International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

# Automated Ontology Building for News Text using Associative Word Properties

**S.M.F.D Syed Mustapha, Abdulmajeed Alsufyani**

***Abstract*: It is known in the literature that ontology had been used extensively in machine learning for performance enhancement for text retrieval. It is also shown that robust ontology with detailed description of the domain knowledge will contribute to the accuracy in the retrieval. Nevertheless, we argue in some domain such as news text retrieval, building an ontology manually can be costly for a large-scale news repository and especially with the changes in content due to the dynamic events. In addition, maintenance can be a dauting task to keep up with new words that are associated with new events. This paper demonstrates the attempt to fully automate the development of an ontology for identifying the news domain and its subdomain. The ontology specification is defined based on the needs of the accuracy in retrieval. The mechanism of generating the ontology specification is defined and the results of the retrieval performance is discussed.***

***Keywords* : news text retrieval, ontology, information extraction and text mining.***

## I. INTRODUCTION

Ontology is a formal specification to describe the concepts of a domain knowledge. Traditionally, ontology is built for a specific knowledge domain and it is called domain ontology, for example in medical, economy, geology and others. Since, domain ontology is a very specialized area where knowledge concepts are defined at a lowest level of granularity, it is usually manually built by domain experts in the respective field. However, ontology has also been built for multiple knowledge domains such as in managing retrieval of news text. News texts are usually classified in various domains such as sports, business, religion, politics, economy, technology etc. In each domain, there are sub-topics which are called sub-domains. For example, in sports, the sub-domain could be football, cricket, golf etc. The sub-domain can be further classified by some associative words that form the event story. For example, the following news text called `Claxton Example`, the associative words are highlighted:

```
Claxton Example:

Claxton hunting first major medal

British hurdler Sarah Claxton is confident she can win
her first major medal at next month's European Indoor
Championships in Madrid.

The 25-year-old has already smashed the British record
over 60m hurdles twice this season, setting a new mark
of 7.96 seconds to win the AAAs title. "I am quite
confident," said Claxton. "But I take each race as it
comes. "As long as I keep up my training but not do
too much I think there is a chance of a medal." Claxton
has won the national 60m hurdles title for the past
three years but has struggled to translate her
domestic success to the international stage. Now, the
Scotland-born athlete owns the equal fifth-fastest
time in the world this year. And at last week's
Birmingham Grand Prix, Claxton left European medal
favourite Russian Irina Shevchenko trailing in sixth
spot.
```

For the first time, Sarah Claxton has only been preparing for a campaign over the hurdles - which could explain her leap in form. In previous seasons, the 25-year-old also contested the long jump but since moving from Colchester to London she has re-focused her attentions. Claxton will see if her new training regime pays dividends at the European Indoors which take place on 5-6 March.

The associative words are referred to as co-occurrence words where these words are in presence together in at least in one or more news texts. The collection of the associative words within the same news text forms a story event (which is Claxton hunting first major medal in this example). The associative words are made of N number of words where N = {1, 2 or 3}. Based on the `Claxton Example`, the words such as `60m hurdles` (N= 2) co-occur more than once in a single news text with `Sarah Claxton` (N = 2) and hence both are considered as associative words. Similarly, the other words such as `major medal`, `European Indoors`, `hurdles`, `British`, etc. It is hypothesized that the collection of these associative words is discriminative enough to classify the subdomain of the news collections. However, there is a challenge in managing words with high frequency such as stop words or function words as they also co-occur in all news text. To resolve this, stop words are easily removed using the stop words list. Nevertheless, in some cases these words are part of the associative words such as “University of York”, “Prince of Wales”,

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
**S.M.F.D Syed Mustapha***, Computer Science Department, College of Computing and Information Technology, Taif University, Makkah, Saudi Arabia. Email: smfdsm@gmail.com/syed.malek@tu.edu.sa
**Abdulmajeed Alsufyani**, Computer Science Department, College of Computing and Information Technology, Taif University, Makkah, Saudi Arabia. Email: a.s.alsufyani@tu.edu.sa

"The Yorkshire Hotel", "Port of New Orleans", and "Duchess of York". Hence, the mechanism in identifying the associative words have to also consider these factors.

The ontology is built to describe the properties of each associative word in terms of its co-occurrence with other associative word/s and the strength value of the co-occurrences; the behavior of the associative words in

terms of its co-occurrence behavior in other news texts within the same sub-domain of the main domain; and the relationships of news texts with other new texts based on the associative words within the same sub-domain and domain. The strength value between associative words is measured using the frequencies of each keyword towards each other and hence two associative words are considered "strong" if the total frequencies for both associative words are among the highest compared to other co-occurrences of other associative words. The other possible strength values are "medium" and "weak". In the `Claxton Example`, the `major medal` and `European Indoor` have strong relationship as both appear twice or having same number of occurrences in the news text and `Claxton` also has a strong relationship with `medal` as both have the highest number of frequencies compared to others. The former implies that the two associative words will be present concurrently as both contribute to the context of the news text while the latter indicates that both words are influential as they are the most frequently used in the news text. It is necessary to check these associative words on the relationship with each of every other associative word. The associative words with high frequencies in a particular news text are referred as *influencer*. Another property value described by the ontology is that associative keywords with high frequencies will have their profiles which contain the information on a) the number of other associative words that has a "strong" "medium" or "weak" relationship with. This information implies that the associative words that have more "strong" relationships with other associative words in a similar news text indicates that it has strong relevancy to the collection of associative words that form the uniqueness of the news text. For example, `Claxton` and `medal` have the highest relevancy compared to other associative words as both have the high frequencies such that they will become influencer and will have relationship with each possible associative word. There are three possible values in describing relationships – strong, medium and weak depending on the frequencies. Identifying associate word with high relevancy is essential as it has more weightage in discriminating the news text with other news texts; b) its presence (i.e. the associative word) in other news text to indicate i) the uniqueness of the associative word that it can play the role as discriminative word against other news texts ii) the representative of the associative word where these words are also present in other news texts that have the similar context. For example, `Claxton` will be a discriminative word that eliminates other news text within the same sub-domain (sports) that do not describe about `Sarah Claxton` and also at the same time, it is representable associative word to determine other news text of similar context to `Sarah Claxton`. The ontology that is described in this paper is much less sophisticated that those ontologies that are built manually but the intention is to be able to extract the relevant news text using the ontology that is fully built automatically using associative words. Hence, the research questions are a) what is the possibility that an ontology can be built automatically as opposed to the manual approach in news text extraction and what are the properties that needed to be described by the ontology b) will the associative words sufficient to discriminate the news text at subdomain level such that there are *AW*s that are unique to a domain or a particular news text? In the subsequent sections, related works are discussed as comparison to our approach, technical descriptions of our approach in building ontology and finally the experimental results in measuring the capabilities in building ontology as well as extracting the relevant news.

## II. ONTOLOGY-BASED TEXT RETRIEVAL

The literature survey focuses on the use of ontology for news categorization and classification. InfoMap is an ontology that represents the concepts and related sub-concepts where two words are associative based on the hierarchy position. For example, "tire" and "car" are associative based on the concept of a car taxonomy [1]. Since it has a detailed taxonomy of a car, it is used by many NLP (Natural Language Processing) applications. The advantage is that it can detect the semantic relationship of words appear at every sentence level. The disadvantage is the ontology requires human expertise to manually build the ontology which can be practically expensive to build for each taxonomy for every domain. Another work on using taxonomy as the ontology for news text is ePaper [2]. It uses the ontology to match the users' profiles against the news content. The ontology depends heavily on the news metadata structures defined by IPTC (www.iptc.org). Our argument is that there is a price to pay as it is a commercial entity and the news are limited to NewsCodes defined by IPTC which are mainly Reuter news. The work done by Goosen et. al [3] is to use concept frequency rather than term frequency where the concepts of the terms are determined from the lookup with WordNet semantic network. It claims that using concepts, the word sense ambiguation could be avoided. Another work on news personalization is reported by Schouten et. al [4] using Hermes framework. The framework is pre-built ontology with concepts that are managed by the experts. The maintenance for this framework requires manual intervention that makes difficult to keep abreast with the new domain in news. Ontology is also used to determine the weightage of a term which was extracted from web pages [5]. Ontology-based term weighting had shown significant improvement to the traditional TF-IDF approach where the F-measure is between 85 – 93% when ontology is used. Wijewickrema and Gamage used ontology to solve the problem with word ambiguity [6]. It used Lucene API to determine the subject of a document in which the subject label was sent to Protégé-OWL API to retrieve the relevant ontology. The initiative to fully automate the subject and minimizing the word ambiguity of the document is commendable but it relies on the availability of the domain defined in the ontology. Similar work by Song, et. al where ontology is used to define the main concepts of the terms and the relationships [7].

It is essential to elaborate here as the same ontology that we could develop closely to this work. For example, the concept is defined as the most common words being used in the domain, the feature of the individual term, the relation to the concept (most common word) and possibly the class and instance of the word. However, to do this, the challenge is that while Song worked on document with rich of terms while our work is on news text which can be as short as few lines. An effort towards developing classes of a domain was reported in by Vogrinčič and Bosnić [8]. This work signifies our effort in developing a method to build classes automatically as the way forward in text or document classification. However, the work was not a fully automated ontology building as several methods such as k-means clustering, querying interactively, visualization using Onto-Gen and manual assignment were used to build an initial ontology. This method requires a retraining of the new data sets and manual update of the ontology which is not appropriate for managing an evolving domain such as news text. Nyberg et. al defined the document ontology in terms of the relationships of terms and documents using the principles of hyponym, meronym and associative [9]. The work by Allahyari et. al [10] aspired the work to be presented in this paper. The three approaches adopted from this work are i. identifying the entities from using the Wikipedia ontology ii. formation of thematic graph iii. categorization of the topics to the thematic graph where these ideas will be altered to suit our problem. Other method in information extraction is mentioned by researchers dealing with various real world problems [12, 13]. The literature had shown that using ontology for news retrieval had improved the performance in terms of retrieval accuracy. Nevertheless, we would argue that in certain domains and languages, building ontology manually can be time consuming or expensive. Building ontology automatically by using merely the available sources within the available internal data and not outsourcing to external data such as Wikipedia, WordNet or any pre-built taxonomy is a new initiative. In our effort, associative words are mainly used in building the ontology. We begin with formalizing the concepts of associative words and demonstrates how they are practically used in determining the domain and subdomain of the news text. The research objectives (RO) are as the followings:

RO1 – determine the associative words for each domain and news text while excluding the function words or stop words;

RO2 – build the properties of each associative word in terms of the relationships with other associative words;

RO3 – determine that the concept of *AW*s is feasible for at least 90% of the total news for each domain;

For RO1, some of the associative words contain function words or stop words, the success is measured based on the ability to extract associative words while excluding other high recurrence words which are not meaningful to be associative words. RO2 investigates whether the concept of associative words exist in every news text as the technique heavily rely on associative words in building ontology. This also determines the unique property of associative words as they are discriminative which is essential to be able to retrieve relevant news texts with higher precision. RO3 will demonstrate the capabilities of the associative words-based ontology in terms of recall and precision.

## III. METHODOLOGY AND TECHNIQUE

In this section, the formal definition and the techniques of determining the associative words are given in this section.

**Definition 1**. Associative Word (*AW*) is defined as a word that has more than one occurrence in a news text and it is not function words or stop words unless it is part of the associative words such as "Duchess of York" . $AW^K$ where $K = (1,2,3)$ is an associative word that contain one or two or three words together. For example, if *Knowledge Management* is associative word, then it is $AW^2$. News text is a single release of news with specific date and place and also may contain one or more paragraph for a single news title. (Note: for convenience, the superscript *K* is not used all the times).

**Definition 2**. Relationship between two associative words, *AW*, is denoted as $\rho(AW_i, AW_j) = \{s \mid m \mid w\}$ where *s*, *m* or *w* is a relationship value indicating *strong*, *medium* and *weak* respectively.

**Definition 3**. Given an associative word $AW_i$, then η is the total number of associative words that has *s*, *m* or *w* relationship with $AW_i$ and be described as $\rho s(AW_i, \eta)$, $\rho m(AW_i, \eta)$ and $\rho w(AW_i, \eta)$. For example, $\rho s(AW_i, 5)$ indicates that the associative word $AW_i$ has 5 associative words that have *strong* relationship with it.

**Definition 4**. A frequency, *f*, of an associative word describes the number of occurrences the associative word AW in a single news text and it is denoted as *f*(*AW*).

**Definition 5**. Two associative words have *strong* relationship denoted as $\rho(AW_i, AW_j) = s$, provided the following conditions are fulfilled:

i. $f(AW_i) \geq f(AW_j) > f(AW_k) > \ldots f(AW_p)$ or
ii. $f(AW_i) = f(AW_j)$.

In (i), $f(AW_i)$ and $f(AW_j)$ are the two associative words with highest frequencies among the *AW*s in the news text; and (ii) states that both AWs have the same number of occurrences in the news text. The other relationships value, $m < s$ and $w < m$ which are determined by the medium and low frequency values. For example, if the highest frequency score is 7 and the lowest is 1, then the medium will be 4. The *AW*s will be assigned to the strength that it is closed to, for example, *AW* with frequency equal to 5 will be assigned to medium value.

**Definition 6**. $AW_i$ and $AW_j$ are *influencer* if $f(AW_i) \geq f(AW_j) > f(AW_k) > \ldots f(AW_p)$ which states that both frequencies for both $AW_i$ and $AW_j$ are among the top two.

**Definition 7**. Given a domain of a collection of news text, $\lambda^i(NT) = \{nt_1, nt_2, \ldots, nt_k\}$, $\exists AW_k \in \lambda^i(NT)$ where $AW_k \in \forall nt_i$ which states that if there are some associative words, $AW_k$, which are elements of associative words in domain $\lambda^i$ and $AW_k$ MUST exist in all news text, $nt_i$, then $AW_k$ are the associative words of domain $\lambda^i$ denoted as $AW_k^{\lambda_i}$.

**Definition 8**. Given a collection of news text, $NT = \{nt_1, nt_2, \ldots, nt_k\}$, such that $\exists AW_k \in nt_j$ and $\exists AW_k \notin nt_i$ where $\{i = 1, 2, \ldots k\}$ and $i \neq j$ which means there are some *AW*s that appear in ONLY one news text i.e. $nt_j$ but not others, then $AW_k$ is unique discriminative associative words, named as UDAW. UDAW are unique associative words of a subdomain for a particular news text.

**Definition 9**. Given a collection of news text, $NT = \{nt_1, nt_2, \dots, nt_k\}$, such that $\exists AW_k \in nt_j$ and $\exists j$ where $\{j = 1, 2, \dots k\}$ which means there are some *AW*s that appear in MORE THAN one news texts but NOT ALL news texts then $AW_k$ is non-unique discriminative associative words, named as NUDAW. NUDAW are associative words of a subdomain for some news texts.

### A. Methodology

The news text under this investigation is a taken from the BBC news repository, hence it is not an online version. There are pre-processing procedures which are modification from the standard text processing method as shown in Fig. 1 which shows the additional step of determining the associative words, $AW^K$ with K = {1 | 2| 3}. The first phase is to determine these associative words based on **Definition 1** using the frequencies as stated in **Definition 4**. For *K* = 3, the word is scanned throughout the news text to find other occurrences. These are associative words with three words adjacent to each other and they may consist of stop words or function words. At this phase, they are extracted and stored as list of potential associative word. This process is repeated for K = 2. Subsequently, the stop words and function words removal are performed before the process K = 1 is performed. The removal is essential to avoid treating stop words or function words as associative words.

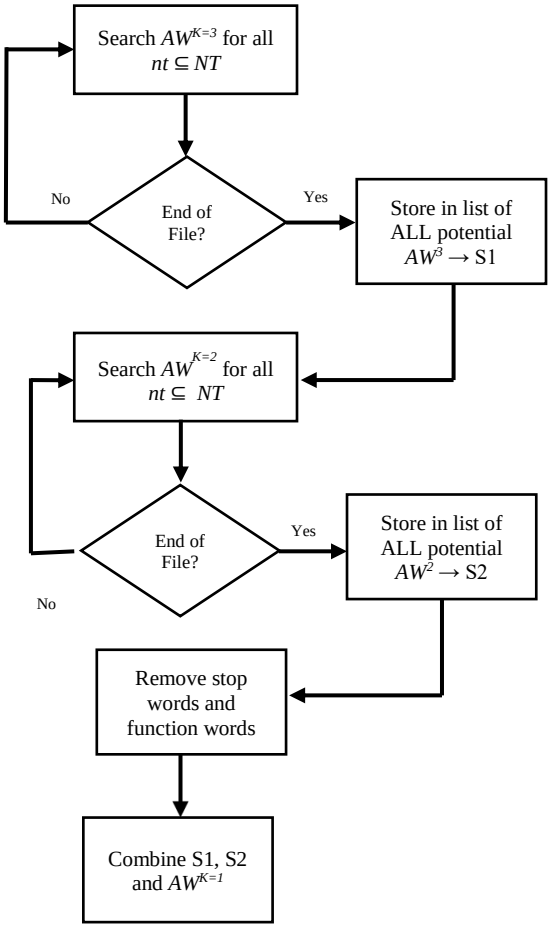


**Fig. 1. Once the AWs have been determined, the subsequent steps**

Process to determine Associative Words are to compute the properties which are the frequency for each *AW*, denoted as *f*(*AW*) and the relationships, denoted as ρ($AW_i$, $AW_j$). The relationship, *s*, is given in **Definition 5**. For relationship *m*, it is the median of the sorted frequencies for all *AW*s and *w* is the lowest in the sorted list as stated in **Definition 2**. The number of *s*, *m* and *w* can be subsequently obtained for each *AW* as given in **Definition 3**. The frequencies are calculated based on associative words of a particular news text, hence, an associative word $AW_k$ may be the highest frequency for one news text but not on the other news text. The frequencies are also used to determine the influencer which are used as

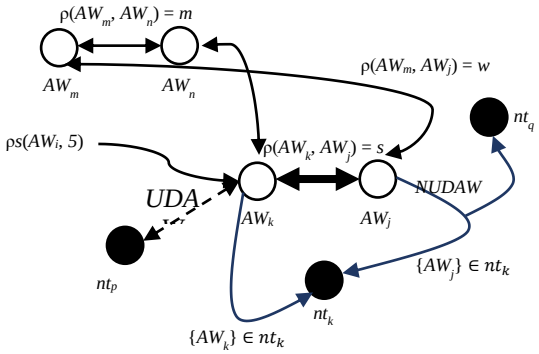


**Fig. 2. Associative Word based Ontology**

the central core *AW* when building ontology for a particular news text as stated by **Definition 6**.

Fig. 2 demonstrate the idea of building ontology to describe the relationship of *AW*s. Each circle represents the associative word which is labelled underneath. It shows three types of relationships in which the double-headed arrows have different thickness to indicate the strength. Each *AW* will be assigned to every *AW*, however it is not shown in the diagram to avoid the diagram from being cluttered. The ontology also describes the relation between the *AW* with another news text in which it is indicated as UDAW ($AW_k$ is UDAW to $nt_p$) or NUDAW ($AW_j$ is NUDAW to $nt_p$ and $nt_q$) as given in **Definition 8** and **Definition 9**. The ontology also implies that $nt_q$ and $nt_k$ are two interrelated news text by the associative words $AW_j$. ρ*m*($AW_i$, *5*) is part of the ontology describing the number of strong relationship with the associative word. Finally, based on the data quantitative data from UDAW and NUDAW, it is possible to determine the collection of *AW*s that exist in domain $\lambda$ as described in **Definition 7**.

### A. Data Preparation and Experiment

There are five domains (business, entertainment, politics, sport, tech) which contain 2225 news text. The data is available at http://mlg.ucd.ie/datasets/bbc.html [11]. Table I. lists the experiments, the expected outcome and the correspond research objectives. The first experiment is essential to prove that associative words can exist in news text as the definitions given in this work are merely a postulate. The second experiment determines that there exist some relationships between associative words that these relationships of the associative words can be used as representatives for extracting the news text. For example, if $AW_1$, $AW_2$ and $AW_3$ have strong relationships, they have more weightage to be used for matching other news text of similar context. The third experiment tested few selected news text that have similar context (by examining manually) and determined their associative words and these news text are called testing news text.

These associative words are used to find the news text in the collection of news text within the same domain by matching with their associative words. For example, $AW_1^T$, $AW_2^T$ and $AW_3^T$ are associative keywords from testing data and $AW_1^C$, $AW_2^C$ and $AW_3^C$ are associative keywords from collection of news text (for conveniences, the number of *AWs* are made the same, but necessary so). If the similarities are of the two are high, then the news texts from the testing group are considered to have the same context to the news texts from the collection of news text. The similarity is calculated based on the number of words that exact match.

**Table- I: Experiments addressing the Research Objectives**

| Experiments | Expected Outcome | Research Objectives |
|---|---|---|
| Determine the the number of *AWs* for each news text | Target: 90% of the news texts that have at least K = 1 to 3 | RO1 – determine the associative words for each domain and news text while excluding the function words or stop words. |
| Apply to the domains that can proof the existence of the properties | At least 80% of the domains that have the $\rho s(AW_i, AW_j)$ and $\rho m(AW_i, AW_j)$. | RO2 – build the properties of each associative word in terms of the relationships with other associative words. |
| Apply to all domain of all news texts | At least in average there are at least 3 *AWs* with *K*= 3 and 10 *AWs* with *K* = 2 | RO3 – determine that the concept of *AWs* is feasible for at least 90% of the total news for each domain. |

## IV. RESULT AND DISCUSSION

The first experiment is to determine whether the concept of Associative Words is realistic for the short document such as news text as the news text is usually short and the chances of having occurrences may be low. The associative words, mainly with strong relationships are essential to be representative words for the news text. We perform the checking on the associative word's extractions for each news text for each domain. For each domain, the number of news text are as follows: Business domain (510 news text), Entertainment domain (386 news text), Politics domain (417 news text), Sport domain (511 news text) and Technology domain (401 news text). The results for each domain are as shown in Table-II.

The associative words that are shown in Table-II are in any form of the K = {1 or 2 or 3}. It is observed that the numbers of news text that do not have any Associative Words are small. The average number of Associative Words per news text are representable since if the average number of words in a news text is within 200 to 500 words then the associative words will be 10 – 20% for medium size length or 3 – 8% for large news text. Subsequently, it can be concluded that the actual outcome has outperformed the predicted outcome which targeted at least 90%.

**Table- II: Results of the Associative Words for Domains**

| Domain | News text with Associative Words | News text with missing Associative Words | The number of Associative Words per Domain | Average number of Associative Words per news text |
|---|---|---|---|---|
| Business | 510 | 0 (0%) | 7557 | 14.5 |
| Entertainment | 385 | 1 (0.25%) | 8768 | 22.77 |
| Politics | 417 | 0 (0%) | 14292 | 34.27 |
| Sport | 509 | 3 (0.58%) | 8894 | 17.47 |
| Technology | 401 | 0 (05) | 14540 | 36.25 |

It is also important to show the finding on the *AW* where *K* = 3 and *K* = 2. The details in terms of the number of *AW* with *K* = 3 and *K* = 2 are tabulated in Table-III in order to have the idea whether such *AWs* exist. The results convincingly show that *AW* with *K* = 3 and *K* = 2 can be found in most news text that the combination of these words can be used as unique combined words as discriminating factors in extracting relevant news. The results show that the minimum average of *AW* with *K* = 3 is 4.21 which is higher than the expected outcome and *K* = 2 is 31.57 which is superior than the expected outcome.

**Table- III: *AW* with *K* = 3 and *K* = 2 for each Domain**

| Domain | *K = 3 (Average per News Text)* | *K = 2 (Average per News Text)* |
|---|---|---|
| Business | 2149 (4.21) | 16105 (31.57) |
| Entertainment | 4073 (10.55) | 18747 (48.56) |
| Politics | 4470 (10.72) | 28579 (68.53) |
| Sport | 2159 (4.22) | 16985 (33.23) |
| Technology | 3967 (9.89) | 26638 (66.42) |

**Table- IV: Strong and Medium Relationships for Domains**

| Domain | Strong Relationship | Medium Relationship |
|---|---|---|
| Business | 100% of all News Text | 31 News Text missing |
| Entertainment | 1 News Text missing | 37 News Text missing |
| Politics | 100% of all News Text | 8 News Text missing |
| Sport | 3 News Text missing | 72 News Text missing |
| Technology | 100% of all News Text | 1 News Text missing |

Another analysis is on the presence of strong relationship and medium relationship. We are interested with these two relationships as our proposition states that they are influential in representing the uniqueness of the news text. Table-IV represent the report on the presence of the strong and medium relationships for each domain. In addition, the concept of relationships between several *AW*s which have strong or medium values will also be used to enhance the retrieval capabilities.

The subsequent results show the findings on the *AW* relationships. The finding in Table-IV has enabled us to make a conjecture that *AW*s can be established in almost all news text as well as the relationship among them. In Business domain, all of the news texts are found to have strong relationship while there are 31 news text have missing medium relationships between the associative words. It is noted that the number of texts that do not have strong relationships are much less than the medium relationships and also the number of news texts that do not have strong relationship are rather very small (less than 1%). It is fortunate for our case that the strong relationships are extensively presence in almost all news text. The words result is on Sport domain where there are 72 (15%) news text that having missing medium relationships. However, the score is 85% of the news text having medium relationship which is higher than the expected outcome of 80%.

Overall, the experiments have addressed the research objectives and achieve desirable outcome from the expectation. However, there are few weaknesses on the finding especially on *AW* with $K = 3$. Since, as mentioned in our methodology, the stop-list words are removed after the *AW* with $K = 3$ is processed, therefore some *AW* do not carry any meaning or proper noun such *AW*s are "to understand the" , "as an electronic" etc. Nevertheless, since these words have high frequency compared to other words in the news text, these words may still be useful to represent the uniqueness of the news text for retrieval purpose.

## V. CONCLUSION

It can be concluded that the postulation which states that associative words with more than one terms specifically three terms and double terms can be established from the news text and the concept of associative words have proven in the five domains. The number of news texts which do not have associative words is negligibly small and will not affect the intention in building an ontology. It is demonstrated also that the relationships between *AW*s can be established as all of the domains show such circumstances. The relationships are essential as the formation of the combined words can be used to represent the text. The work contributes to the notion that ontology can be fully built automatically especially in the domain where the content changes dynamically such as news text, blog text, forums and others. The extension of the work could be applying to other types of documents (short or long) on various disciplines such as scientific paper, meeting minutes, essays and others. It will be interesting to determine the accuracy and precision of the ontology built using this technique on various disciplines.

## ACKNOWLEDGMENT

We would like to acknowledge our token of appreciation to the Computer Science Department and Taif University for the facilities and support during the period where the research was conducted. We also would like to thank our colleagues for the useful feedbacks on our research.

## REFERENCES

1. Shih-Hung Wu, Richard Tzong-Han Tsai and Wen-Lian Hsu, "Text categorization using automatically acquired domain ontology," International Review of Applied Linguistics in Language Teaching., Volume 41 (4) – Nov 6, 2003, pp 138 – 145.
2. L. Tenenboim, B. Shapira and P. Shoval, "Ontology-Based Classification Of News In An Electronic Newspaper", International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008, pp 89 – 97.
3. F. Goosen, W. IJntema, F. Frasincar, F. Hogenboom, and U. Kaymar "News Personalization Using the CF-IDF Semantic Recommender" WIMS'11 Proceedings of the International Conference on Web Intelligence, Mining and Semantics, May 25 – 27, Sogndal, Norway, Available: https://dl.acm.org/citation.cfm?id=1988701
4. K. Schouten, P. Ruijgrok and J. Borsje, "A Semantic Web-based Approach for Personalizing News" SAC 10, March 22 – 26, Sierre, Switzerland, 2010, Available: https://personal.eur.nl/frasincar/papers/SAC2010a/sac2010a.pdf
5. A. Qazi and R.H. Goudar, "An Ontology-based Term Weighting Technique for Web Document Categorization", International Conference on Robotics and Smart Manufacturing (RoSMA 2018), Procedia Computer Science 133, 2018, pp 75 – 81.
6. C.M. Wijewickrema and R. Gamage, "An Ontology-based Fully Automatic Document Classification System Using an Existing Semi-Automatic System" 79th IFLA General Conference and Assembly, Suntec Singapore, August 2013, pp 1 – 13.
7. M.H Song, S.Y Lim, D.J. Kang, and S.J Lee, "Ontology-based Automatic Classification of Web Documents" in D-S Huang, K. Li, and G.W. Irwin (Eds): ICIC 2006, LNAI 4114, pp 690 – 700.
8. S. Vogrinčič and Z. Bosnić, "Ontology-based Multi-label Classification of Economic Articles" COMSIS, Vol 8, No 1, Jan 2011, pp 101 – 119.
9. K. Nyberg, T. Raiko and T. Tiinanen, "Document Classification Utilising Ontologies and Relations between Documents", MLG'10 Washington, 2010, Available: https://dl.acm.org/citation.cfm?id=1830264
10. M. Allahyari, K.J. Kochut and M. Janik, "Ontology-based Text Classification into Dynamically Defined Topics" 2014 IEEE International Conference on Semantic Computing, 16 – 18 June, Newport Beach, CA, 2014, pp. 273-278.
11. D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.
12. Shabanaunnisa Begum. "Information Extraction from Text using Text Mining", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-9 Issue-2S5, December 2019, pp 23 – 24.
13. Ameen Abdullah Qaid Aqlan, B.Manjula. "Extraction and Analyze Text in Twitter using Naive Bayes Technique", International Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-9 Issue-4, February 2020, pp 1635 – 1639.

## AUTHORS PROFILE

**S.M.F.D Syed Mustapha** He is a full professor at Computer Science Department, College of Computing and Information Technology, Taif University. He obtained his BSc (Computer Science) from University of Texas, USA, MPhil and PhD from University of Wales, UK. He has published more than 80 papers in both conferences and journals in the area of Ontology Building, Knowledge Sharing Systems, Knowledge Management, Information Extraction and Text Retrieval. He had served as Dean of Faculty of Information Technology and School of ICT at two different universities and also Vice President (Operations and Technology).

**Abdulmajeed Alsufyani.** Abdulmajeed Alsufyani received the bachelor's degree (Hons.) in computer science from Taif University, Saudi Arabia, in 2006, and the master's degree in Computer Science and the Ph.D. degree in computer science from the University of Kent, U.K., in 2010 and 2015, respectively. He is currently an Assistant Professor of Computer Science at the College of Computers and Information Technology, Taif University, Saudi Arabia. He is also the Vice-Dean of Community College at Taif University. His research interests include Computational Intelligence, Computational Neuroscience, Machine Learning Algorithms.