Investigating the Efficiency of WordNet as Background Knowledge for Document Clustering

Iyad AlAgha[1], Rami Nafee[2]

[1]*Faculty of Information Technology, The Islamic University of Gaza, ialagha@hotmail.com*
[2]*Faculty of Information Technology, The Islamic University of Gaza, raminafe2002@hotmail.com*

Abstract—Traditional techniques of document clustering do not consider the semantic relationships between words when assigning documents to clusters. For instance, if two documents talk about the same topic but by using different words, these techniques may assign documents to different clusters. Many efforts have approached this problem by enriching the document's representation with background knowledge from WordNet. These efforts, however, often showed conflicting results: While some researches claimed that WordNet had the potential to improve the clustering performance by its capability to capture and estimate similarities between words, other researches claimed that WordNet provided little or no enhancement to the obtained clusters. This work aims to experimentally resolve this contradiction between the two teams, and explain why WordNet could be useful in some cases while not in others, and what factors can influence the use of WordNet for document clustering. We conducted a set of experiments in which WordNet was used for document clustering with various settings including different datasets, different ways of incorporating semantics into the document's representation and different similarity measures. Results showed that different experimental settings may yield different clusters: For example, the influence of WordNet's semantic features varies according to the dataset being used. Results also revealed that WordNet-based similarity measures do not seem to improve clustering, and that there was no certain measure to ensure the best clustering results.

Index Terms—Document Clustering, WordNet, Similarity Measure, Ontology

I INTRODUCTION

Document clustering is a technique that aims at grouping document collections into meaningful groups. In traditional techniques of document clustering, documents are represented as bag of words, and are then assigned to clusters according to the similarity scores obtained from a document similarity measure. These techniques ignore the semantic relationships between the document words, and thus cannot accurately group documents based on meaning similarity. For example, a document that only contains the word "plane" and another that only contains the word "jet" are assigned to different clusters as the two words will be considered different.

Existing research has tried to overcome this limitation by proposing clustering techniques that are based on meaning similarities. The similarity in meaning can be measured by exploiting background knowledge in form of domain ontologies or lexical resources such as WordNet. Similarity scores obtained from WordNet can be used to enhance the document's representation by giving more weight to words that are semantically related [1]. With the enhanced document's representation, the clustering algorithm can better assign documents to clusters based on their semantic similarities to each other. Several efforts have investigated different approaches to incorporate the semantic features of ontologies in an attempt to improve document clustering, and have shown that information semantics have the potential to improve the quality of the obtained clusters [2-5].

WordNet [6] is one of the most popularly used semantic networks for determining semantic relations between words. WordNet has an ontology alike structure: words are represented as having several meanings (each such meaning forming a synset, which is the atomic structure of WordNet), and relations between words (hyponymy, hyperonymy, antonymy, and other relations) are represented as links in a graph. Many similarity measures use the relations defined in WordNet to determine the semantic relatedness between words. Due to its wide coverage as compared to other restricted domain ontologies, many efforts used it as background knowledge for document clustering [7-9].

Despite the significant amount of research on WordNet-based clustering, existing approaches often resulted in conflicting results: while some approaches showed that WordNet could enhance the document's representation with semantic features, yielding to better clustering [9-11], other approaches claimed that WordNet resulted in little or no improvement, or might even degrade the clustering results due to the introduced noise [7, 12, 13]. Given this contradiction, the objective of this research is to try to resolve this issue by seeking answers to the following questions:

- What potential factors could make WordNet useful for document clustering in particular situations while not in others?
- Do different experimental settings, i.e. different datasets, document's representations and similarity measures affect the potential of WordNet to improve clustering?
- What is the best similarity measure to use with WordNet-based clustering?

II Related Work

The idea of incorporating semantic features from the WordNet has been widely explored to improve document clustering techniques. However, there were major differences among the findings of these efforts: while some efforts affirmed the value of WordNet in improving document clustering, other efforts indicated the opposite. For example, Hotho et al. [1] discussed different strategies for representing text documents that take background knowledge from WordNet into account. Their performed evaluations indicated improved clustering results. Gharib, Fouad, and Aref [14] matched the stemmed keywords to terms in WordNet for word sense disambiguation. Their approach outperformed the traditional clustering techniques; however, it seemed to over generalize the affected keywords [15]. Fodeh et al. [13] addressed the effect of incorporating the polysemous and synonymous nouns into document clustering, and showed that they play an important role in clustering. Chen et. al [16] proposed a document clustering approach that combines fuzzy association rule mining with WordNet for grouping documents. They also proposed a technique to filter out noise when adding hypernyms into documents. Wei et. al [17] presented a WordNet-based semantic similarity measure for word sense disambiguation whereas lexical chains are employed to extract core semantic features that express the topic of documents.

In contrast, some studies indicated that the use of WordNet as background knowledge does not necessarily lead to better clusters, and may even produce noise that degrades the clustering performance. For example, Dave et al. [18] used synsets as features for the document's representation and subsequent clustering, and reported that WordNet synsets actually decreased or added no value to clustering performance. Amine et al. [19] found that the mapping of document words to concepts in WordNet might increase ambiguity and induce loss of information. Passos, A. and J. Wainer [20] showed that many similarity measures between words derived from WordNet are worse than the baseline for the purposes of text clustering, and indicated that WordNet does not provide good word similarity data. However, they worked on a single dataset, and did not examine other approaches of incorporating WordNet's features into the document's representation. Sedding, J. and D. Kazakov [7] showed that synonyms and hypernyms disambiguated only by Part-of-Speech tags are not successful in improving clustering effectiveness. This could be attributed to the noise introduced by all incorrect senses that are retrieved from WordNet.

The above discussion reveals inconsistent results regarding the ability of WordNet to operate as background knowledge for document clustering. This demands further investigation into the factors and circumstances causing this inconsistency.

Approaches that exploit WordNet or any other ontology for clustering often rely on some types of semantic similarity measures to estimate the similarity between document words. These measures can be classified into four groups: path length based measures, information content based measures, feature based measures, and hybrid measures. An exhaustive overview of these approaches can be found in [21]. A previous study [22] compared the use of different similarity measures with medical ontologies for document clustering, and indicated that there was no a certain type of similarity measure that significantly outperforms the others. Our study also compares the use of similarity measures for clustering but with WordNet rather than domain-specific ontologies. We also examine the effect of WordNet's semantics with different datasets and document's representations. Amine et al. [19] compared three different clustering algorithms which were all based on the synsets of WordNet as terms for the representation of documents. While their study aimed to determine the best clustering algorithm to use with WordNet, this study aims to explain the opposite findings regarding the efficiency of WordNet for document clustering.

III Using WordNet to Enhance the Document's Representation

Clustering of a document collection typically starts by representing each document as a bag of words. The simple bag of words representation may be enhanced by weighting the terms according to their information content by, for example, tf-idf. Subsequently, a similarity measure, such as the cosine similarity, is used to assign a score to each pair of documents, and similar documents are accordingly assigned to the same cluster. There are two approaches that are commonly used to enhance the document's representation with WordNet, which are explained in what follows:

A Enhancing the Document's Representation by Replacing Synonyms

One limitation of using the traditional bag of words representation is that words are weighted separately without considering the similarity between them. For example, the terms <smart, brilliant, bright> are weighted separately despite of all being synonyms. This leads to information loss as the importance of a determinate concept is distributed among different components of the document's representation.

Existing approaches [1, 13] addressed this issue by referring to lexical databases such as WordNet to identify synonyms. Subsequently, the document's bag of words is modified by replacing all synonyms with a single descriptor. For

example, the terms: <smart, brilliant, bright> may be replaced with the term <intelligent>. Afterwards, the document is represented by using the tf-idf scheme. Therefore, the replacing term will have a cumulative weight that is equal to the sum of the tf-idf weights of replaced synonyms. Finally, a clustering algorithm, such as K-means, is applied.

B Enhancing the Document's Representation by Using Similarity Measures

Having documents with different term sets does not necessarily mean that documents are unrelated. Document terms can be semantically related even though they are syntactically different. For example, the terms: cow, meat, milk and farms are all related with some relations which cannot be captured without using background knowledge. As discussed earlier, the document's representation can be enhanced by identifying and replacing synonyms of the same term. However, this approach only considers synonyms, while terms that are not synonyms but are semantically related are ignored. For example, the words: <camel> and <desert> are related in meaning, and this relation will not be captured by simply identifying and replacing synonyms. To overcome this limitation, it is necessary to represent the document in a way that reflects the relatedness in meanings between the document terms.

Similarity measures have been commonly used to measure the semantic relatedness between documents words, and then relatedness scores are incorporated into the document's representation. Similarity measures exploit knowledge retrieved from a semantic network (i.e., WordNet) to estimate the similarity between term pairs according to the topology structure of WordNet. Similarity scores are then incorporated into the document's tf-idf representation so that terms are related will gain more weight. Reweighting terms according to their semantic relatedness may help discount the effects of class-independent general terms and aggravate the effects of class-specific "core" terms [22]. This can eventually help to cluster documents based on their meanings. Employing similarity measures on WordNet is an idea that have been explored by several efforts [3, 17, 23] for the purpose of improving document clustering.

IV Experimental Study

After presenting the approaches for enhancing the document's representation with knowledge in WordNet, the following subsections report on the experimental study we conducted with the following objectives in mind: 1) Compare between the approaches previously explained and examine their influence on document clustering by testing with different datasets. 2) Examine the use of different ontology-based similarity measures in order to identify the best measure(s) to use with WordNet. 3) Explain, in light of the results obtained from 1 and 2, the contradiction between existing works regarding the value of WordNet's semantics for document clustering.

A Datasets

Two datasets were used for the study, which were: Reuters-21578, and OHSUMED. Details of each dataset are given below. In addition, the rationale behind using these particular datasets is illustrated.

- *Reuters-21578 [24]* : The documents in the Reuters-21578 collection appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by personnel from Reuters Ltd[24]. Reuters-21578 dataset has been widely used for evaluating document clustering algorithms. Its domain is not specific, therefore it can be understood by a non-expert [7].
- *OHSUMED [25]:* The OHSUMED test collection is a set of 348,566 references from MEDLINE, the on-line medical information database, consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). The available fields are title, abstract, MeSH indexing terms, author, source, and publication type. The National Library of Medicine has agreed to make the MEDLINE references in the test database available for experimentation [26].

The above datasets were chosen because they have different characterises that might lead to different clustering performance: the Reuters-21578 is considered a heterogeneous dataset with no specific domain, covering a wide verity of dissimilar topics from the newswire. In contrast, OHSUMED is a domain-specific dataset strictly covering the domain of medicine. The intention was to explore how the use of datasets of different homogeneity could yield different clustering performance.

B Experiments

We conducted three experiments, each of which used a different approach of representing documents. K-means clustering algorithm was applied in the three experiments. These experiments were as follows:

- Traditional clustering without background knowledge: This represented the baseline case. Documents were pre-processed by applying tokenization, stemming and stop-word removal. Documents were then represented in tf-idf prior to applying K-means for clustering. Note that the conceptual relations between document terms were ignored, and terms were weighted only according to their frequency of co-occurrence in the document collection.
- *Enhancing the document's representation by identifying and replacing synonyms:* Before finding synonyms in documents, the following pre-processing steps were applied: First, all documents were broken down into sentences which undergone part of speech tagging. Part of speech tags were essential to correctly identify synonyms which should have the same POS tag. After tagging the document words, other pre-processing steps including tokenization, stemming and stop-word removal were applied. The following step was to search the document collection for terms that are synonyms with the help of WordNet. Synonyms of a particular concept were replaced by a unique term in the document's bag of words. The modified bag of words of each document was

then represented in tf-idf.

- *Enhancing the document's representation by using ontology-based similarity measures:* First, pre-processing steps consisting of tokenization, stop-word removal and stemming were applied on the document collection. Documents were then represented in tf-idf scheme. Ontology based similarity measures were used to measure the WordNet-based similarity between each pair of words in the document collection. Similarity scores were then used to augment the tf-idf weights so that terms gained more weight according to their similarity to each other. This process is formally represented as the following: Let $d = \{w_1, w_2, w_3, .., w_n\}$ be the document's representation where w_i is the weight of term t_i in document d, and is computed using the tf-idf scheme. The similarity between each pair of terms in the document was calculated by using each similarity measure shown in Table 1. Afterwards, if-idf weights were reweighted using the following equation [27]:

$$w'_i = w_i + \sum_{j=0, j\neq i}^{m} w_j * sim(i, j) \qquad (1)$$

where: w'_i stands for the augmented tf-idf weight of term i, w_j is the tf-idf weight of term j of the same document, and $sim(i, j)$ is the semantic similarity score between terms t_i, t_j rated from 0 to 1, where 1 represents the highest similarity. This equation assigns more weight to terms that are semantically related. Weights of terms that are not related to any other terms or that are not included in WordNet remain unchanged.

After augmenting the tf-idf document's representation with similarity scores, K-means algorithm was applied. Since it was of our objectives to assess different similarity measures, the above process was repeated for every similarity measure shown in Table 1. These measures have been widely used for semantically-enhanced text clustering. Short descriptions of these measures are also given in Table 1.

TABLE 1
Similarity measures used in the study.

ID	Description
LCH	Leacock and Chodorow [28]: This measure relies on the length of the shortest path between two terms. It is limited to IS-A links, and the path length is scaled by the overall depth of the taxonomy.
WUP	Wu and Palmer [29]: This measure calculates similarity by considering the depths of the two terms in WordNet, along with the depth of the least common subsumer.
JCN	Jiang and Conrath [30]: This measure uses the notion of information content, but in the form of the conditional probability of encountering an instance of a child-synset given an instance of a parent synset.
LIN	Lin[31]: Math equation is modified a little bit from Jiang and Conrath: 2 * IC(lcs) / (IC(synset1) + IC(synset2). Where IC(x) is the information content of x.
RES	Resnik [32]: This measure defined the similarity between two terms to be the information content of the most specific common subsume.
LESK	Banerjee and Pedersen [33]: The relatedness of two terms is proportional to the extent of overlaps of their dictionary definitions.
HSO	Hirst and St-Onge [34]: Two terms are semantically related if their WordNet synsets are connected by a path that is not too long and that does not change direction too often.

V Results and Discussion

The clustering performance was evaluated by using F-measure [35] and purity[36]. Table 2 summarizes the results whereas rows indicate the three experiments and columns indicate the two datasets used. When using similarity measures with WordNet, the clustering process was repeated several times while varying the similarity measure and the best result was considered for comparison. The ID of the similarity measure giving the best result is shown alongside the result between brackets (IDs of similarity measures are shown in Table 1).

TABLE 2
Clustering results of the three experiments.

Experiment	Reuters-21578		OHSUMED	
	Purity	F-measure	Purity	F-measure
Without Background Knowledge	0.57	0.64	0.36	0.47
With Replacing Synonyms	**0.64**	**0.77**	**0.49**	**0.65**
With WordNet-based Similarity Measures	0.59 (LCH)	0.70 (LCH)	0.43 (RES)	0.60 (RES)

Table 3 lists the different similarity measures we used for the third experiment, i.e. clustering with WordNet-based similarity measures, and the clustering performance per each measure. Results are discussed in the following subsections, and related efforts are revisited, where appropriate, in light of our results.

TABLE 3
Clustering results for each similarity measure.

Similarity Measures	Reuters-21578		OHSUMED	
	Purity	F-measure	Purity	F-measure
LCH	**0.59**	**0.70**	0.39	0.49
WUP	0.56	0.64	0.41	0.55
JCN	0.40	0.48	0.30	0.39
LIN	0.48	0.55	0.41	0.55
RES	0.48	0.61	**0.43**	**0.65**
LESK	0.54	0.67	0.42	0.57
HSO	0.46	0.58	0.42	0.62

A Comparing the Document's Representation Techniques

In this subsection, we compare results across the three experiments, i.e. clustering without background knowledge (baseline), clustering with replacing synonyms and clustering with similarity measures.

In the case of Reuters dataset, clustering with replacing synonyms outperformed other approaches (F-measure =0.77 and purity =0.64), followed by clustering with similarity measures (F-measure=0.70, Purity=0.59).When using the OHSUMED dataset, the best results were also achieved by replacing synonyms (F-measure=0.65, Purity=0.49), followed by clustering with similarity measures (F-measure=0.60, Purity=0.43). In general, this result shows that potential of WordNet to improve the clustering results, either by replacing synonyms or by using similarity scores, as compared to clustering without background knowledge. This result conforms to other studies which indicated the value of WordNet semantics for document clustering [1, 7, 11, 17].

However, the use of similarity measures with WordNet has unexpectedly produced results worse than those produced by replacing synonyms, but slightly better than the baseline case, i.e. clustering without background knowledge. It should be noticed that Table 2 shows the top result obtained from all the seven similarity measures. This result concurs with some studies which indicated that the use of similarity measures with WordNet had little impact on text clustering and may produce worse results [20, 23].

The above result suggests that WordNet-based similarity measures do not seem to improve the clustering results. We think that this can be attributed to the structure of WordNet taxonomy which is mainly designed to represent specific relations (e.g. hyponymy, hyperonymy) but is not designed to capture similarity between words. For example, when measuring the similarity between the words: “camel” and “desert”, or between the verb “sit” and the noun “chair”, the similarity scores were close to 0.

B Comparing the Influence of Datasets

Comparing results across the two datasets, we can see that the improvement resulted from semantic-based approaches (synonyms and similarity measures) was more obvious in the case of the OHSUMED dataset than in the Reuters dataset. We think this difference can be explained by the nature of the dataset in terms of the disparity of its content: For example, the Reuters dataset is heterogeneous in the sense that it covers news from unrelated domains, a thing that makes it difficult to identify semantic relations between the document words. It was noticed experimentally that the scores obtained by applying the similarity measures on the Reuters dataset were often low. Considering the fact that most similarity measures rely, mainly or partially, on the taxonomical distances within WordNet, the similarity scores will decrease as the differences between terms increases. In contrast, OHSUMED is a domain-specific dataset with different classes of documents belonging to the medical domain. This makes it easy to identify terms that belong to a specific domain and measure similarities between them. This explains the better results obtained from the OHSUMED dataset as compared to the results obtained from the Reuters dataset.

The above discussion reveals that the use of different datasets can result in different results: The more homogeneous and domain-specific the dataset is, the easier it becomes to capture similarities between terms included in the dataset, hence the more influence the WordNet has on the clustering results. We should also bear in mind that the WordNet is a general-purpose lexical database of English terms but it does not provide a thorough coverage of every domain of knowledge. Although its use has improved the clustering performance in our experiment, WordNet is not meant to be used with domain specific applications. It is always recommended to use domain-specific ontologies to cover domain-specific datasets.

C Comparing the WordNet-based Similarity Measures

Comparing the use of different similarity measures, results vary as shown in Table3: in the case of Reuters dataset, the LCH measure achieved the best results followed by the PATH and WUP measures. In the case of OHSUMED, the RES measure gave the best results, followed by the HSO and LESK. However, the improvement on the results was not significant [t-test, $p>0.05$]. In addition, the clustering performance with some similarity measures was even lower than the performance of the baseline case where no background knowledge was used (e.g. refer to JCN measure in Table 3). These results indicate that there was no certain measure to ensure the best clustering results. They also support our conclusion about the inadequacy of WordNet to be used with ontology-based similarity measures. However, this result does not generalize to other types of ontologies as our study focused strictly on WordNet.

VI Conclusions and Recommendations

The different outcomes of exsiting approaches regarding the influence of WordNet on document clustering have motivated us to conduct this work. Multiple experiments on document clustering were conducted with multiple datasets, document’s representations and similarity measures. In summary, our study found that the characteristics of the datasets being clustered in terms of the disparity of its topics may reduce the ability to capture the semantic relations between terms from the taxonomy of WordNet. Results also indicated that augmenting the document’s representation by replacing synonyms may achieve better results than those achieved by using similarity measures or the baseline case, i.e. clustering without background knowledge.

Based on these findings, we draw some recommendations to be considered when using WordNet as background knowledge for text clustering: First, experimenters should consider the nature of the dataset in hand and the diversity of its topics before deciding to use WordNet for measuring similarities. Second, the WordNet structure does not seem to support the application of similarity measures. Alternatively, WordNet can be better exploited by capturing specific types of relations such as “IS-A”, “hyponymy” and hypernymy,

and then use them to enhance the document's representation. For example, capturing and replacing synonyms in the document collection outperformed other approaches in our experiments.

REFERENCES

[1] A Hotho, S Staab, andG Stumme (2003). Ontologies Improve Text Document Clustering, *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*: IEEE (pp. 541-544).

[2] A Charola, andS Machchhar. (2013). Comparative Study on Ontology Based Text Documents Clustering Techniques. *Data Mining and Knowledge Engineering*, 5(12), 426.

[3] H H Tar, andT T S Nyunt. (2011). Ontology-Based Concept Weighting for Text Documents. *world Academy of Science, engineering and Technology*, 81(249-253.

[4] G Bharathi, andD Venkatesan. (2012). Study of Ontology or Thesaurus Based Document Clustering and Information Retrieval. *Journal ofEngineerir1g and Applied Sciences*, 7(4), 342-347.

[5] Q Dang, J Zhang, Y Lu, andK Zhang (2013). Wordnet-Based Suffix Tree Clustering Algorithm, *2013 International Conference on Information Science and Computer Applications (ISCA 2013)*: Atlantis Press (pp.

[6] G Miller, andC Fellbaum, "Wordnet: An Electronic Lexical Database", MIT Press Cambridge1998.

[7] J Sedding, andD Kazakov (2004). Wordnet-Based Text Document Clustering, *Proceedings of the 3rd Workshop on RObust Methods in Analysis of Natural Language Data*: Association for Computational Linguistics (pp. 104-113).

[8] H-T Zheng, B-Y Kang, andH-G Kim. (2009). Exploiting Noun Phrases and Semantic Relationships for Text Document Clustering. *Information Sciences*, 179(13), 2249-2262.

[9] D R Recupero. (2007). A New Unsupervised Method for Document Clustering by Using Wordnet Lexical and Conceptual Relations. *Information Retrieval*, 10(6), 563-579.

[10] A Hotho, Staab, S. and Stumme, G (2003). Wordnet Improves Text Document Clustering, *in Proc. of the Semantic Web Workshop at 26th Annual International ACM SIGIR Conference*, Toronto, Canada (pp.

[11] Y Wang, andJ Hodges (2006). Document Clustering with Semantic Analysis, *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*: IEEE (pp. 54c-54c).

[12] L Jing, L Zhou, M K Ng, andJ Z Huang, "Ontology-Based Distance Measure for Text Clustering", *Proceedings of the Text Mining Workshop, SIAM International Conference on Data Mining*2006.

[13] S Fodeh, B Punch, andP-N Tan. (2011). On Ontology-Driven Document Clustering Using Core Semantic Features. *Knowledge and information systems*, 28(2), 395-421.

[14] T F Gharib, M M Fouad, andM M Aref, "Fuzzy Document Clustering Approach Using Wordnet Lexical Categories", *Advanced Techniques in Computing Sciences and Software Engineering*Springer. pp. 181-186. 2010.

[15] C Bouras, andV Tsogkas. (2012). A Clustering Technique for News Articles Using Wordnet. *Knowledge-Based Systems*, 36(115-128.

[16] C-L Chen, F S Tseng, andT Liang. (2011). An Integration of Fuzzy Association Rules and Wordnet for Document Clustering. *Knowledge and information systems*, 28(3), 687-708.

[17] T Wei, Y Lu, H Chang, Q Zhou, andX Bao. (2015). A Semantic Approach for Text Clustering Using Wordnet and Lexical Chains. *Expert Systems with Applications*, 42(4), 2264-2275.

[18] K Dave, S Lawrence, andD M Pennock (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, *Proceedings of the 12th international conference on World Wide Web*: ACM (pp. 519-528).

[19] A Amine, Z Elberrichi, andM Simonet. (2010). Evaluation of Text Clustering Methods Using Wordnet. *Int. Arab J. Inf. Technol.*, 7(4), 349-357.

[20] A Passos, andJ Wainer (2009). Wordnet-Based Metrics Do Not Seem to Help Document Clustering, *Proc. of the of the II Workshop on Web and Text Intelligente, São Carlos, Brazil*, (pp.

[21] L Meng, R Huang, andJ Gu. (2013). A Review of Semantic Similarity Measures in Wordnet. *International Journal of Hybrid Information Technology*, 6(1), 1-12.

[22] X Zhang, L Jing, X Hu, M Ng, andX Zhou, "A Comparative Study of Ontology Based Term Similarity Measures on Pubmed Document Clustering", *Advances in Databases: Concepts, Systems and Applications*Springer. pp. 115-126. 2007.

[23] L Jing, L Zhou, M K Ng, andJ Z Huang (2006). Ontology-Based Distance Measure for Text Clustering, *Proceedings of the Text Mining Workshop, SIAM International Conference on Data Mining*, (pp. 537-541).

[24] D D Lewis. (1997). Reuters-21578 Text Categorization Test Collection, Distribution 1.0. *http://www. research. att. com/~ lewis/reuters21578. html*,

[25] W Hersh, C Buckley, T Leone, andD Hickam (1994). Ohsumed: An Interactive Retrieval Evaluation and New Large Test Collection for Research, *SIGIR'94*: Springer (pp. 192-201).

[26] X U Group, 07/15/2005 [cited 2014 1/12/2014]; Available from: http://davis.wpi.edu/xmdv/datasets/ohsumed.html.

[27] G Varelas, E Voutsakis, P Raftopoulou, E G Petrakis, andE E Milios, "Semantic Similarity Methods in Wordnet and Their Application to Information

Retrieval on the Web", *Proceedings of the 7th annual ACM international workshop on Web information and data management*ACM: NewYork. pp. 10-16. 2005.
[28] C Leacock, andM Chodorow. (1998). Combining Local Context and Wordnet Similarity for Word Sense Identification. *WordNet: An electronic lexical database*, 49(2), 265-283.
[29] Z Wu, andM Palmer (1994). Verbs Semantics and Lexical Selection, *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*: Association for Computational Linguistics (pp. 133-138).
[30] J J Jiang, andD W Conrath. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *arXiv preprint cmp-lg/9709008*,
[31] D Lin (1998). An Information-Theoretic Definition of Similarity, *ICML*, (pp. 296-304).
[32] P Resnik. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *arXiv preprint cmp-lg/9511007*,
[33] S Banerjee, andT Pedersen, "An Adapted Lesk Algorithm for Word Sense Disambiguation Using Wordnet", *Computational Linguistics and Intelligent Text Processing*Springer. pp. 136-145. 2002.
[34] G Hirst, andD St-Onge. (1998). Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. *WordNet: An electronic lexical database*, 305(305-332.
[35] B Larsen, andC Aone, "Fast and Effective Text Mining Using Linear-Time Document Clustering", *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*ACM. pp. 16-22. 1999.
[36] Y Zhao, andG Karypis. (2001). Criterion Functions for Document Clustering: Experiments and Analysis. *Machine Learning*,

Iyad M. AlAgha received his MSc and PhD in Computer Science from the University of Durham, the UK. He worked as a research associate in the center of technology enhanced learning at the University of Durham, investigating the use of Multi-touch devices for learning and teaching. He is currently working as an assistant professor at the Faculty of Information technology at the Islamic University of Gaza, Palestine. His research interests are Semantic Web technology, Adaptive Hypermedia, Human-Computer Interaction and Technology Enhanced Learning.

Rami H. Nafee received his BSc from Al-Azhar University-Gaza and his MSc degree in Information technology from the Islamic University of Gaza. He works as a Web programmer at Information Technology Unit at Al-Azhar University-Gaza. He is also a lecturer at the Faculty of Intermediate Studies at Al-Azhar University. His research interests include Data Mining and Semantic Web.