find the common linguistic pattern where location and person are named. Some of the observations that are found to be the challenges in dealing

with the person or location's name in the news text are as follows:

Ambiguous verbal action - common English refers to some of the verbal words to be actioned by the human such as "says", "declares", "announced" etc. These words create ambiguity between names of a person or non-person such as organization or society. For example, "Viacom International Inc said National Amusements Inc. has…"

Common names - common names that are used as a person, month, building or location. For example, "June" as the name of the month and a person's name or "Lincoln" as the name of a person or building such as "Lincoln Inn".

Non-physical names - names (location or person) in the news are detected from the capitalization of the spelling. However, there are other names that are capitalized in the spelling format that are non-physical such as "Canadian Dollars", "World FIFA" or "Democrats". These names are difficult to be differentiated with other physical names without having the background knowledge. For example, the preposition that commonly used to indicate a name as location such as "at" and "in" can somehow be used for these words.

Rules are developed subsequently based on the subjective observations. For each case, a rule is constructed to deal specifically for a particular situation. There two categories of rules:

- **Concrete rule:** rule which is universally acceptable and valid for all cases.
- **Soft rule:** rule that works for some cases and can be superseded with other rules.

For example, the following rule is a concrete rule for the basic Noun Rule using BNF notation:

Noun ::=<Word> { <Word>}
Word ::=<letter> { <letter>}
letter ::=<upper_case> {"." | <lower_case>}
upper_case ::= "A - Z"
lower_case ::= "a - z"

The Noun Rule defines that all words that begins with capital letter will be considered a noun which includes the name of a person, building, society, political party and so forth.

Soft rule is described using the following rule template which determines the name of a place if the word is followed by the preposition "at", "the", "in" or prepositional phrase such as "beside the". However, the rule is not valid for all cases as shown in Case 1 (valid) and Case 2 (invalid):

Case 1: "The tourist stands beside the Grand Mall to take picture"

Case 2: "Mrs Jones is feeding the baby while her 10 years old son is standing beside the Oak tree"

In both cases, nouns are followed by the prepositional phrase "beside the" but the latter is not a location. Since location can be ambiguous, it requires several rules to counter check from other sources. Currently, Google Map is one of the possible sources for cross-referencing on physical location.

RESULTS

Two experimental results are demonstrated in which the first part is on extracting the associated keywords and followed by the frequency determination of the keywords obtained from the earlier part based on the theories discussed above.

Associated keywords extraction

The algorithm is applied on eight articles that contain the same story. The articles are hand-picked to ensure that they describe the same story but may contain different keywords to describe the same context. For the illustration purpose, we show the outcome for two selected articles. Table-1 shows the list of keywords that are extracted based on the threshold set as 0.1. It is observed that since the algorithm uses word frequency to determine the keyword, some important words that have significance role in describing the story may be omitted when the threshold is set higher. Each Reuter represents an article and since there are keywords which overlap between these articles, they can be considered as the potential of associated words. The overlap value for this example is set to 0.3 that means 30% of the collection of Keywords in Reuter No *i* is the same as Reuter No *j*.

Table-1. Collection of potential associated keywords.

Reuter No	Keywords
1	anti-Sammy, gerakan, chief, Indian, member, MIC, Mugilan, party, president, Sammy, Vellu, youth
2	gerakan, chief, Indian, member, MIC, Mugilan, party, president, Sammy, Vellu, youth,step
3	gerakan, Indian, member, MIC, Mugilan, president, Datuk, Seri, Sammy, Vellu, youth,anti-Sammy
4	Indian, member, MIC, Mugilan, president, Datuk, Seri, Sammy, step, chief, Vellu, youth
5	anti-Sammy, Datuk, gerakan, chief, Indian, member, MIC, Mugilan, party, president, Seri, Sammy, Vellu, youth
6	anti-Sammy, Datuk, gerakan, chief, Indian, member, MIC, Mugilan, party, president, Seri, Sammy, Vellu, youth,step
7	anti-Sammy, Datuk, gerakan, chief, Indian, member, MIC, Mugilan, party, president, Seri, Sammy, Vellu, youth,step,said
8	gerakan, member, MIC, chief, MIC, party, president Datuk, Seri, Sammy, Vellu, youth

The overlap of keywords among the eight reuters articles is measured to ensure a minimum percentage of overlapping is fulfilled. That means, if there are 100 reuters articles, only those articles that have their collections of keywords to be the same of a certain percentage will be selected. Once the set of reuters articles are determined (in this case, Reuters No 1 - 8), the mapping of each article to one another is again performed to search for the keyword with the highest occurrence for the whole set (in this case, eight articles). Based on the results in Table-1, the following keywords are selected as associated words if 100% is setas the required percentage: Sammy, Vellu, President, MIC, Indian in which in the real-world context, Sammy Vellu is the President of MIC representing Indian community.

News network on frequency-based keywords extraction

The algorithm for the frequency-based keywords extraction is performed on 925 articles. From these articles, 1097 keywords have been determined in which some of them are repetitions. Table-2 shows the example of the two articles (Reuters 1 and Reuters 2) where 10 keywords are duplicated in both articles. So, for this case, the strength $\theta$is calculated as 0.47.

Table-2. Keywords based on frequency.

Article	Word	Count	Level	Ordering	Frequency
reut 1	anti	1	2	2	0.1
reut 1	anti-sami	1	1	4	0.1
reut 1	datuk	1	1	5	0.1
reut 1	ga	2	2	2	0.3
reut 1	gerakan	2	1	2	0.3
reut 1	Indian	4	1	1	0.5
reut 1	mic	2	2	3	0.3
reut 1	parti	1	1	5	0.1
reut 1	presid	1	2	3	0.1
reut 1	said	3	2	3	0.4
reut 1	sami	8	1	4	1
reut 1	seri	1	1	4	0.1
reut 1	vellu	3	2	2	0.4
reut 2	anti	1	2	1	0.1
reut 2	chief	2	1	4	0.2
reut 2	ga	9	1	1	0.9
reut 2	gerakan	2	2	1	0.2
reut 2	Indian	1	2	1	0.1
reut 2	member	2	1	5	0.2
reut 2	mic	6	1	1	0.6
reut 2	mugilan	5	1	2	0.5
reut 2	parti	3	1	3	0.3

Event detection and named-entity news matching

Events are described by distinctive words that may represent the activity, the name of person, building, society, location or building. The words are not necessarily those with high frequency or appear in significant position such as title. These words are difficult to be detected and they are recognized based on the background knowledge of the reader. There are two phases to identify these words - manually and automatically. In the manual approach, eight articles that describe similar events are identified and plugged into 925 reuters news articles. Five keywords are determined from the subjective analysis of the eight articles. The event detection algorithm searches the articles of similar events based on these five keywords. Our result has shown that the precision is 0.89 and recall is 1. While this result for the precision and recall are considered to be high, the performance may vary on different data sets.

CONCLUSION AND FUTURE WORKS

The technology for news delivery has received lots of attention by the industrial and academic researchers. The common interest among the technologies is to deliver the right news to the right recipient with the right preferences and at the right time. Hence, the right keywords are essentials to ensure the right news are retrieved and using the syntactical or synonymic approach is inadequate. The context of words associated by the person, location and event is necessary to ensure that words are related by how they appear in the news in that contemporary. For example, news that related "president of united states" and "donald trump" may not be relevant prior to 2017. It is also argued that using common thesaurus or words taxonomy may not be sufficient to understand the context. The work presented in this paper describes the approach of using words association which is contextual-based. Greater accuracy can be obtained with more detail information given to the associated keywords such as descriptive labeling of news network relationship, stratified news structure and event and named entity. The recall and precision are reported to be reasonably high because the associated keywords are selected based on the context and not merely using statistical approach. In the case of the news stories about the movement to protest "Datuk Sammy Vellu", the retrieval of the news is highly accurate since the context of the story is rather unique. It is expected that the precision and recall may vary depending on the variations of the context of the news story. For example, if there are more than one movement who makes the protest, then there will be more sets of associated keywords that describe about the same context. In this case, if we want to achieve high precision and recall value, the set of associated keywords must be extracted from separate set of news. This leads to the limitation of our approach that the introduced algorithm works well for news which is focused to single issue and event. A future enhancement to the algorithm is to look at the news with multiple events such that the aggregation of keywords association could be more complex. Hence, this requires the process of disassociation of words which has lower probability of being an associated keyword. Another extension of the work is to improve on the named-entity relationship. Since the focus in this work is on keywords extraction and words association, the naming of the entity is tailor made for the experimental set. Name-entity relationship is another area by itself and focus on news is an interesting area (Shrestha and Vulic, 2013; Parvez, 2017). Another recommended future work is to build a communicative news reader that allows natural language interaction between the news reader and the users on the content of the news considering solid understanding of the news content on the contextual aspects.

ACKNOWLEDGEMENT

We would like to thank Taif University for giving the opportunity to perform the research using available resources.

REFERENCES

Abdul-Hamid A, Darwish K. 2010. Simplified feature set for Arabic named entity recognition. In: Proceedings of the 2010 named entities workshop. Association for Computational Linguistics, Uppsala. pp. 110-115.

Alotaibi F, Lee M. 2012. Mapping Arabic Wikipedia into the named entities taxonomy. In: Proceedings of COLING 2012: posters, Mumbai. The COLING 2012 Organizing Committee. pp. 43-52.

Becker M., Hachey B., Alex B. and Grover C. 2005. Optimising selective sampling for bootstrapping named entity recognition. In Proceedings of the Workshop on Learning with Multiple View, ICML, Bonn, Germany. pp. 5-10.

Beck-Fernandez H., Nettleton D.F., Recalde L., Saez-Trumper D and Barahona-Penaranda A. 2017. A System for Extracting and Comparing Memes in Online Forums. Expert Systems and Applications. 82, 231-251.

Biswas S.K., Bordoloi M. and Shreya J. A graph based keyword extraction model using collective node weight. Expert Systems with Applications. 97: 51-59.

Brenna L., Demers A., Gehrke J., Hong M., Ossher J., Panda B., Riedewald M., Thatte M. and White W. 2007. Cayuga: a high-performance event processing engine. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data (SIGMOD''07). ACM, New York, NY, USA, 1100-1102. DOI: https://doi.org/10.1145/1247480.1247620

Chen Ping-I and Lin Shi-Jen. 2010. Automatic Keyword Expansion using Google similarity distance. Expert Systems with Applications. 37, 1928-1938.

Chen Yi-Hui, Lu, E. Jui-Lin and Tsai M.F. 2014. Finding keywords in blogs: Efficient keyword extraction in blog mining via user behaviors. Expert Systems with Applications. 41, 663-670.

Davies M. 2011. Word frequency data from the Corpus of Contemporary American English (COCA). http://www.wordfrequency.info

Ekbal A and Bandyopadhyay S. 2010. Named Entity Recognition using Support Vector Machine: A Language Independent Approach. International Journal of Electrical, Computer and Systems Engineering. 4(2): 155-170.

Fan XX., Chow KP., Xu F. 2014. Web User Profiling Based on Browsing Behavior Analysis. In: Peterson G., Shenoi S. (eds) Advances in Digital Forensics X. DigitalForensics 2014. IFIP Advances in Information and Communication Technology, vol 433. Springer, Berlin, Heidelberg.

Gabbard R., De Young J., Lignos C., Freedman M and Weischedel R. 2017. Combining rule-based and statistical mechanisms for low-resource named entity recognition Machine Translation (2017). https://doi.org/10.1007/s10590-017-9208-0.

Habibi M., Weber L., Neves M.,Wiegandt D.L., Leser U. 2017. Deep learning with word embeddings improves biomedical named entity recognition, Bioinformatics. 33(14): 37-48.

Hüser C., Weber A. 1992. The Individualized Electronic Newspaper: An Application Challenging Hypertext Technology. In: Cordes R., Streitz N. (eds) Hypertext und Hypermedia 1992. Informatikaktuell. Springer, Berlin, Heidelberg.

Kozareva Z. 2006. Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists. Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, Trento, Italy. pp. 15-21.

Lample G., Ballesteros M., Subramanian S., Kawakami K. And Dyer C. 2016. Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260-270, San Diego, California, June. Association for Computational Linguistics

Lee H.J. and Park S.J. 2007. MONERS: A News recommender for the Mobile Web. Expert Systems with Applications. 32(1): 143-150.

Lee L.H., Isa D., Choo W.O and Chue W. Y. 2012. High Relevance Keyword Extraction facility for Bayesian text classification in different domains of varying characteristics. Expert Systems and Applications. 39, 1147-1155.

Matsuo Y. and Ishizuka M. 2004. Keyword Extraction from a single document using word co-occurrence statistical information. International Journal of Artificial Intelligence Tools. 13(1): 157-169.

Mohit B. 2014. Named Entity Recognition. In: Zitouni I. (eds) Natural Language Processing of Semitic Languages. Theory and Applications of Natural Language Processing. Springer, Berlin, Heidelberg.

Montes-García A, Álvarez-Rodríguez J.M., Labra-Gayo J.E., Martínez-Merino M. 2013. Towards a journalist-based news recommendation system: The Wesomender approach. Expert Systems with Applications. 40(17): 6735-6741.

Morales G. D. F., Gionis A., Lucchese C. 2012. From chatter to headlines: harnessing the real-time web for personalized news recommendation, Proceedings of the fifth ACM international conference on Web search and data mining, February 08-12, 2012, Seattle, Washington, USA. [doi>10.1145/2124295.2124315].

Oflazer K and Tür G. 1996. Combining hand-crafted Rules and Unsupervised Learning in Constrained-based Morphological Disambiguation. Proceedings of ACL-SIGDAT Conference on Empirical Methods in Natural Language Processing. pp. 69-81.

Parvez, Shamima. 2017. Named Entity Recognition from Bengali Newspaper Data. International Journal on Natural Language Computing (IJNLC). 6(3): 47-56.

Rennison E. 1994. Galaxy of news: an approach to visualizing and understanding expansive news landscapes, Symposium on User Interface Software and Technology, Proceedings of the 7th annual ACM symposium on User interface software and technology. pp. 3-12.

Romero M., Moreo A., Castro J.L and Zurita J.M. 2012. Using Wikipedia concepts and frequency in language to extract key terms from support documents. Expert Systems and Applications. 39, 13480-13491.

Sebastiani F. 2002. Machine Learning in automated text categorization. ACM Computing Survey. 34(1): 1-47.

Shrestha N. and Vulic I. 2013. Named Entity Recognition in Broadcast News Using Similar Written Texts. Proceedings of the Student Research Workshop associated with RANLP 2013, pages 142-148, Hissar, Bulgaria, 9-11 September.

Time Magazine / CNN. The 10 most endangered newspapers in America. (2009, March 9). Available at: http://www.time.com/time/business/article/0,8599,1883785,00.html

Trampus M. and Novak B. 2012. Internals of an aggregated web news feed. In Proceedings of 15th Multiconference on Information Society 2012 (IS-2012).

Wen H., Fang L. and Guan L. 2012. A hybrid approach for personalized recommendation of news on the Web. Expert Systems with Applications. 39(5): 5806-5814.

Zhang C., Wang H., Cao L., Wang W. and Xu F. 2016. A hybrid term–term relations analysis approach for topic detection. Knowledge-based Systems. 109-120.