Clustering Hashtags Based on New Hybrid Method and Power Links

Mahmoud Rokaya, Hamza Turabieh, Sanaa Al Azwari, Abdullah Alharbi, Mrim Alnfiai, Mohammed Alzahrani, Wael Osaimi and Wajdi Alhakam

***Abstract*— It is very important for various apps to cluster hashtags accurately. Some clustering methods depend on text properties. Since in social media there is complete freedom for users, there much spelling and grammar errors that might make dependence on lexical properties is useless. On the other hand, depending on the metadata of wordnet also affected by the users spelling and grammar errors. Hybrid methods might improve the accuracy of clustering for some extent. In this work, an un-supervised method for clustering hashtags based on text properties, semantic metadata, and power links is presented. The semantic method and lexical method will be combined in a strictly different way to produce a new hybrid method. The proposed hybrid method is supported through Power links to refine the clusters. The experiments proved that the proposed method outperforms each method individually and, also, outperform past hybrid methods. In all results, it is never happened that previous method achieved better results than the proposed method.**

***Index Terms*— Hashtag, Power Link, Semantic methods, Lexical methods, clustering, hybrid approach, wordnet, Social media.**

Manuscript received October 26, 2020; revised June 26 2021. This research was supported by Taif University Researchers Supporting Project number (TURSP-2020/254), Taif University, Taif, Saudi Arabia.

Mahmoud Rokaya is an associate professor at Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia. He works also at Tanta University, Faculty of Science, Tanta, Egypt. (corresponding author: Mahmoud Rokaya, +966502216186, mahmoudrokaya@tu.edu.sa)

Hamza Turabieh is a full professor at Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia. (h.turabieh@tu.edu.sa)

Sanaa Al Azwari is an assistant professor at Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944 (alazwari.s@tu.edu.sa)

Abdullah Alharbi is an associate professor at Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944 (amharbi@tu.edu.sa)

Mrim Alnfiai is an assistant professor at Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944 (m.alnofiee@tu.edu.sa)

Mohammed Alzahrani is an associate professor at Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944 (marzahrani@tu.edu.sa)

Wael Osaim is an associate professor at Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944 (w.osaimi@tu.edu.sa)

Wajdi Alhakam is an associate professor at Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944 (whakami@tu.edu.sa)

I. INTRODUCTION

CLUSTERING can be defined as the process of grouping similar items in groups. The similarity measure is different from one method to another. The most famous similarity measure is the Euclidian distance [1 and 2]. Clustering is very important for many areas of research and applications. Clustering is one main task in artificial intelligence, data mining, image processing, and machine learning. There are many methods to achieve the clustering task. Among the clustering methods, the famous techniques are Hierarchical Clustering Techniques [3], the K-means Technique [4], the Fuzzy C-means Technique [5], the Gaussian Expectation-Maximization Technique [6 and 7], the K-harmonic Means Technique [8, 9 and 10], Non-iterative Partitional technique [11] and Hybrid technique [4]. Among these methods, the Hierarchical Clustering Technique is adopted in this work.

A hashtag is a word, or a set of concatenating words prefixed by the symbol # to give the impact or related topic of the text message accompanying the hashtag [1]. Clustering of hashtags has a positive impact on real-world activities as it helps in improving the content rendering of timelines that appears for specific users. The methods of hashtag clustering can be categorized into two groups: lexical methods and metadata semantic clustering. The hybrid method combines the advantages of both the lexical method and the metadata semantic method. These methods are introduced in the related work section which explores some relevant work to the proposed method.

The proposed method, in this work, develops a modified hybrid method. Instead of forming the similarity matrix by giving an equal weight to each method, the actual similarity value in the similarity matrix related to the method is used. Also, the modified hybrid algorithm performance is supported by the Power Link measure. In the related work, we will explore what we mean by a Power Link and its history. Through explaining the method description, we will show how the Power Link will be used to enhance the modified hybrid algorithm performance. The expected impact of the Power Link came from the co-occurrence relationships that Power Link calculates between two words, two tweets and two hashtags. The experiments show that a modified hybrid clustering algorithm could overcome the performance of the other algorithms in all experiments in average and in the most individual experiments. Since the Power Link is related to the lexical content of the tweets it was expected to enhance the performance of the lexical

algorithm more than its ability to enhance the semantic metadata algorithm.

In section 2, the related work to the clustering methods and Power Links is presented. The details of the proposed and related methods are presented in section 3. Two types of experiments are presented in section 4. Gold standard test and pairwise disagreement test are used to evaluate the proposed method. Finally, the results are explored in section 5. The conclusion and future work are presented in section 6.

II. Related Work

This section reviews related work to Power Links, methods of lexical clustering, metadata semantic clustering and hybrid clustering.

A. Related work to Power Links

Power Link is a quantitative approach that uses the normalized and co-occurrence frequencies as well as the relative locations between various successive terms through a document to compute the relation between two words and to reflect the distribution of a term that occurs in a document [12]. The high value of the Power Link measure between two terms indicates a strong relation between these terms. If the value of the Power Link measure is low, the words in the text are not related to each other. The Power Link approach has various applications in different fields. In 2010, Rokaya and Atlam developed the Power Link approach as a field association tool, which uses the co-frequency and the distances of different incidence of words across the text to measure the association power relation between words in a document. The purpose of the Power Link algorithm is to determine and calculate the tendency of two words to occur together. Refining search engine results are one of the valuable applications that apply the Power Link approach. Rokaya developed an algorithm to enhance ranking of the results of search engines that use the Power Link approach [13 and 14]. This algorithm is based on the classification of terms in a certain field. Power Link was used in text summarization field as well. Rokaya presented an automatic method of summarization based on collocations and Power Link, which provides a custom summary based on users' needs [15]. Rokaya, Nahla and Aljahdali proposed a dynamic field association terms method for automatic summarization and text extraction based on Field Association terms and Power Links [16]. Moreover, the application of the Power Link approach was integrated into dictionary field and can support several languages. Rokaya and Nahla introduced a Multi-languages dictionary based on the Power Links field association method, which improved the quality of field association terms (FATs) to expose everyday words in languages [17]. Furthermore, Rokaya and Aljahdali proposed a real word spell checker using Power Link approach [18]. In addition. The approach of Power Links was used to extract semantic information from Arabic text. In 2017, Rokaya and Ghiduk developed an Arabic ontology tool with the support of Power Link concept, where Power Links were used to generate automatic ontology based on the co-frequency and the terms' dynamic distribution over a given text [21]. In 2016, researchers applied Power Link approach to extract patterns between terms in nutrition field by computing the link between two terms to find out how much they are related to each other and how often they occur together [22]. Rokaya [23] also proposed a new spam reduction method with the support of Power Link approach. This method was implemented to classify and detect the important and non-important emails. The results in that study showed that using Power Link approach generates improved spam reduction and more efficient classification for emails. Atlam et al., presented an approach using Power Link and co-word algorithms to improve the collocations by giving different weights to words in a document. This approach supports the application to achieve a 90% precision [24]. A considerable amount of research has been applied Power Links approach and it showed improvements over other Field Association methods. For the Power Link algorithm details we can refer to [12].

B. Related work to lexical clustering

In lexical clustering of hashtags, the tweets are the source of hashtags features [25]. Most of the adopted lexical methods are getting the features from tweets by adding the related tweets to a specific hashtag into one document. This document is called the tweets bag. This bag is represented as a vector space [26 and 33] which differs from one method to another depending on the co-occurrence of the hashtag senses in the same tweets [34 and 33]. The idea in lexical clustering depends on the fact that hashtags with a similar pattern of usage are semantically related. The temporal senses are derived from the content of the tweets not from the lexical pattern of the hashtag itself. Many works used the lexical clustering, for example, Wang et al. [36] used the co-occurrence of the hashtags for the purpose of sentiment analysis purposes. Similarly depending on the graph theory to represent the co-occurrence relations between words on tweets, Teufland Kraxberger [37] used clustering for event detection purposes. For tweet search, Park and Shin [27] used the graph representation for searching similar tweets. Rosa et al. [32] and Bhulai et al. [28] used clustering depending on document bag to cluster a set of predefined topics. The interesting point in this work that they found that expanding the URL found in tweets affects the clustering performance in a negative way. Also, since the size of the document bag becomes very large for most of the hashtags, the clustering process based on the document bag becomes a challenging task [38]. Tsur et al. [30 and 31] and Muntean et al. [26] used the concept of virtual documents for the purpose of hashtags clustering. The virtual document was built by appending tweets that contain a specific hashtag in one document. Then these documents represented using a vector space model. By implementing the concept of a virtual document with weights, Feng et al. [39] presented a method for hashtags clustering. Costa et al. [29] implemented a clustering of hashtags as a tool to classify tweets. Two tweets are considered similar if they belong to two hashtags in the same cluster.

C. Related work to Metadata based Semantic Clustering

Clustering of hashtags creates a controlled usage of the hashtags that can improve the quality of semantics and increase the frequency of their usage [40]. The variations

resulting from the uncontrolled use of hashtags present a challenge to clustering of semantics [41,45]. One of the two main approaches for clustering hashtags on social media is semantic metadata approach. Semantic metadata is defined as the metadata that defines the value of data and the names of things that can be articulated to represent such value [46]. It identifies the lexical semantics of hashtags based on external resources that are separated from tweet texts [29, 41-44]. Semantic metadata clustering is the ideal tool to convey information through social media [47]. For proper and effective performance of metadata-based method, hashtag quality and metadata quality factors play an important role in this approach. Javed and Lee [41] claim that metadata quality has a significant direct effect on the performance of this approach.

The approach of metadata-based semantic clustering is currently considered a comparatively new research area despite the increase in metadata availability. Therefore, there are a few existing studies that adopted this approach in clustering hashtags on social media [41-43]. For instance, Vicient and Moreno [43] applied semantic metadata approach to clustering hashtags. A metadata source proposed in their work used WordNet and Wikipedia to classify hashtag's lexical semantics. The proposed approach follows three phases, namely semantic grounding, construction of similarity matrix and semantic clustering [43]. In their work, clustering decisions are taken at the level of the word. The calculation of semantic similarity matrix was performed on the hashtags containing a minimum of one concept which presents a challenge of having incorrect clusters unless the similarity of concepts is found using the correct sense of the hashtag. To adjust on this, clustering is outlaid at the sense level. We should consider that similar words could have different meanings. Therefore, Javed and Lee [41 and 42] argue that clustering can be done based on the calculated similarity and correct senses or concepts. In the work of Javed and Lee [41], the clustering of hashtags was made at sense level. They identified a hybrid approach that performs clustering through complementing the contextual and lexical semantics of a given hashtag. Their approach presents a method that depends on two primary algorithms joined by a consensus where the first algorithm focuses on lexical semantics of metadata while the second method focuses on contextual semantics from text. The two results are pooled in a consensus to prove the semantics used for hashtags clustering [41]. Moreover, another related work carried out by Costa et al. [29] addressed the classification of a hashtag-based tweet by using semantic metadata. They utilized a crowdsourcing platform in their work to provide metadata and classify the tweets based on the metadata.

As it is evident above, metadata based semantic clustering is considered as an important resource that can be applied to resolve the limitation of lack of consistency in the data or the complexity of the challenge to be solved, and it has well proved to be working as applied [41-43].

D. Related work to Hybrid approach

Despite there are some existing studies in the literature that apply separate algorithms for data analytic [48-54], other use a combination of multiple approaches within the same domain [46]. The primary objective of using hybrid methods is to produce noticeable accurate findings belonging to accuracy and performance speed which could not be generated using a single method. For example, the author in [46] uses hybrid approaches of bio-inspired to identify spam profiles of twitter. Another work presented in [44] in which a hybrid algorithmic approach based on Naïve Bayes and Random Forest is applied on Twitter datasets. The produced findings indicate that both accuracy and efficiency factors were improved using the hybrid approach compared to the Naïve Bayes classifier algorithm when applied separately. Although, there are a number of available approaches based on textual and lexical conducted by researchers, a combination of these two methods is introduced recently and has not received much study in the literature [41]. For instance, Javed and Lee [41] use a hybrid semantic clustering algorithm where the lexical and contextual semantics of a hashtag is used to create accurate clusters of input data. Their results demonstrated that the hybrid algorithm performs better than both the text-based and the metadata-based algorithms alone. The Method

The proposed algorithm in this work depends on sematic clustering, lexical clustering, hybrid clustering and Power Links. This section introduces the details of each individual algorithms and how these methods are combined as a full clustering method.

A. Clustering based on semantic metadata

Our proposed clustering method for hashtags depends on Wordnet and Wikipedia information related to the hashtag name [41]. Fig. 1 summarizes the main steps of the algorithm, which can be divided into three steps as follows:

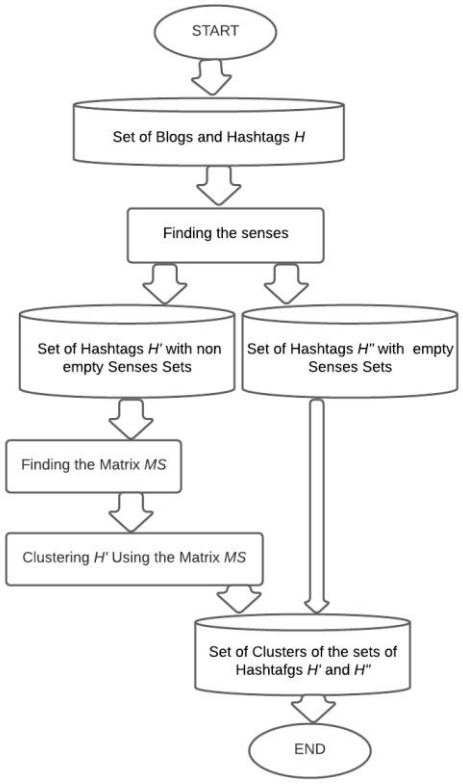

Fig. 1. Semantic Clustering Algorithm

a. Finding the senses.

Suppose we have a set H of hashtags, for each hashtag h we find the senses related to h as follows:

i) If there are senses in WordNet corresponding to h, these senses are retrieved and listed in a set h_s (hashtag senses set). If h_s is empty go to step ii.

ii) If there are no senses corresponding to h and h can be divided into multiple words, we drop the rightmost word and go to step i till all words that form h are exhausted and return h_s. If h_s is still empty, go to step iii

iii) Look in Wikipedia, if a topic related to h is found, find all auxiliary topics and retrieve the nouns in the title of these topics. For each noun, retrieve the senses from WordNet and append them to h_s.

iv) If $h_s$is still empty drop h itself from further calculations.

b. Finding the similarity matrix.

After dropping the hashtags with empty senses sets, the set of hashtags H is divided into two sets H' and H''. H' contains all hashtags with non-empty senses sets and H'' is the set of hashtags with empty senses set. The similarity between each two hashtags h_i and h_j is defined as the max sense similarity between pairs of senses, one from the senses set $h_{si}$of h_i and one from the senses set h_{sj} of h_j. Let s_p from the senses set of h_i and s_q from the senses set of h_j that corresponding to the max sense similarity between h_i and h_j a list called similarity list $LHsim$ can be defined as $LHsim =< h_i.s_p, h_j.s_q, sim(s_p, s_q) >$, i and j carry over all hashtags in H'.

c. Clustering:

i. We form the similarity matrix MS between hashtags based on the similarity list $LHsim$ as follows: let $SS = \{s_1, s_2, \ldots\ldots, s_n\}$ be the set of all differences, the dimension of the similarity matrix MS is $n \times n$ the value element $MS(i,j)= sim(s_i, s_j)$ if $< h_l.s_i, h_k.s_j, sim(s_p, s_q) > \in LHsim$ for some hashtags h_l and h_k in H', otherwise, $MS(i,j)$ is set to be zero.

ii. Perform hierarchical clustering of the H'using the similarity matrix MS and extract the flat clusters using a given threshold.

B. Clustering based on text properties.

Fig. 2 shows the basic step of clustering based on textual properties; these steps can be explained as follows:

a. Hashtag document: For each blog b_k that references a hashtag h_i append b_k to the hashtag document D_i of h_i. If b_k references two different hashtags h_i and h_j, then b_k will be appended to both D_i and D_j.

b. Preprocessing step: A spelling checker will be applied to correct the misspelled words. If there is no possible correction, the word will be added to the current dictionary. All stop words are deleted. Here a stop word is any word that appears in the majority of the blogs. Controlling the term majority to be a variable percent β%. Also, a word will be replaced by its stem if we can get the corresponding, otherwise the word will be replaced by itself.

c. Vectorizing: All unique words in all hashtag's documents are extracted and sorted. Let n be the number of unique words. For each hashtag h_i a vector v_{hi} is initialized to be n zeros, $v_{hi}(l) = f(w_l)$ in D_i, where $f(w_i)$ is the frequency of w_i in the document D_i

d. The similarity matrix ML between hashtags is calculated based on the cosine between the corresponding hashtag vectors.

$$ML(h_i, h_j) = \cos(v_{hi}, v_{hj})$$

Clustering: Perform hierarchical clustering of H using the similarity matrix ML and extract the flat clusters using a given threshold

C. Hybrid Approach

Semantic methods are affected by the formality of WordNet, and Wikipedia. In many cases, the users in their writing are very far from this formality and hence the target hashtag might be clustered in the wrong place. From the other hand, the informality, and the great freedom of users in writing their blogs might mislead the clustering process since the user made many mistakes that might lead to the wrong correction or the wrong stem and hence to the wrong cluster. To reduce the side effects of each method categories [41] proposed a hybrid method. This method will be adopted with some modifications. The original method gave an equivalent weight to both methods. In our proposed method, a weight is given for each method based on the distance between the hashtag and the center of the cluster. The steps below and Fig. 3 explain the detail of the algorithm.

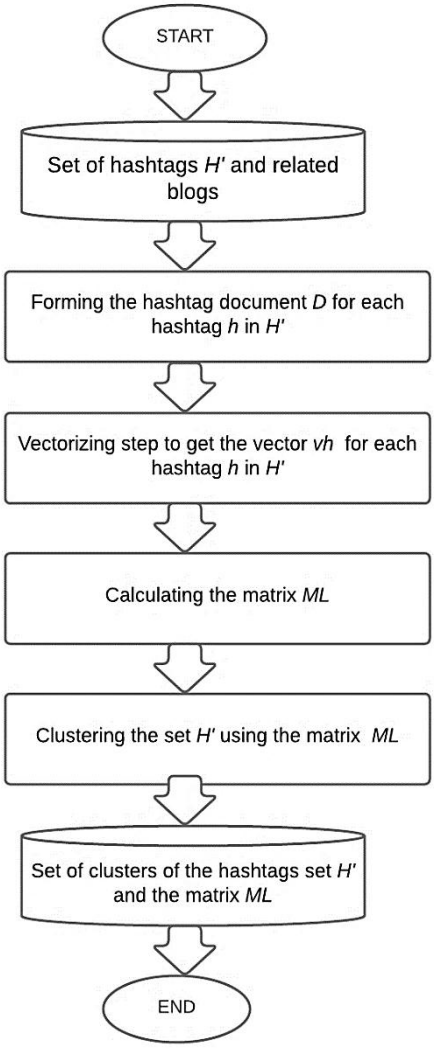

Fig. 2. Lexical clustering algorithm

a. Sematic Clustering: Perform semantic clustering to the set H to get the ground base hashtag set H' and the sets of flat clusters CSH.

b. Lexical Clustering: Perform a lexical clustering to the set H' and retrieve the set of flat clusters CLH

c. Hybrid similarity matrix: Initiate the similarity matrix MH to be zeros. The dimension of MH is $|H'| \times |H'|$. If the cluster of the hashtags h_i and h_j based semantic clustering is the same cluster then the corresponding element $MH(i,j)$ is increased by $MS(i,j)/2$ and if the cluster of the hashtags $h_i$and $h_j$based on the lexical cluster is the same cluster then the corresponding element $MH(i.j)$ is increased by $ML(i,j)/2$. So, instead of adding 0.5 corresponding to each algorithm, weighted values based on the original similarity matrices are added. It is clear if neither sematic algorithm nor the lexical algorithm mapped $h_i$and h_j to the same cluster then the value of $MH(i,j)$ is set to zero.

Hybrid clustering: Perform hierarchical clustering of the H' using the similarity matrix MH and extract the flat clusters using a given threshold.

D. Power Links

For two words w_1 and w_2, the Power Link $PW(w_1, w_2)$ is given by the number of blogs where w_1 and w_2 appeared together divided by the total number of blogs in a given corpus.

$$PW(w_1, w_2) = NB(w_1, w_2)/N$$

Where, $NB(w_1, w_2)$ is the number of blogs that each of them contains w_1 and w_2 and N is the number of blogs. For all words in a corpus, the Power Link matrix is defined as

$$M(i,j) = PW(w_i, w_j)$$

For each blog b a Power Link vector Pb is defined as

$$Pb(i) = \begin{cases} \dfrac{\sum_{j \neq i} M(i,j)}{n} & if\ w_i \in b \\ 0 & if\ w_i \notin b \end{cases}$$

,

Where, n is the total number of words in the current corpus, $i = 1,2,3, \dots, n$.

Suppose the number of blogs is N, $B = (b_1, b_2, b_3, \dots, b_N)$ is an ordered tuple of all blogs.

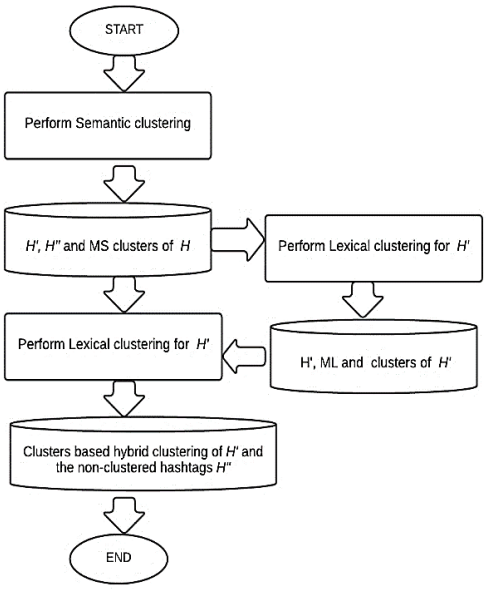

Fig. 3. Hybrid Clustering Algorithm.

For each hashtag hg, a Power Link vector Phg is defined as

$$Phg(k) = \begin{cases} \dfrac{\sum_l Pb_k(l)}{n} & if\ hg \text{ referenced by } b_k \\ 0 & if\ hg \text{ not referenced by } b_k \end{cases}$$

Where $k = 1,2, \dots, N$

Now, each hashtag hg is represented by a vector Phg and it is possible to calculate the Euclidian distance between the vectors Phgs.

Consider that we have C clusters and suppose a given hashtag hg belongs to C_i cluster, the distance between hashtag hg and a cluster C_j is average distance between the vector Phg corresponding to the hashtag hg and all vectors corresponding to the hashtags in C_j, if the distance between hg and C_j is less than the distance of hg and other clusters, then hg will immigrate from C_i to C_j. This process is called refining clusters based on Power Links.

The method keeps track of the original cluster C_i of each hashtag till all hashtags are reallocated in their proper cluster. This process can be applied recursively on the output cluster sets till the output becomes identical to the input. Fig. 4 explains the steps of refining based on Power Links.

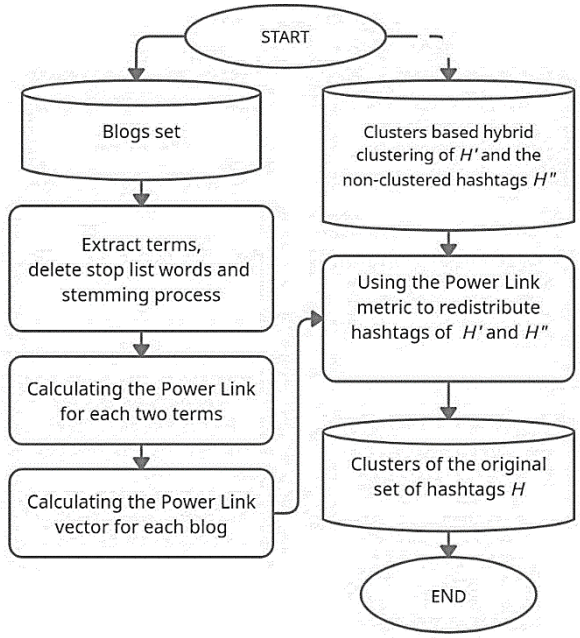

Fig. 4. Clustering based on power link metric.

E. Full clustering Method

Fig. 5 summarizes the steps of the full clustering method. The method works in three subsequent stages. First the clusters are defined based on semantic metadata method, followed by redefining the clusters based on the lexical method. Then, the modified hybrid method is used to combine the results of the semantic method and the lexical. Finally, the Power Link metric will be used to redistribute the hashtags based on the power link relationship of words in all blogs.

To illustrate how the modified hybrid clustering algorithm and the power link works, we will use a dummy example with artificial values.

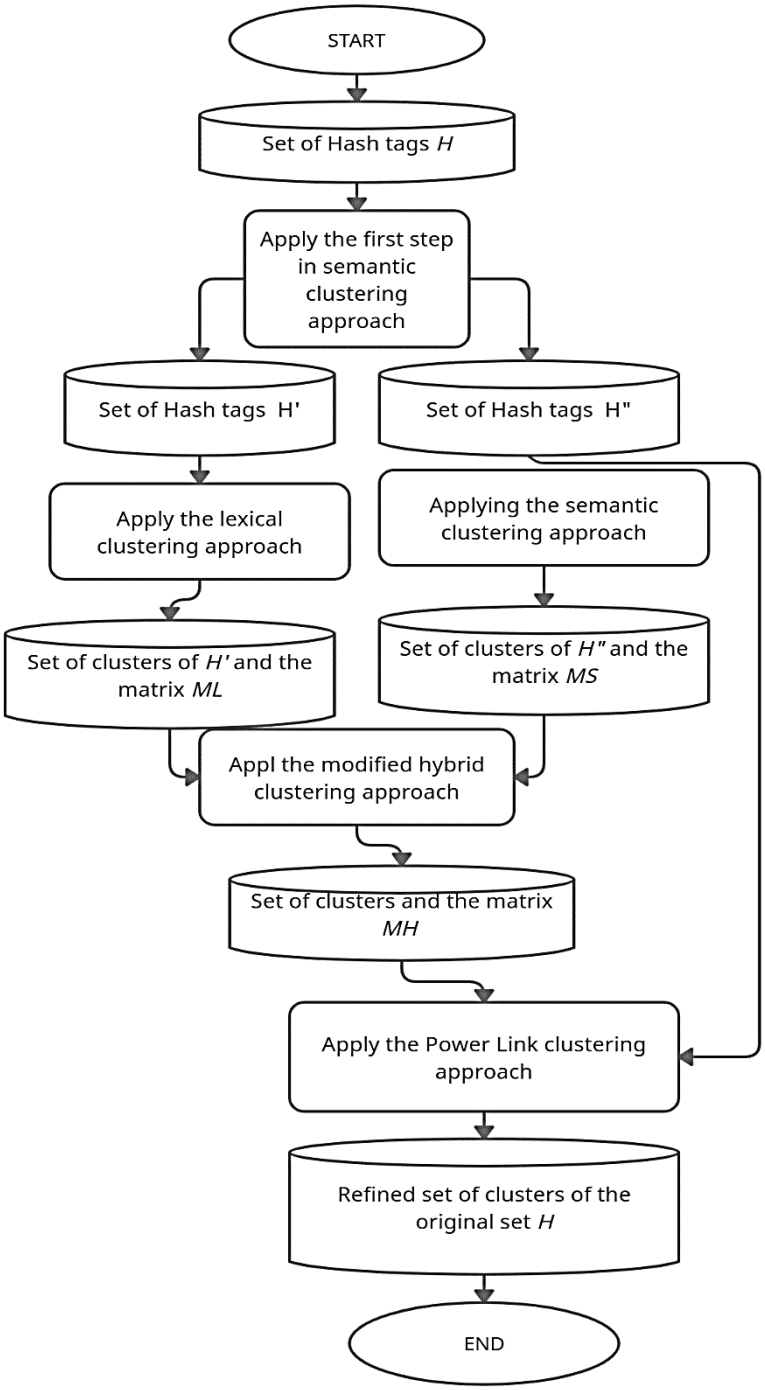

Fig. 5. Steps of full clustering method

Suppose that we have 8 hashtags #h1, #h2, ..., #h8. Based on the Semantic metadata clustering algorithm $\#h_7$ and $\#h_8$ has no senses and therefore will not be considered for further calculations in semantic metadata clustering or lexical clustering. Suppose that the similarity matrix MH of these hashtags is given by table I.

This matrix can be explained using a non-directed graph as shown in Fig. 6. Each vertex represents a hashtag, and each edge represents the similarity between the hashtags connected using the edge.

We used a distance measure between any two hashtags to be 1-similarity. In other words, if we have two hashtags h_1, h_2 and the similarity between h_1, h_2 is $MH(h_1, h_2)$ then the distance between h_1, h_2 is $D(h_1, h_2) = 1 - MH(h_1, h_2)$.

TABLE I
DUMMY EXAMPLE FOR THE HYBRID SIMILARITY MH MATRIX

	$\#h_1$	$\#h_2$	$\#h_3$	$\#h_4$	$\#h_5$	$\#h_6$
$\#h_1$	XXX	0.875	0.745	0.234	0.112	0.456
$\#h_2$	0.875	XXX	0.744	0.122	0.145	0.455
$\#h_3$	0.745	0.744	XXX	0.00	0.00	0.313
$\#h_4$	0.234	0.122	0.00	XXX	0.677	0.888
$\#h_5$	0.112	0.145	0.00	0.677	XXX	0.698
$\#h_6$	0.456	0.455	0.313	0.888	0.698	XXX

Using this measure, we can get the flat clusters for a given distance threshold dth. For example, if $dth \in (0.17,0)$, it is clearly that the clusters will be $\{\#h_1, \#h_2\}$, $\{\#h_3\}$, $\{\#h_4, \#h_6\}$ and $\{\#h_5\}$. However, if we considered more relaxed distance for example $dth \in (0.4,0)$ then we get two clusters, namely, $C_1 = \{\#h_1, \#h_2, \#h_3\}$ and $C_2 = \{\#h_4, \#h_5, \#h_6\}$. Applying the Power Link clustering algorithm can refine the clusters distribution and can map the hashtags that are removed by semantic metadata clustering method to their proper cluster. Here, considering the relaxed distance $dth \in (0.4,0)$ the Power link will modify the clusters to be $\{\#h_1, \#h_2, \#h_3, \#h_5, \#h_7\}$ and $\{\#h_4, , \#h_6, \#h_8\}$. The Power Links moved the hashtag $\#h_5$ from C_1 to C_2 and also mapped $\#h_7$ to C_1, and $\#h_8$ to C_2. This example illustrates the effect of Power Links in distributing the deleted hashtags H'' to their proper clusters and in redistributing the clustered hashtags in the set H', hopefully to get a better clustering.

III. Experiments

The target of the experiments is to compare the performance of the modified hybrid algorithm enhanced by the Power Links to the base (semantic metadata and lexical) clustering algorithms as well as the hybrid clustering algorithm proposed by [41]. Two types of tests will be used: "gold standard test" and "pairwise disagreement test". Each test will be used with different sets of experiments. The goal of these experiments is to test the performance of the proposed modified hybrid algorithm in different situations that affect the performance of the base algorithms. In the gold standard test, a set of clusters will be used to represent the truth set.

C. Data sets

To reflect the differences between the two base algorithms, two data sets are collected. The first data set is collected manually from specific domains (5 domains: Education, Policy, Medicine, Economy and Elections). These domains are clear, and the language used in these domains tends to be formal and respects the syntax and grammar rules. The second data set is collected randomly from Twitter API during the period from February 2018 till February 2019. Table II shows the number of tweets for each data set and the corresponding number of hashtags. It is expected that the lexical clustering algorithm will give better results in the first data set since the used terms tends to give the real relation between the hashtag and the written words. On the other hand, the random data set was chosen based on unbaised scale and this enabled covering many topics more than the first data set. However, the random data set contains many more unformal tweets and the lexical relation between the hashtag and the terms contained in the tweets seems to be strong enough to reflect a correct relation between the hashtags and the tweets that contain the hashtags. The above discussion shows the contrast between the two data sets and gives a good base to test the modified hybrid clustering algorithm.

TABLE II
THE NUMBER OF TWEETS FOR EACH DATA SET AND THE CORRESPONDING NUMBER OF HASHTAGS

Data Set	Number of tweets	Number of hashtags
Manual Data Set (MDS)	9873	1465
Random Data Set (RDS)	68521341	214569

D. Parameters

We need to set two types of parameters. The distance measure for hierarchy clustering and the threshold to choose the flat clusters. The approach here is to use a gradually incremented threshold that begins with 0.2 and gradually incremented by 0.05 in each iteration. The value that gives the best results will be considered. F-measure is used to evaluate the performance of all algorithms results.

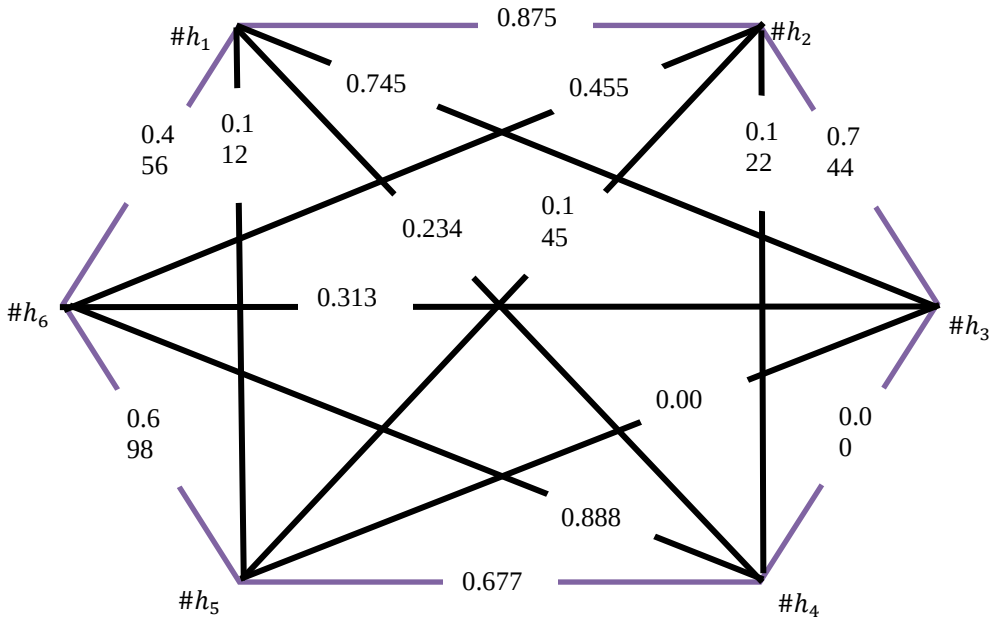

Fig. 6. Non-directed graph of the hybrid similarity $\boldsymbol{MH}$ matrix of the dummy example

E. Platform

Python language is used to implement the algorithms. All programs were run on a computer with windows 10 operating system, Intel® Core™ i7-7700HQ CPU@ 2.8 GHZ and 16 GB DDR memory.

F. Experiments for gold standard test

To form the ground truth sets, we followed the following steps as [41]

Two groups of ground truth are formed, the random ground truth sets (RGT), and the manual ground truth sets (MRT). To build the ground truth sets RGT, we apply the following steps:

- From the random set about 3.0 million tweets were randomly chosen. The number of hashtags that each of them appeared in more than 20 tweets is 1245.
- Approximately 50 hashtags are chosen based on their lexical properties to form the ground truth cluster RGT1. This process is repeated 3 times to choose the ground clusters RGT2, RGT3 and RGT4. The theme of each ground truth cluster is independently generated. Around 200 hundred hashtags clustered into 4 random ground truth clusters.
- From the manual tweet set, depending on the semantic properties, 300 hashtags are extracted from 1465 hashtags and classified into 20 clusters

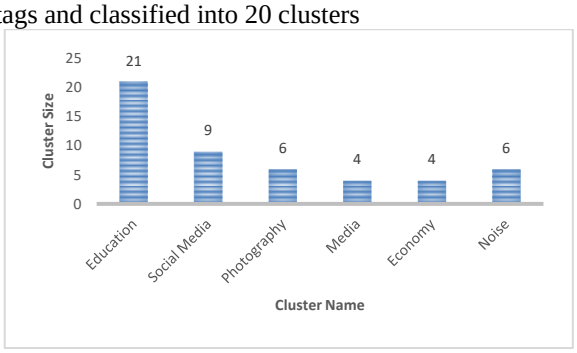

Fig. 7(a). Distribution of hashtags over different domains in the first random ground truth set (RGT1)

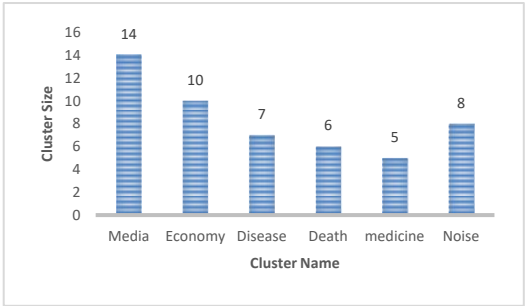

Fig. 7(b). Distribution of hashtags over different domains in the second random ground truth set (RGT2)

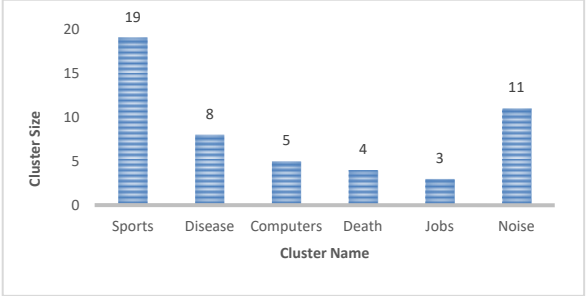

Fig. 7(c). Distribution of hashtags over different domains in the third random ground truth set (RGT3)

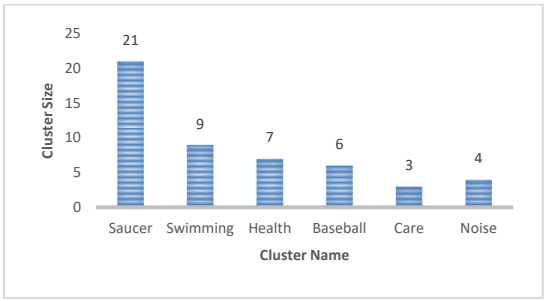

Fig. 7(d). Distribution of hashtags over different domains in the fourth random ground truth set (RGT4)

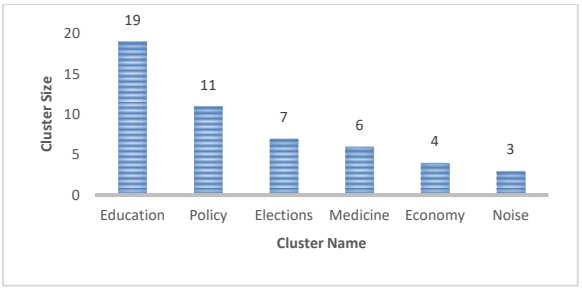

Fig. 7(e). Distribution of hashtags over different domains in the first manual ground truth set (MGT1)

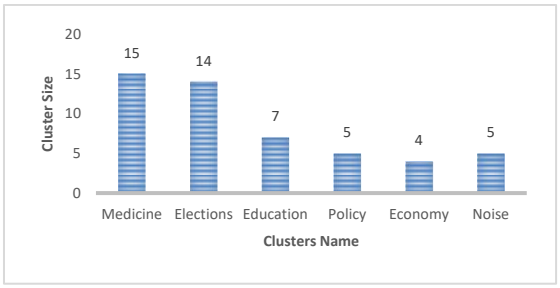

Fig. 7(f). Distribution of hashtags over different domains in the second manual ground truth set (MGT2)

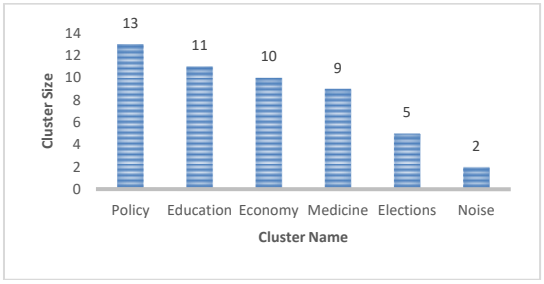

Fig. 7(g). Distribution of hashtags over different domains in the third manual ground truth set (MGT3)

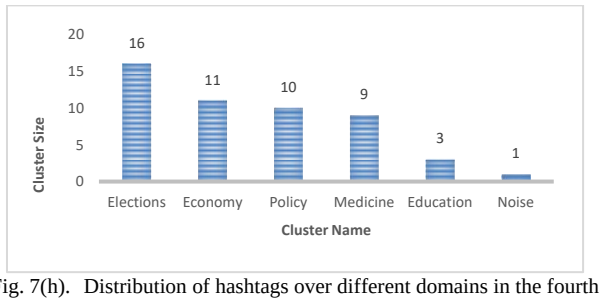

Fig. 7(h). Distribution of hashtags over different domains in the fourth manual ground truth set (MGT4)

Fig. 7. Distribution of the size of clusters in Random and Manual ground trust clusters

.

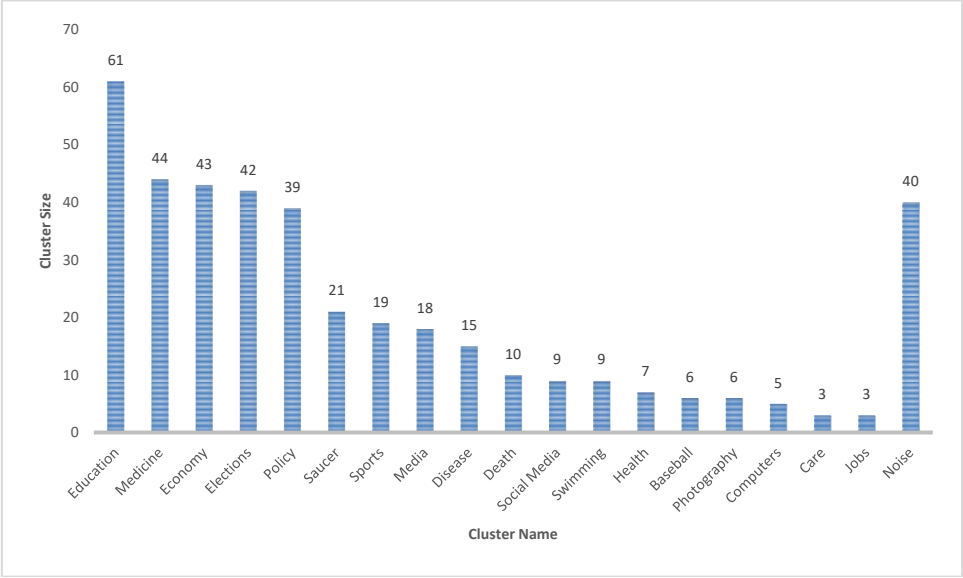

Fig. 8. Combined distribution of all hashtags in all manual and random ground truth sets (MRGT)

TABLE III.
PROFILES OF THE GROUND TRUTH CLUSTER SETS

	No of Hashtags	Source	Min no. of tweets
RGT1	50	Random	20
RGT2	50	Random	20
RGT3	50	Random	20
RGT4	50	Random	20
MGT1	50	Manual	20
MGT2	50	Manual	20
MGT3	50	Manual	20
MGT4	50	Manual	20
MRGT	400	Random ∪ Manual	Total of Random and Manual min tweets

- From these 20 clusters, one cluster was randomly chosen then another one was chosen and merged with the first one. The loop of choosing a cluster then merging it with the previous picked clusters is repeated until the total number of hashtags in the resulting set become around 50. The resulting set was given the name MGR1. The pervious step was repeated to form the trust ground clusters MG2, MG3 and MG4
- The themes of all ground trust sets were produced independently based on the work of 4 different people.
- All random ground clusters and manual ground clusters were merged to produce an additional ground cluster set MRGT.

Fig. 7, Fig. 8 and Table III summarizes the distributions of clusters and the themes of each ground trust clusters including the combined manual and random ground trust MRGT cluster.

E. Experiments for pairwise disagreement test

To test a huge number of hashtags (Large scale evaluation) it will be infeasible to use ground trust sets. The infeasibility is due to the burden of the required manual work. To overcome the overhead of the manual work pairwise disagreement test is used. We will use the disagreement to compare the performance of the modified hybrid clustering algorithm supported by the Power Link in contrast with the original hybrid clustering algorithm and the modified hybrid clustering algorithm. The pairwise disagreement test gives the ability for an algorithm to guess the right decision in contrast of the two other algorithms. The aim here is to evaluate the effect of Power Links on the modified clustering algorithm, and we need to check to what extent it will affect the performance of the modified algorithm in contrast with the performance of the hybrid algorithm.

The data in Table IV and Table V show that many clusters contain very few numbers of hashtags, in many cases one or two. This means that we have a small granularity level, so the chance that two different clusters will be allocated in the same cluster is very low. In other words, the chance that any two clusters will be clustered into two different clusters is very high. So, it seems that the probability that all algorithms will classify any two hashtags to different clusters is high.

Based on this fact, we will check the disagreement between different algorithms to classify hashtags to different clusters. In fact, we have two cases of disagreement:

1- Each of the hybrid and modified hybrid algorithms classify two hashtags to one cluster, while the modified algorithms supported by the Power link classifies the two hashtags to different clusters.

TABLE IV.
CLUSTERS PROFILE FOR CONTROLLED DATA SET

	Number of clusters per size range					
	>=100	50-99	20-49	10-19	5-9	1-4
Lexical	0	0	67	111	73	1325
Semantic Metadata	0	16	81	74	92	1942
Hybrid	3	6	73	238	201	1658
Modified Hybrid	4	7	23	251	78	1458
Modified Hybrid and Power Links	1	5	51	189	194	1788

TABLE V.
CLUSTERS PROFILE FOR UNCONTROLLED DATA SET

	Number of clusters per size range					
	>=100	50-99	20-49	10-19	5-9	1-4
Lexical	1	14	20	99	250	1451
Semantic Metadata	3	10	69	89	149	1125
Hybrid	0	6	48	125	211	1891
Modified Hybrid	4	3	66	109	189	1521
Modified Hybrid and Power Links	0	5	48	143	56	1753

2- Each of the hybrid and modified hybrid algorithms classify the two hashtags to different clusters, while the modified algorithms supported by the Power link classifies the two hashtags to one cluster.

The word "two clusters are in one cluster" means that there is at least one cluster that contains the two hashtags. The word "two clusters are in two different clusters" means that there is no cluster that contains the two hashtags at the same time. This reflects the fact that the clusters are overlapping.

The experiments are set as follows:

Having two types of data sets; the controlled data set and the uncontrolled data sets. A controlled data set was formed by using all hashtags in the manual data set. The uncontrolled data set was formed by extracting 2000 hashtags randomly based on the timestamp ~~randomly~~. To evaluate the performance of the modified hybrid algorithm supported by the Power Links, 100 pairs of hashtags are extracted manually from the controlled and from the uncontrolled sets in each case. These pairs are evaluated manually based on the tweets meaning to check whether each pair should be in one cluster or to be in two different clusters. The parameters in this group of experiments are set to be the same as the best values that we got during the golden standard experiments.

IV. RESULTS

For evaluation purpose, F measure is used. Maximum score F(FMAX) and average FAVG of FMAX for all ground trust sets is used. For a given set of clusters C, the FMAX of a member G_i in the set of ground trust sets G is given by:

$$FMAX(G_i, C) = \max_{c_j \in C} F(c_j, G_i)$$

The average FAVG of the F measure accuracy between the trust set G and an output clusters C is given by:

$$FAVG(G, C) = \frac{\sum_{G_i \in G} FMAX(G_i, C) \times |G_i|}{\sum_{G_i \in G} |G_i|}$$

Fig. 9 and Fig. 10 show the weighted average FAVG and the results of pairwise FMAX for the ground trust sets corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid, and modified hybrid enhanced with Power Links clustering methods.

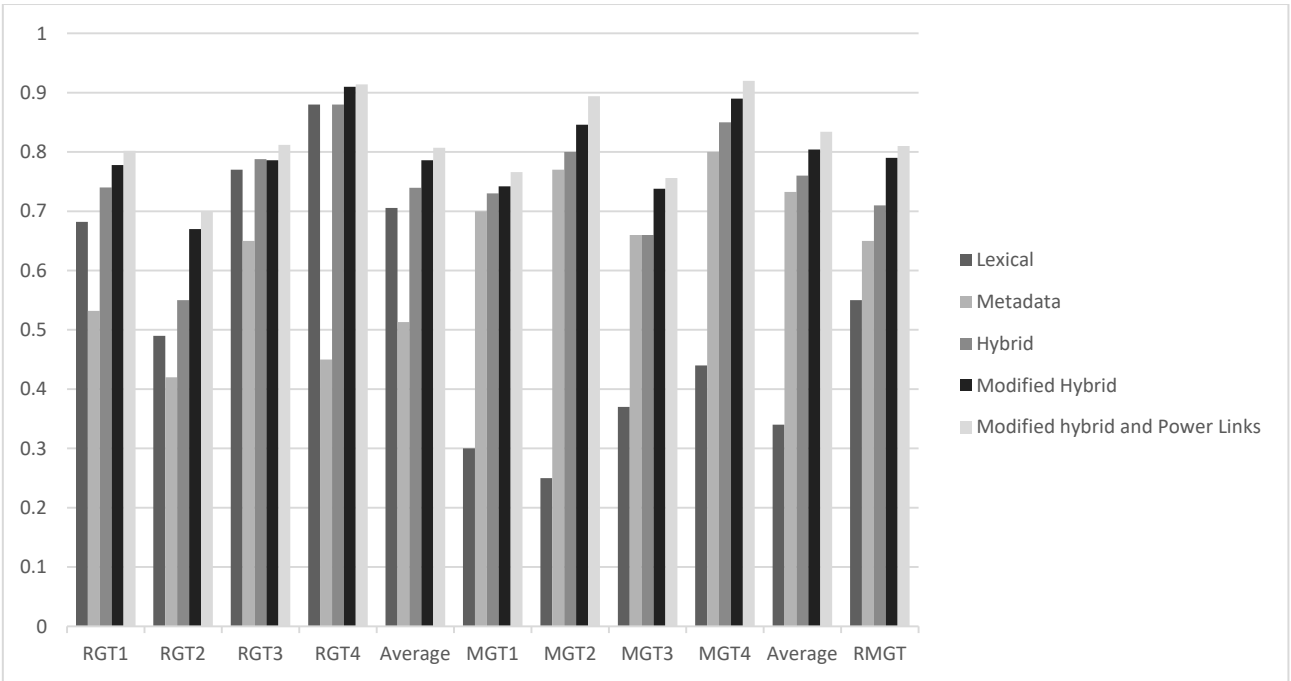

Fig. 9. The weighted average FAVG of F measure for the ground trust sets corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid and modified hybrid enhanced with Power Links clustering methods.

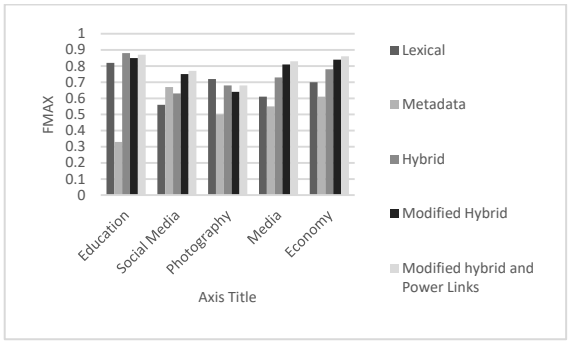

Fig. 10(a). The results of pairwise Maximum score F (FMAX) for the ground trust set (RGT1) corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid and modified hybrid enhanced with Power Links clustering methods

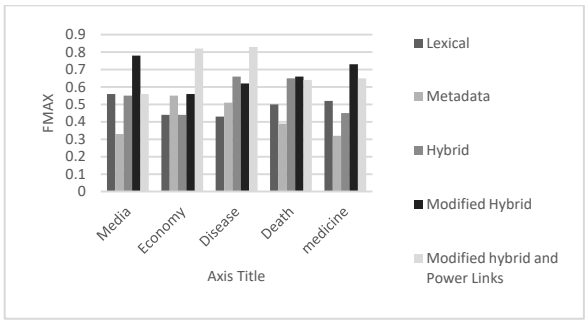

Fig. 10(b). The results of pairwise Maximum score F (FMAX) for the ground trust set (RGT2) corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid and modified hybrid enhanced with Power Links clustering methods.

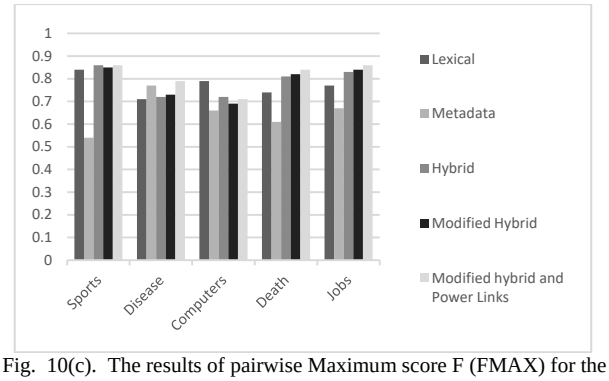

Fig. 10(c). The results of pairwise Maximum score F (FMAX) for the ground trust set (RGT3) corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid and modified hybrid enhanced with Power Links clustering methods.

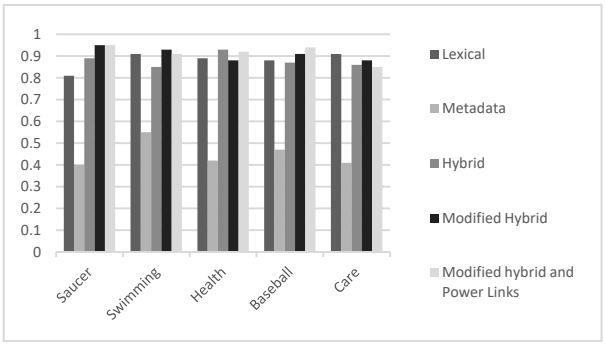

Fig. 10(d). The results of pairwise Maximum score F (FMAX) for the ground trust set (RGT4) corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid and modified hybrid enhanced with Power Links clustering methods.

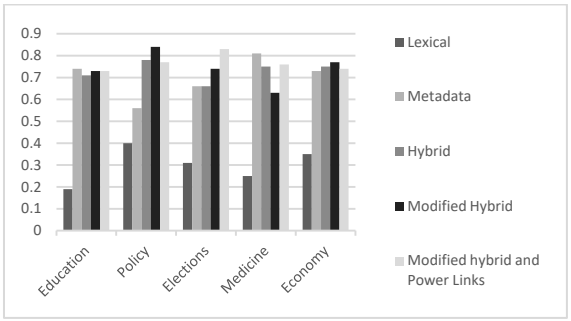

Fig. 10(e). The results of pairwise Maximum score F (FMAX) for the ground trust set (MGT1) corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid and modified hybrid enhanced with Power Links clustering methods.

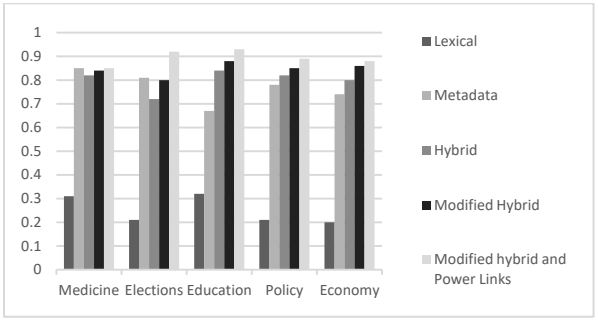

Fig. 10(f). The results of pairwise Maximum score F (FMAX) for the ground trust set (MGT2) corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid and modified hybrid enhanced with Power Links clustering methods.

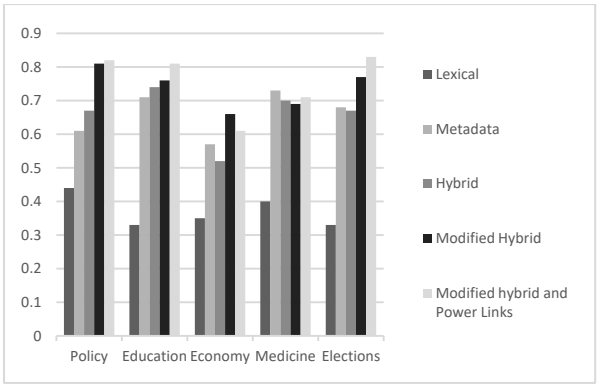

Fig. 10(g). The results of pairwise Maximum score F (FMAX) for the ground trust set (MGT3) corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid and modified hybrid enhanced with Power Links clustering methods.

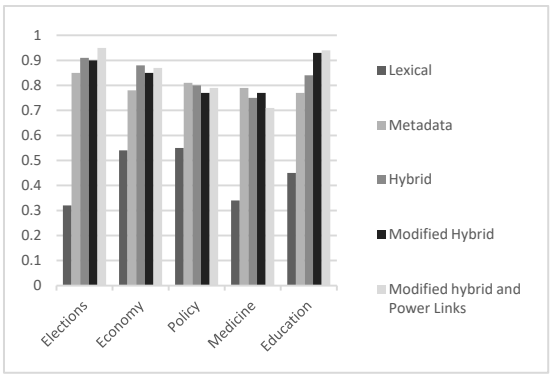

Fig. 10(h). The results of pairwise Maximum score F (FMAX) for the ground trust set (MGT4) corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid and modified hybrid enhanced with Power Links clustering methods.

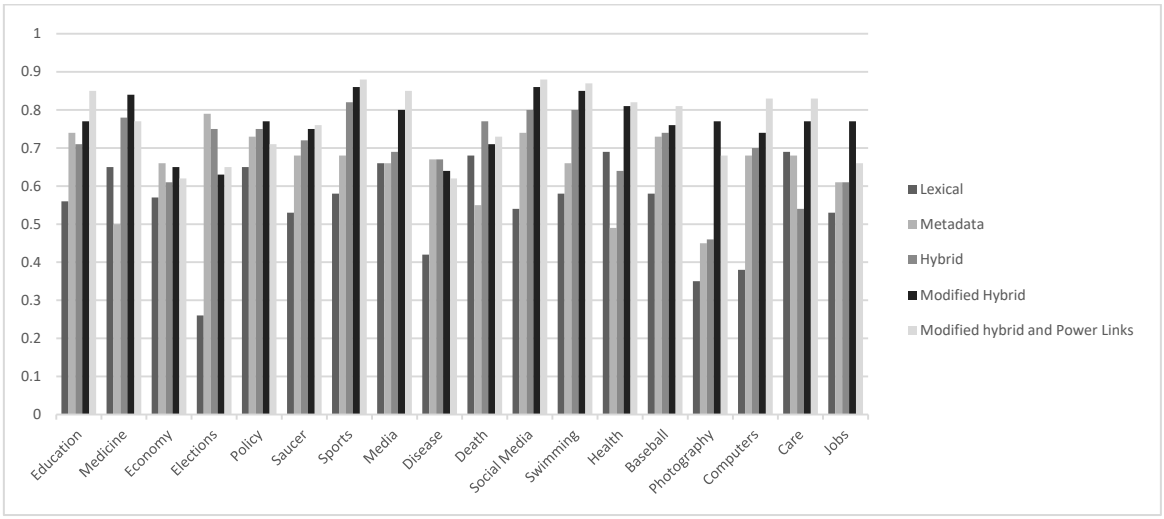

Fig. 10 (i). The results of pairwise Maximum score F (FMAX) for the combined ground trust set (MRGT) corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid and modified hybrid enhanced with Power Links clustering methods.

Fig. 10. The results of pairwise Maximum score F (FMAX) for the ground trust sets corresponding to the two base algorithms (Lexical and metadata semantic), hybrid, modified hybrid and modified hybrid enhanced with Power Links clustering methods.

The results show that the hybrid clustering method performed better than the two bases clustering algorithms in most of the above results.

Also, the modified hybrid clustering method outperformed the hybrid method in most cases and the modified clustering enhanced with power link clustering method succeeded to outperform all other clustering methods in most cases. From Fig. 9, FAVG values show that, in average, the best algorithm is the modified hybrid algorithm enhanced with the Power Links which produces the best results.

If we looked in detail for the random ground trust sets RGTs in Fig. 10, the hybrid algorithm outperformed the base algorithms in 9 cases out of twenty cases. Even in the cases when one of the base algorithms outperform the hybrid algorithm, the performance of the hybrid algorithm is equal to or differs by a minor margin from the higher algorithm. On the other hand, the modified hybrid algorithm outperformed the hybrid algorithm in 13 out of 20 cases. Also, supporting the modified algorithm with the Power Links helped to enable the modified hybrid algorithm to outperform the hybrid algorithm in 15 out of 20 cases.

For the manual ground trust sets MGTs, in Fig. 10, the hybrid algorithm outperformed the base algorithms in 11 out of 20 cases. Even in the other 9 cases, the hybrid algorithm performance was very near the higher base algorithm. On the other hand, the modified hybrid algorithm outperformed the hybrid algorithm in 15 out of 20 cases and the Power Links helped the modified hybrid algorithm to outperform the hybrid algorithm in 16 out of 20 cases.

For the combined manual and random ground trust set, the hybrid algorithm outperformed the base algorithms in 12 out of 18 cases. Also, the modified hybrid algorithm outperformed the hybrid algorithm in 13 out of 18 cases and the Power Link helped the modified algorithm to outperform the hybrid algorithm in 10 out of 18 cases. So, in most cases, the modified hybrid algorithm outperformed the hybrid algorithm or gave a very near performance from the hybrid algorithm in most cases.

The lexical algorithm outperformed the metadata semantic algorithm in most cases (16 out of 20) in the random ground trust sets RGTs. On the other hand, the metadata semantic algorithm outperformed the lexical algorithm in all cases for the manual ground trust sets MGTs. For the combined trust set MRGT, the lexical algorithm outperformed the semantic algorithm in 4 out of 18 cases which means that the semantic metadata algorithm outperformed the lexical algorithm in most of the cases. [41] explained why the lexical algorithm outperformed the semantic metadata algorithm in the random trust sets to the lexical structure of the tweets contains a given hashtag. Since the hashtag to be considered must appear in more than 20 tweets in the random sets and relaxing this condition for the manual sets. In our experiments, we restricted the condition in both manual and random sets; however, we still found that the lexical outperforms the semantic metadata in the random sets and conversely the semantic metadata outperforms the lexical method in the manual sets and the combined sets. This clearly due to the formal language used in the manual sets, since they are written by specialist who write specific and related tweets to the topic of the hashtags. On the other hand, the random sets contain many informal languages, and in many cases, there is no clear relation between the written tweet and the expected topic of the hashtag. In addition, there is great freedom in using the words without considering any syntax or grammar rules.

TABLE VI.
PAIRWISE DISAGREEMENT TEST RESULTS FROM THE CONTROLLED DATASET.

Method	Case	No. of instances	No. of Modified Hybrid and Power Links Correct
Hybrid algorithm	Case1	16124	95/100=95%
	Case2	2314	96/100=96%
Modified Hybrid Algorithm	Case1	11156	90/100=90%
	Case2	1256	91/100=91%

Table VI and Table VII represent the results of the gold. These results show that the modified hybrid algorithm supported by the Power Links was able to give the correct decision in average 93 % for the uncontrolled set and 97 % in the case for the controlled data set. This explains that the proposed algorithm gave promising results.

TABLE VII.
PAIRWISE DISAGREEMENT TEST RESULTS FROM THE UNCONTROLLED DATASET.

Method	Case	No. of instances	No. of Modified Hybrid and Power Links Correct
Hybrid algorithm	Case1	819	100/100=100%
	Case2	256	99/100=99%
Modified Hybrid Algorithm	Case1	55624	94/100=94%
	Case2	953	95/100=94%

To give a near insight of the effect of using the hybrid modified clustering algorithm, we choose one important example. In this example, we will show that the modified hybrid algorithm supported with the power link enhancement makes the right decision. In this example the algorithm will beat the Lexical clustering and the semantic meta data clustering algorithm.

Example: This example, explores the hashtag #digitalhealth. This hashtag consists of two terms "digital" and "health". The hashtag #digitalhealth might belong to different domains, computer or health. Based on the three algorithms, the resulting group that contains this hashtag for each algorithm are as follows.

- Lexical algorithm list: {#digitalhealth, #BCSM, #MedEd, #HCLDR, #HITsm, #hpm, #LCSM, #BTSM, #hcsmSA, #JACR, #NephJC, #ASEchoJC, #SPSM, #NurChat, #KareoChat, #EofL, #Path2Path, #LaMPSCymru, #HealthyMindChat, #PubertyEndoChat, #JJDiabetesInst, #HIC18, #abimf2018, #ANMS18, #WE2012, #MOGA18, #AFibPatient18, #AAFPNC, #AAPM2018, #AANAM, #PHASA2018, #OchreDay2018, #BCVS18, #JergeDental, #47MillerMed, #ASH17, #OW2018, #MonitoreoHem2018, #OCEIONonClinModels, #MEWomenshealth, #EAHM2018, #BCSM, #Lyphoma, #BTSM, #pwme, #ChildhoodCancer, #hidradenitissuppurativa, #ProstateCancer, #OVCA, #AFib, #Migraine, #SpinalCordInjury, #ColonCancer, #Diabetes, #Parkinsons, #colds, #HipDysplasia, #iamabariatricpatient, #BoxonslaSEP, #digitalhealth, #vacunas, #hcsm, #NNM, #massagetherapy, #PlasticSurgery, #skincare, #Cytopath, #pharma, #SelfCare, #homecare, #KeepTalkingMH, #pharmacy, #doctors, #computerscience #software #programming #pcgaming }
- Semantic Metadata: {#digitalhealth, #homecare, #SelfCare, #computerscience, #software, #programming #pcgaming, #windows, #coding, #computers, #PlasticSurgery, #skincare, #Cytopath, #pharma, #NurChat, #KareoChat, #EofL, #Path2Path, #LaMPSCymru, #HealthyMindChat, #PubertyEndoChat, #JJDiabetesInst, #abimf2018, #JergeDental, #47MillerMed, #MonitoreoHem2018, #OCEIONonClinModels}
- Modified Hybrid Algorithm: {#digitalhealth, #computers, #SpinalCordInjury, #ColonCancer, #Diabetes, #Parkinsons, #colds, #HSDD, #HipDysplasia, #iamabariatricpatient, #BoxonslaSEP, #digitalhealth, #hcsmeu, #vacunas, #hcsm, #NNM, #massagetherapy, #PlasticSurgery, #skincare, #Cytopath, #pharma, #SelfCare, #homecare, }.

We note that the semantic metadata failed to distinguish between the domain of the computer and the domain of health and tends to collect hashtags from on the domain in one category based on their appearance. The lexical clustering algorithm succeeded in performing a better clustering since most of the hashtags come from the health domain but the set still contains many hashtags from the computer field. The modified clustering algorithm put the hashtag #digitalhealth in a set where all hashtags belong to the health domain except the hashtag #computers. The semantic algorithm based heavily on breaking the hashtag to its individual terms then search for the related hashtags for each term. Lexical approach depends on the content of the related blogs and since most of the blogs that contain the hashtag #digitalhealth come from the health domain and also contain some hashtags from the computer domain, the algorithm clustered the hashtag #digitalhealth in a set with a majority of hashtags from the health domain. The Power Link measure helped the hybrid algorithm to exclude many of the hashtags that belong to the computer domain and added other hashtags from the health domain based on the co-occurrence properties. This explains why the modified hybrid algorithm can outperform the primary algorithms.

V. CONCLUSION

In this paper, we explored the problem of clustering hashtags. Based on two traditional lexical and metadata semantic clustering algorithms, a hybrid algorithm combines the advantages of the two base algorithms to produce a better clustering. Our work focused on improving the hybrid algorithm by considering the real values of similarity provided by the two base algorithms to build a new similarity matrix. The proposed modified hybrid algorithm overcomes the drawbacks of the hybrid algorithm as well as the two base algorithms in the average evaluation and in many cases with respect to the random data sets and the manually extracted data sets. The modified hybrid algorithm was enhanced by the Power Links. Power Links made use of co-occurrence to define a link between hashtags and these links were used to define a metric measure. This metric measure was used to refine the clusters produced by the modified hybrid algorithm. The results show that the Power Links metric was able to improve the performance of the modified hybrid algorithm by a significant value. Future work will concentrate on enhancing the base algorithms with a new source for semantic metadata clustering like google new tools. Also, a specialized languages processor to detect the language of related tweets related to the hashtags and choose a specific pre-processing manner suitable for different languages to extend the approach for different languages from English and Multilanguage tweets. Moreover, the method can imbed into applications to enhance the power of these applications by knowing the clusters and related topics to different tweets.

REFERENCES

[1] Jain, M. Murty and P. Flynn. Data Clustering: A Review. ACM Computing Surveys, vol. 31, no. 3, pp. 264-323,1999.
[2] Jain, R. Duin and J. Mao. Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no.1, pp. 4-37, 2000.

[3] Y. Leung, J. Zhang and Z. Xu. Clustering by Space-Space Filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no.12, pp. 1396-1410, 2000.

[4] G. Hamerly and C. Elkan. Alternatives to the K-means Algorithm that Find Better Clusterings. In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM-2002), pp. 600-607, 2002.

[5] E. Forgy. Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classification. Biometrics, vol. 21, pp. 768-769, 1965.

[6] J. Bezdek. A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 2, pp. 1-8, 1980.

[7] J. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, 1981.

[8] G. McLachlan and T. Krishnan. The EM algorithm and Extensions. John Wiley & Sons, Inc., 1997.

[9] R. Rendner and H. Walker. Mixture Densities, Maximum Likelihood and the EM Algorithm. SIAM Review, vol. 26, no. 2, 1984.

[10] C. Bishop. Neural Networks for Pattern Recognition. Clarendon Press, Oxford, 1995.

[11] Y. Leung, J. Zhang and Z. Xu. Clustering by Space-Space Filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no.12, pp. 1396-1410, 2000.

[12] Mahmoud Rokaya and El-Sayed Atlam. (2010). Building of field association terms based on links, Int. J. Computer Applications in Technology, vol. 38, no. 4, 2010.

[13] Rokaya M., and Nahla A., Building a Multi-lingual Field Association Terms Dictionary, International Journal of Computer Science and Network Security,11-3 (2011), 208-213.

[14] Rokaya M., Nahla A. and Aljahdali S., Context-Sensitive Spell Checking Based on Field Association Terms, IJCSNS International Journal of Computer Science and Network Security. 12- 3(2012), 64-68.

[15] Mahmoud B. Rokaya. (2013). Automatic text extraction based on field association terms and power links," International Journal of Computer and Information Technology (ISSN: 2279 – 0764) vol. 02– no 06, November 2013.

[16] Rokaya M. and Aljahdali S. (2013). Building a Real Word Spell Checker Based on Power Links, International Journal of Computer Applications, 65-7 (2013), 14-19.

[17] Rokaya, M.B. (2013) Automatic text extraction based on field association terms and power links.

[18] (2013) International Journal of Computer and Information Technology, 2 (6).

[19] Mahmoud Rokaya. (2014). Improving Ranking of Search Engines Results Based on Power Links IPASJ International Journal of Information Technology (IIJIT), vol 2, no 9, September 2014.

[20] Mahmoud Rokaya. (2015). Arabic Semantic Spell Checking Based on Power Links, International Information Institute, vol.18, no.11,2015.

[21] Mahmoud Rokaya, Ahmed S. Ghiduk. (2017). Developing Arabic ontology based on power links. International Information Institute (Tokyo). Information 20 (10A), 7429-7444, 2017 .

[22] Mahmoud Rokaya, Dalia I. Hemdan. (2016). "Bibliometric cartography of nutrition science researches based on Power Links analysis," International Information Institute, vol.19, no.9(B), 2016.

[23] Mahmoud Rokaya. (2016). "Spam Reduction Based on Power Link Analysis," International Information Institute, 19(6A), 1921–1932

[24] Atlam, E.-S., Abo-Shady, D., Mohamed, D. A. A., & Ghaleb, F. (2018). An improvement of FA terms dictionary using Power Link and co-word analysis. International Journal of Advanced Computer Science and Applications, 9(2), 236–241. https://doi-org.sdl.idm.oclc.org/10.14569/IJACSA.2018.090233.

[25] H. Saif , Y. He , H. Alani ,Semantic sentiment analysis of Twitter, Proceedings of the 11th International Conference on The Semantic Web - Volume Part I, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 508–524 .

[26] C.I. Muntean , G.A. Morar , D. Moldovan , Exploring the meaning behind Twitter hashtags through clustering, Lect. Notes Bus. Inf. Process. 127 (2012) 231–242 .

[27] S. Park , H. Shin , Identification of implicit topics in Twitter data not containing explicit search queries, Proceedings of the 25th International Conference on Computational Linguistics, 2014, pp. 58–68 .

[28] S. Bhulai , P. Kampstra , L. Kooiman , G. Koole , M. Deurloo , B.K. CCing , Trend visualization on Twitter: what's hot and what's not?, Proceedings of the 1st International Conference on Data Analytics, Springer-Verlag, 2012, pp. 43–48 .

[29] J. Costa , C. Silva , M. Antunes , B. Ribeiro , Defining semantic meta-hashtags for Twitter classification, Lect. Notes Comput. Sci. 7824 (2013) 226–235 .

[30] O. Tsur , A. Littman , A. Rappoport , Scalable multi stage clustering of tagged mi- cro-messages, Proceedings of the 21st International Conference on World Wide Web, 2012, pp. 621–622 .

[31] O. Tsur , A. Littman , A. Rappoport , Efficient clustering of short messages into general domains, in: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, 2013, pp. 621–630 .

[32] K.D. Rosa , R. Shah , B. Lin , Topical clustering of tweets, Proceedings of the 3rd Workshop on Social Web Search and Mining, 2011, pp. 133–138 .

[33] W. Feng , C. Zhang , W. Zhang , J. Han , J. Wang , C. Aggarwal , J. Huang , STREAM- CUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream, Proceedings of the IEEE 31st International Conference on Data Engineering, 2015, pp. 1561–1572 .

[34] G. Stilo , P. Velardi , Temporal semantics: time-varying hashtag sense clustering, Proceedings of the 19th International Conference on Knowledge Engineering and Knowledge Management, Springer, 2014, pp. 563–578

[35] G. Stilo , P. Velardi , Hashtag sense clustering based on temporal similarity, Com- put. Linguist. 43 (1) (2017) 181–200 .

[36] X. Wang , F. Wei , X. Liu , M. Zhou , M. Zghan , Topic sentiment analysis in Twit- ter: a graph-based hashtag sentiment classification approach, Proceedings of the 20th ACM Conference on Information and Knowledge Management, 2011, pp. 1031–1040 .

[37] P. Teufl, S. Kraxberger , Extracting semantic knowledge from Twitter, Proceed- ings of the 3rd IFIP WG 8.5 International Conference on Electronic Participa- tion, Springer-Verlag, 2011, pp. 48–59 .

[38] I.S. Dhillon , Y. Guan , J. Fan , Efficient Clustering of Very Large Document Collec- tions, Kluwer Academic Publishers, 2001 .

[39] W. Feng , C. Zhang , W. Zhang , J. Han , J. Wang , C. Aggarwal , J. Huang , STREAM- CUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream, Proceedings of the IEEE 31st International Conference on Data Engineering, 2015, pp. 1561–1572

[40] Merriam Webster, Simple definition of hashtag. [cited 2016-05-01]. URL http: //www.merriam-webster.com/dictionary/hashtag .

[41] Javed, A., & Lee, B. S. (2018) 'Hybrid semantic clustering of hashtags', Online Social Networks and Media, 5, 23-36.

[42] A. Javed , B.S. Lee (2017) 'Sense-level semantic clustering of hashtags', Communications in Computer and Information Science, 1–16 .

[43] C. Vicient , A. Moreno. (2014) 'Unsupervised semantic clustering of Twitter hashtags', Proceedings of the 21st European Conference on Artificial Intelligence, pp. 1119–1120 .

[44] Srivastava, A., Singh, V., & Drall, G. S. (2019) 'Sentiment Analysis of Twitter Data: A Hybrid Approach', International Journal of Healthcare Information Systems and Informatics (IJHISI), 14(2), 1-16.

[45] Mika, P. (2004) 'Social networks and the semantic web', IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), IEEE.

[46] Aswani, R., Kar, A. K. and Ilavarasan, P. V. (2018) 'Detection of Spammers in Twitter marketing: A Hybrid Approach Using Social Media Analytics and Bio Inspired Computing', Information Systems Frontiers, 20(3), pp. 515–530

[47] Ramzan, N., van Zwol, R., Lee, J. S., Clüver, K., & Hua, X. S. (Eds.). (2012) 'Social media retrieval', Springer Science & Business Media.

[48] Tao Ma, Heng Liu, and Yu Zhang, "A Method for Establishing Tropospheric Atmospheric Refractivity Profile Model Based on Multiquadric RBF and k-means Clustering," Engineering Letters, vol. 28, no.3, pp733-741, 2020

[49] Md Kamrul Islam, Md Manjur Ahmed, and Kamal Zuhairi Zamli, "i-CODAS: An Improved Online Data Stream Clustering in Arbitrary Shaped Clusters," Engineering Letters, vol. 27, no.4, pp752-762, 2019

[50] Lu Li, Yun Lin, Xu Wang, Tian Guo, Jie Zhang, Hua Lin, and Fuqian Nan, "A Clustering-Classification Two-Phase Model on Service Module Partition Oriented to Customer Satisfaction," Engineering Letters, vol. 26, no.1, pp76-83, 2018

[51] Jiemin Huang, Jiaoju Ge, and Yixiang Tian, "Analysis of Corporate Bond Yield Spread Based on the Volatility Clustering Effect," IAENG International Journal of Applied Mathematics, vol. 49, no.4, pp605-611, 2019

[52] A. Ragozin, V. Telezhkin, and P. Podkorytov, "Hierarchical Cluster-analysis of Transient Heart Rate using a Digital Spectral Analysis in

the Complex Frequency Plane," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2019, 22-24 October, 2019, San Francisco, USA, pp1-3
[53] Hirofumi Miyajima, Hiromi Miyajima, and Norio Shiratori, "Proposal of Fast and Secure Clustering Methods for IoT," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2019, 13-15 March, 2019, Hong Kong, pp1-6
[54] Maaz Ahmed, Mohsin Khan, Waseem Ahmed, Rashid Mehmood, Abdullah Algarni, Aiiad Albeshri, and Iyad Katib, "PSim: A Simulator for Estimation of Power Consumption in a Cluster," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2018, 23-25 October, 2018, San Francisco, USA, pp102-107

BIBLIOGRAPHY

MAHMOUD B. ROKAYA was born in Tanta city, Egypt in 1971. He received the B. S and M.S. degrees in mathematics from Tanta University, Egypt in 2003 and the Doctor of Engineering in information science From Tokushima University, Japan in 2009.

From 1997 to 2003, he was assistant of teaching in department of mathematics, Tanta University, Egypt. From 2003 to 2009, he was a researcher in the advanced engineering institute, Tokushima University, Japan. Since 2009-2020, he was assistant professor in information technology, Taif University, KSA. Currently, he is associate professor in informatics, Taif University. His research interests related to AI, information retrieval, natural language processing and data science.

Dr. Rokaya was a recipient of the outstanding in scientific research from Taif University for 5 subsequent years from 2010 to 2016. He also was the chair of the committee that got the ABET accreditation in the college of computers and information technology, Taif University, KSA from 2018 to 2025.

HAMZA TURABIEH received his B.A., M.Sc. degrees in Computer Science from Balqa Applied University in 2004 and 2006 respectively, in Jordan and Ph.D. from National University of Malaysia (UKM) in 2010.

From 2011 to 2014, he was Assistant professor in Computer Science Department- College of Computers and Information Technology, Taif University, KSA. Since 2014, he was Associate professor in Computer Science Department, College of computers and Information Technology, Taif University, KSA.

Dr. Turabieh research interests and activities lie at the interface of Computer Science and Operational Research. Intelligent decision support systems, search and optimization (combinatorial optimization, constraint optimization, multi-modal optimization, and multi-objective optimization) using heuristics, local search, hyper-heuristics, met heuristics (in particular memetic algorithms, particle swarm optimization), hybrid approaches and their theoretical foundations. Minor interest in machine learning, computational geometry, pattern recognition, image processing, intelligent user-interfaces, and Bioinformatics.

SANA AL AZWARI received her PhD degree in Computer and Information Sciences in June 2016 from the University of Strathclyde, UK. She also completed her master's degree in computer and Internet Technologies in 2010 from the University of Strathclyde, UK.

She is an assistant professor and the vice dean of the Computer and Information Technology College at Taif University. Her research focuses on minimizing the number of updates (deltas) necessary for updating Semantic Web data. This work leads to the introduction of a new updating method that exploits implicit information to produce sound and small in size deltas, which are both important properties for a sufficient delta.

Dr. AL Azwari won the best paper Award in SEMANTiCS15 in Vienna and the best paper Award in the 8th Saudi Conference in London. During her PhD studies she won a number of prizes for the Best Student Award from the Saudi Cultural Bureau in London. Before that, in 2006, she awarded the King Abdulaziz and His Companions Foundation for the Gifted Award. She is now an International Science Ambassador for the University of Strathclyde.

ABDULLAH ALHARBI was born in Saudi Arabia in 1885. He received his Ph.D. degree in Information Technology from the University of Technology Sydney, Australia.

He is currently an Assistant Professor with Information Technology Department, College of Computers and Information Technology, Taif University in Saudi Arabia . His research interests include big data analytics, data mining, the Internet of Things, web design, human computer interaction, information systems.

Dr. ALHARBI is currently the vice dean of training, College of computers and information technology, Taif University, KSA.

WAJDI ALHAKAMI received, the M.Sc. degree in computer network, and the Ph.D. degree in network security from the University of Bedfordshire, U.K in 2011 and 2016 respectively.

He is currently an Assistant Professor with the College of Computers and Information Technology, Taif University, KSA.

Dr. ALHAKAMI research interests include the Internet of Things, cyber security, and computer networking.

MRIM M. ALNFIAI was born in Taif, Makkah, SA in 1987. She received the B.S. degree in Information Technology from Taif University, Taif, in 2009. From 2012 to 2014, she was a master student in Computer Science at Dalhousie University, CANADA. From 2014 until 2018, she was a PhD student in Computer Science at Dalhousie University, CANADA.

She is an Assistant Professor of Information Technology at the Taif University in Saudi Arabia. Her research interests are in assistive technology, Human Computer Interaction and Accessibility. Mrim publishes several papers at assistive technology, HCI and accessibility conferences including ASSETS, ANT, FNC, CIST, JAIHC, and ICCA. Currently, her research focuses on designing accessible tools for visually impaired people including people with no or low vision. She has conducted several studies and experiences to understand visually impaired abilities and behaviors and design accessible systems that help them interact easily with technology. From 2018 until now, she is an Assistant Professor of Information Technology at the Taif University and she is working as the vice president of IT department in Taif university.

Dr. Alnfiai was a recipient of the the 2nd International Conference on Computer and Information Science and Technology (CIST'16) Best Paper Award in 2016 and she has received the Saudi Bureau Award in 2014.

WAEL ALOSAIMI was born in Saudi Arabia in 1979. He received the BSc in Computer Engineering from King AbdulAziz University in 2002. In 2011, He received the MSc in Computer Systems Security and the PhD in Cloud Security from the University of South Wales in November 2016.

From 2002 to 2004, he worked at Saline Water Conversion Corporation (SWCC) as an instrument and control engineer. Then, he served as a trainer for the Technical and Vocational Training Corporation (TVTC) until 2008. Next, he joined Taif University as a teaching assistant. It provides him with a scholarship to pursue his studies in the UK. Since 2017, He has been an assistant professor at computer engineering department in Taif University. He has many publications in peer-reviewed conferences and journals. His research interests and current activities focus on Cloud Computing, Cloud Security, Information Security, Network Security, E-Health Security, Internet of Things Security, and Fog Computing.

Dr. Alosaimi has been awarded as a distinguished student from Saudi Cultural Bureau in London twice in 2014 and 2016, and he has been awarded as a distinguished student in 2015 from his highness prince Mohammed Bin Nawaf the Ambassador of Saudi Arabia to the United Kingdom.

Mohammed Ali R was born in Taif, Saudi Arabia in 1980. He received the B.S. in Computer Science from Umm Alqura University and M.S. degrees in IT-Management in 2012 from University Technology Malaysia, Kuala Lumpur, and the Ph.D. degree in Computer Science from University Technology Malaysia, Kuala Lumpur, in 2016.

From 2016 to 2020, he was a Professor Assistant in Taif University. He is IEEE member of Western Section of Saudi Arabia. He is one of Editorial board member of Korean Journal of Next Generation Information Technology, (JNIT), Korea. Also, he is a consultant of Multaqa Makkah in the Governorate Makkah Region. He is author of more than 10 articles. Moreover, he is the founder of OneNation company which include some mobile applications products such as Localjoyksa and OneNation website and more tourism activities. His research interests include Social Media Mining, Big Data, Human Computer Interaction, IT-Project Management and Machine Learning.

Dr. Alzahrani was a recipient of the Guinness World Record Certificate of The largest IT-Hackathon in the world under the Saudi Federation for Cyber Security Programing and Drones in 2018. He has a Project Management Professional (PMP) certificate form EC-council. Also, he has R-language professional certificate from Prince Norah university, 2019. International Conference on Society and Information Technologies: ICSIT 2012 top 10 Paper Award.