

Comparison of Keyword-based and Semantic-based Web Page Clustering Systems

Ei Ei Moe¹, Hnin Hnin Htun²

¹Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar

Abstract: Today, web page clustering is useful for many applications such as categorization, cleaning, schema detection and automatic extractions. Web page clustering is classified into different categories that are hierarchical and flat clustering, online and offline clustering, soft and hard clustering, and document-based and keywords-based clustering. Among them, keyword-based web page clustering uses the single words or compounds words occurring in the web page set as the features for clustering. In this situation, these words can't precisely represent the content of the web page because the synonyms and polysemous of the word can lead the ambiguity problems. Semantic analysis is useful to solve this ambiguity problem. So, this system proposes both keyword-based and semantic-based web page clustering system, and then compares the performance between them. In the semantic analysis, words in each web page are first mapped to word senses by using supervised based word sense disambiguation method. Then, semantic-based web page clustering system uses both keywords and semantic features for clustering. After performing each cluster process, this system points out the semantic-based web page clustering system is more precise and effective than the keyword-based clustering system.

Keywords: Semantic, Word Sense Disambiguation, Clustering.

1. Introduction

With the enormous success of the Information Society and the World Wide Web, the amount of textual electronic information available has significantly increased. The increasing size and dynamic content of the world wide web has created a need for automated organization of web-pages. Web page clustering plays an important role in providing intuitive navigation and browsing mechanisms by organizing such large amounts of information into a small number of meaningful clusters. Web page clusters can provide a structure for organizing large bodies of text for efficient browsing and searching.

Web page clustering is the automatic discovery of web page groups in a document collection, where the formed clusters have a high degree of association between members, whereas members from different clusters have a low degree of association. Web page clustering groups the documents in an unsupervised way where there is no label or class information. Web page clustering has many applications such as information retrieval; cluster based browsing and bio-informatics applications.

Web page clustering is very important and useful in the information retrieval area. Web page clustering can be applied to a web page database so that similar web pages are related in the same cluster. During the retrieval process, web pages belonging to the same cluster as the retrieved web pages can also be returned to the user. Applying web page clustering to the retrieved web pages could make it easier for the users to browse their results and locate what they want.

By using semantic technology, this system supports better performance about web page clustering system. So, this system is proposed to provide information retrieval and cluster-based browsing application. By applying the combination use of word sense disambiguation (WSD) and clustering methods, this system proposes the semantic-based

web page clustering process. In this semantic-based web page clustering process, this system uses both the supervised based word sense disambiguation method and enhanced Agglomerative algorithm. To point out the semantic that is effective for web page clustering system, this system compares the performance between the semantic-based and keyword-based web page clustering system.

The rest of the paper is organized as follows: related work is described in section 2. Pre-processing of web page is shown in section 3. Keyword-based web page clustering is presented in section 4. Semantic-based web page clustering is expressed in section 5. Proposed system design is presented in section 6. Explanation of the system is described in section 7. Experimental results are shown in section 8. Finally, conclusion is given in section 9.

2. Related Work

In 2014, I. Alagha and R. Nafee [2] presented an efficient approach for semantically enhanced document clustering by using Wikipedia link structure. Traditional techniques of document clustering do not consider the semantic relationships between words when assigning documents to clusters. They presented a new approach to enhance document clustering by exploiting the semantic knowledge contained in Wikipedia. They first map terms within documents to their corresponding Wikipedia concepts. Then, similarity between each pair of terms is calculated by using the Wikipedia's link structure. The document's vector representation is then adjusted so that terms that are semantically related gain more weight. Empirical results showed that their approach improved the clustering results as compared to other approaches.

In 2014, S. Romeo, A. Tagarelli and D. Ienco [3] presented semantic-based multilingual document clustering via tensor modeling. A major challenge in document clustering research arises from the growing amount of text data written in

different languages. Previous approaches depend on language-specific solutions (e.g., bilingual dictionaries, sequential machine translation) to evaluate document similarities, and the required transformations may alter the original document semantics. To cope with this issue, they proposed a new document clustering approach for multilingual corpora that (i) exploits a large-scale multilingual knowledge base, (ii) takes advantage of the multi-topic nature of the text documents, and (iii) employs a tensor-based model to deal with high dimensionality and sparseness.

In 2015, G. Tang and Y. Xia [4] proposed the cross-lingual document clustering that is the task of automatically organizing a large collection of multi-lingual documents into a few clusters, depending on their content or topic. It is well known that language barrier and translation ambiguity are two challenging issues for cross-lingual document representation. To this end, they proposed to represent cross-lingual documents through statistical word senses, which are automatically discovered from a parallel corpus through a novel cross-lingual word sense induction model and a sense clustering method.

According to literature and concepts pointed out from the previous works, this system is intended to present the performance comparison between keyword-based and semantic-based web page clustering system.

3. Pre-processing of Web Page

Web pages are initially in the unstructured format. The first and foremost step is to preprocess this text to form numeric vectors (features). The preprocessing phase consists of the following steps [1]:

- Tokenization: In the step, the web page is broken into words or tokens.
- Stop-Word Removal: Stop words are the words that are irrelevant to the processing to be performed. Words like articles, conjunctions, and verbs etc. which have no impact what so ever in calculating similarity between web pages are removed.
- Stemming: It is the process of converting each word to its root form.
- TF-IDF Weighting Scheme: Once the document has gone through the first 3 steps, a bag of words is got where each word is in the root form. TF-IDF weighing scheme is used to convert the web page from bag of words to numeric vectors where each word in the vocabulary now represents a dimension [1]. TF-IDF term weight is given as follows:

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \dots, f_{vj}\}} \quad (1)$$

$$idf_i = \log \frac{N}{df_i} \quad (2)$$

$$w_{ij} = tf_{ij} \times idf_i \quad (3)$$

In this scheme, N is total number of web pages in the system. The df_i is number of web pages in which term t_i appears at least once. The f_{ij} is the raw frequency count of term t_i in web

page d_j . The tf_{ij} is the normalized term frequency. The idf_i is the inverse web page frequency of term t_i .

4. Keyword-based Web Page Clustering

Clustering analysis is an important technique in data mining. Hierarchical clustering is a classical and popular clustering algorithm. Agglomerative hierarchical clustering is the method to build a bottom-to-top hierarchical decomposition of the data set on the basis of dissimilarities between objects. The clustering result is illustrated using a dendrogram offering easy interpretation by a decision maker [5]. Web page clustering is the process which is carried on the organization or division to the web page set under the condition of un-learning [6].

Hierarchical clustering accuracy is relatively high, but when each class merges, it needs to compare all classes' similarity in the global and selecting the most similar of two classes, so it's relatively slow. The defect of hierarchical clustering is that once a step (merge or split) completed, it cannot be revoked, so it can not correct the wrong decision. Hierarchical clustering methods are divided into bottom-up (merge) and top-down (split) hierarchical clustering methods. Top-down agglomerative hierarchical clustering method starts from the object's complete works, and gradually be divided into more categories [6].

Agglomerative hierarchical clustering process starts by placing each object in its own cluster and then merges these atomic clusters into larger clusters, until certain termination conditions are satisfied. In this process, there are four steps. These are as follows [7]:

- Step 1: Start with N clusters, each containing a single entity, and an $N \times N$ symmetric matrix of similarities (using Jaccard coefficient similarity measure method). Let d_{ij} = similarity between item i and item j.
- Step 2: Search the distance matrix for the nearest pair clusters (i.e., the two clusters that are separated by the smallest distance). Denote the distance between these most similar clusters A and B by d_{AB} .
- Step 3: Merge clusters A and B into a new cluster, labeled S. Update the entries in the distance matrix by deleting the rows and columns corresponding to clusters A and B, and adding a row and column giving the distances between the new cluster S and all the remaining clusters. For merging process, this system uses single linkage method.
- Step 4: Repeat steps (2) and (3) until a total of N-1 times.

4.1 Linkage Method

Linkage methods are hierarchical methods that merging of clusters is based on distance between clusters. Three important linkage methods are single linkage, average linkage and complete linkage. Among them, this system uses the single linkage method. In the single linkage method, the link between two subsets is the shortest distance between them. This single linkage method is as follows:

$$d_{\min}(C_K, C_L) = \min_{p \in C_K, p' \in C_L} |p - p'| \quad (4)$$

In this equation, $|p-p'|$ is the distance between two objects or points, p and p' ; CK and CL are cluster K and L [7].

4.2 Jaccard Coefficient Similarity Method

Jaccard coefficient similarity method gives a useful measure of how similar between two documents. The technique is also used to measure cohesion within clusters. This similarity method is as follows:

$$\text{sim}(d_j, q) = \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sum_{i=1}^{|V|} (w_{ij})^2 + \sum_{i=1}^{|V|} (w_{iq})^2 - \left(\sum_{i=1}^{|V|} w_{ij} \times w_{iq} \right)} \quad (5)$$

where, $\text{sim}(d_j, q)$ is the similarity between training document d_j and testing document q . The w_{ij} is weight of the term t_i within training document d_j and the w_{iq} is weight of the term t_i within testing document q [7].

5. Semantic-based Web Page Clustering

In the semantic-based web page clustering, both the keyword (ambiguous word) and its semantic meaning (disambiguous word) are used as the features for clustering. For solving ambiguous word in each web page, this system uses WordNet and KNN classifier that is the supervised-based WSD method. Then, this system uses the enhanced agglomerative method for web page clustering.

5.1 WordNet

WordNet encodes concepts in terms of sets of synonyms (called synsets). WordNet 3.0 contains about 155,000 words organized in over 117,000 synsets [8]. A synset is a set of word senses all expressing the same meaning. In WordNet, words and their relationships to each other are organized in a hierarchical manner similar to the taxonomies which may be found in the natural sciences. Words which have multiple meanings or "word senses" appear in more than one synset. Each word belongs to a set of synonyms, also known as a synset. WordNet tracks several different semantic relationships for synsets. WordNet is now considered to be a valuable resource for researchers in linguistics, text analysis, and artificial intelligence, among others [9].

5.2 Supervised-based Word Sense Disambiguation

Supervised-based word sense disambiguation (WSD) method uses annotated training corpora as a source of knowledge. WSD can be viewed as a classification task: word senses are the classes, and an automatic classification method is used to assign each occurrence of a word to one or more classes based on the evidence from the context and external knowledge sources. Supervised classification task is used by neural networks, deep learning neural networks, K-nearest neighbours (KNN), support vector machines (SVM), decision lists and naive Bayes classifier [10].

KNN algorithm is one of the most famous classification algorithms used for predicting the class of a record or

(sample) with unspecified class based on the class of its neighbor records [11]. Classification of a new instance is based on the classes of the k most similar stored examples which are referred to as nearest neighbors [12]. KNN algorithm is as follows:

- Step 1: Determine k
- Step 2: Calculate the similarity or distance between the testing data and all the training data.
- Step 3: Sort the distance and determine k nearest neighbors based on the K^{th} minimum distance.
- Step 4: Gather the categories based on majority vote.
- Step 5: Determine the categories based on majority vote.

For similarity calculation, this system uses the Jaccard coefficient similarity method.

5.3 Enhanced Agglomerative Clustering Method

Enhanced agglomerative clustering method has two sub-processes: hierarchical clusters producing process and user desired k -clusters searching process. The hierarchical clusters producing process is performed according to the Agglomerative hierarchical clustering process. Then, the user desired k -clusters searching process is as follows:

- Step 1: Accept the user inputted K value to produce clusters according to the number of K .
- Step 2: Assign the content value to calculate the distance between each hierarchical cluster. For distance calculation, this system uses the Euclidean distance measure method. This method is as follows:

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2} \quad (6)$$

where $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ are two p -dimensional data objects.

- Step 3: Select hierarchical cluster objects randomly. These objects represent initial group centroids. The number of centroid value must be equal to the number of user inputted K value.
- Step 4: Assign each hierarchical cluster object to the group that has the closest centroid.
- Step 5: Recalculate the positions of the centroids after all hierarchical cluster objects have been assigned.
- Step 6: Repeat Steps 4 and 5 until the centroids no longer move.

Finally, this enhanced agglomerative clustering method can produce k -clusters that contain hierarchical cluster result.

6. Proposed System Design

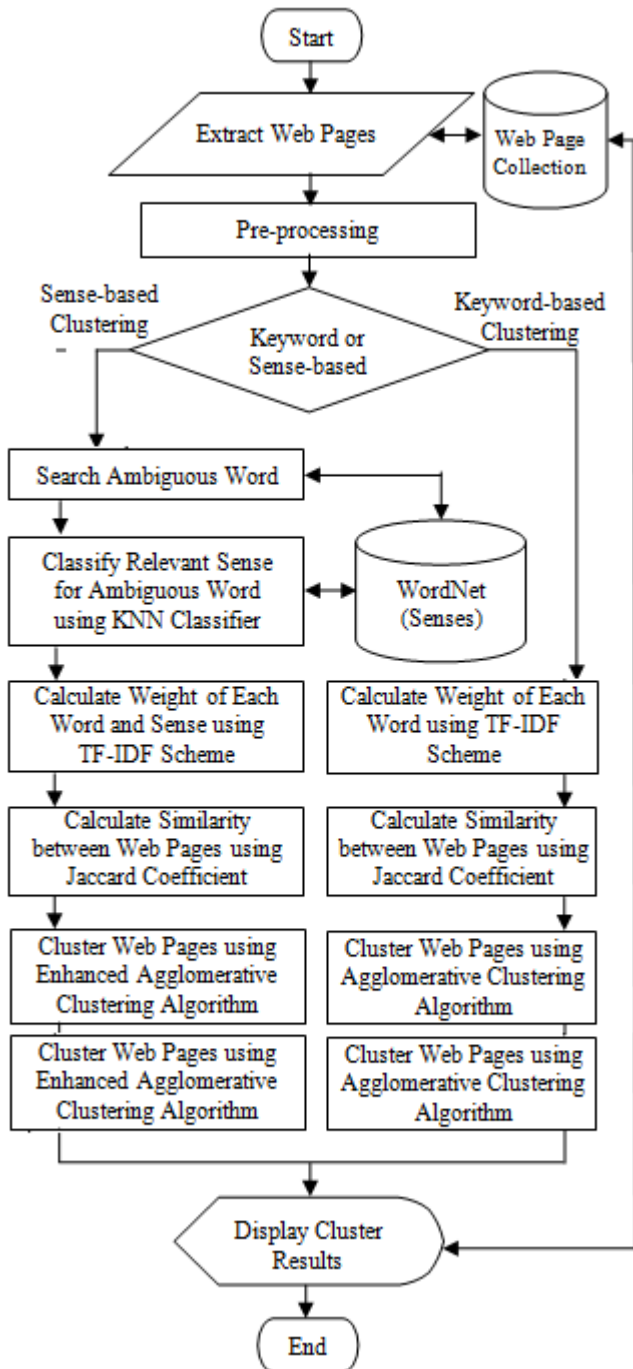


Figure 1: Proposed System Design

The proposed system design is shown in Figure 1. This system is proposed as the performance comparison system between keyword-based and semantic-based web page clustering.

Semantic-based web page clustering system consists of five processes. The first process is the pre-processing that consists of tokenization and stopwords removal. In the second process, this system searches the ambiguous words and then classifies its relevant sense by using WordNet and KNN classifier. Based on TF-IDF (term frequency-inverse document frequency), the third process is weight calculation about each keyword and sense. In the fourth process, this system calculates the similarity between web pages by using Jaccard coefficient similarity method. Finally, this system clusters the

same web page according to the enhanced Agglomerative clustering algorithm.

Keyword-based web page clustering system consists of four processes that are pre-processing, weight calculation about each keyword, similarity calculation between each web pages and clustering to group the same web pages. This clustering system doesn't consider the semantic logic. So, the precision of this system is more eliminate than the semantic-based web page clustering system.

7. Explanation of the System

To show the performance comparison result, sample seven web pages (ambiguous web pages) are tested by using both the keyword-based and semantic-based web page clustering method. These sample web pages are as follows:

- Web page 1: Resultant model is used for unknown classes.
- Web page 2: Classification is the act of distributing things into classes.
- Web page 3: In categorization, a group of people is arranged by category.
- Web page 4: Classification is the basic cognitive process of distributing into classes.
- Web page 5: Prediction is a statement about the future.
- Web page 6: It is a forecast about how the weather will develop.
- Web page 7: There is a need to forecast values of variables.

Keyword-based web page clustering system clusters the same web pages based on only keyword in each web page. Keyword-based cluster results are shown in Table 1.

Table 1: Keyword-based Clustering Results

Hierarchical Cluster Name	Cluster Content (Ambiguous Web Pages)
H - Cluster 1	Web page 2, Web page 4
H - Cluster 2	Web page 2, Web page 4, Web page 1
H - Cluster 3	Web page 6, Web page 7
H - Cluster 4	Web page 3, Web page 5
H - Cluster 5	Web page 6, Web page 7, Web page 3, Web page 5
H - Cluster 6	Web page 2, Web page 4, Web page 1, Web page 6, Web page 7, Web page 3, Web page 5

Semantic-based web page clustering system clusters the same web page based on both keyword and sense in each web page. As a sample, the sense of the "classification" keyword is the "categorization". In this situation, web page 2 and 3 are more similar among other web pages. Sense result for "classification" keyword is shown in Table 2.

Table 2: Sense Result for "Classification" Keyword

ID	Vector Name	Sense Name	Similarity Result
1	Testing vector and Training Vector 1	Categorization, Categorisation, compartmentalization, compartmentalisation, assortment	0.3372
2	Testing vector and Training Vector 2	Categorization, categorisation	0.04828
3	Testing vector and Training Vector 3	Categorization, categorisation, sorting	0.00415

After disambiguating each keyword in each web page, the semantic-based clustering system groups the same web page based on semantic-logic. Semantic-based cluster results are shown in Table 3.

Table 3: Semantic-based Clustering Results

Hierarchical Cluster Name	Cluster Content (Disambiguous Web Pages)
H- Cluster 1	Web page 2, Web page 3
H - Cluster 2	Web page 5, Web page 6
H - Cluster 3	Web page 5, Web page 6, Web page 7
H - Cluster 4	Web page 2, Web page 3, Web page 4
H - Cluster 5	Web page 2, Web page 3, Web page 4, Web page 1
H -Cluster 6	Web page 5, Web page 6, Web page 7, Web page 2, Web page 3, Web page 4, Web page 1

In the semantic-based web page clustering system, the user wants to obtain not only desired cluster number but also the cluster content as hierarchical cluster content. In this sample, the user inputted K value is “2”. Then, this system performs the user desired 2-clusters searching process that is shown in Table 4 and 5.

Table 4: User Desired K-Clusters Producing Process

Step	Cluster 1		Cluster 2	
	Individual	Centroid	Individual	Centroid
1	H-Cluster 2	(0, 0, 0, 0, 1, 1, 0)	H-Cluster 3	(0, 0, 0, 0, 1, 1, 1)
2	H-Cluster 2, H-Cluster 1	(0, 0.5, 0.5, 0, 0.5, 0.5, 0)	H-Cluster 3, H-Cluster 4	(0, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)
3	H-Cluster 1, H-Cluster 2	(0, 0.5, 0.5, 0, 0.5, 0.5, 0)	H-Cluster 3, H-Cluster 4, H-Cluster 5	(0.333, 0.666, 0.666, 0.666, 0.333, 0.333, 0.333)
4	H-Cluster 1, H-Cluster 2, H-Cluster 3	(0, 0.333, 0.333, 0, 0.666, 0.666, 0.333)	H-Cluster 4, H-Cluster 5	(0.5, 1, 1, 1, 0, 0, 0)
5	H-Cluster 2, H-Cluster 3	(0, 0, 0, 0, 1, 1, 0.5)	H-Cluster 4, H-Cluster 5, H-Cluster 1	(0.333, 1, 1, 0.666, 0, 0, 0)
6	H-Cluster 2, H-Cluster 3	(0, 0, 0, 0, 1, 1, 0.5)	H-Cluster 4, H-Cluster 5, H-Cluster 1	(0.333, 1, 1, 0.666, 0, 0, 0)

Table 5: User Desired “2” Cluster Result

Cluster Name	Hierarchical Cluster Content	Cluster Content (Disambiguous Web Pages)
Cluster 1	H-Cluster 2, H-Cluster 3	Web page 5, Web page 6, Web page 7
Cluster 2	H-Cluster 4, H-Cluster 5, H-Cluster 1	Web page 2, Web page 3, Web page 4, Web page 1

8. Experimental Result of the System

There are numerous evaluation measures to validate the cluster quality. To evaluate the clustering results of the proposed system, precision method has been used.

$$\text{Precision (i, j)} = \frac{n_{ij}}{n_j} \quad (7)$$

where, n_{ij} is the number of members of class i in cluster j and n_j is the number of members of class i .

For performance analysis, this system is tested by using 200 web pages from the technology domain, sport domain and

hazard domain. The experimental result of the system is shown in Table 6.

Table 6: Experimental Result of the System

Domain Name	Precision (%)	
	Keyword-based Web Page Clustering Result	Semantic-based Web Page Clustering Result
Technology Domain	88.5%	95.7%
Sport Domain	88.7%	96.3%
Hazard Domain	87.5%	98.5%

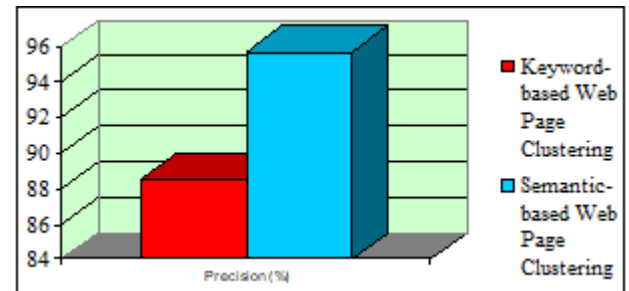


Figure 2: Comparison Result about “Technology Domain”

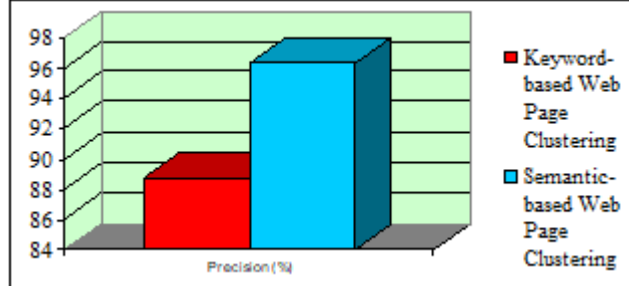


Figure 3: Comparison Result about “Sport Domain”

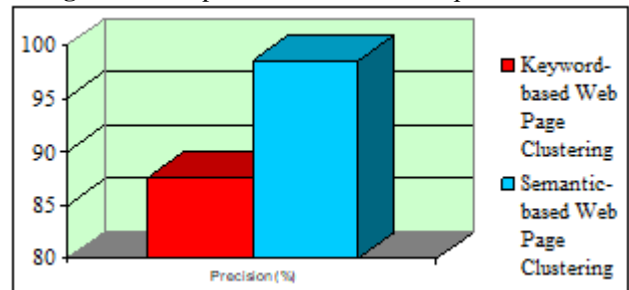


Figure 4: Comparison Result about “Hazard Domain”

Figure 2, 3 and 4 shows the performance comparison result about “Technology Domain”, “Sport Domain” and “Hazard Domain”.

9. Conclusion

In conclusion, this system points out the better performance between the keyword-based and semantic-based web page clustering system. Semantic-based web page clustering system captures the content of a web page more precisely when the senses are used. This system can support for the information retrieval and cluster-based browsing application. Moreover, this system proposed the enhanced Agglomerative hierarchical clustering algorithm that allows the user can choose the desired cluster number. Moreover, the user can view the content in the cluster as the hierarchical level. So, the effectiveness of semantic-based web page clustering

system is more than keyword-based web page clustering system.

References

- [1] M. Gupta and A. Rajavat, "Comparison of Algorithms for Document Clustering", Sixth International Conference on Computational Intelligence and Communication Networks, IEEE, pp. 541-545, 2014.
- [2] I. Alagha and R. Nafee, "An Efficient Approach for Semantically Enhanced Document Clustering by using Wikipedia Link Structure", International Journal of Artificial Intelligence & Applications (IJAIA), vol. 5, no. 6, 2014.
- [3] S. Romeo, A. Tagarelli and D. Ienco, "Semantic-Based Multilingual Document Clustering via Tensor Modeling", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), October, 2014.
- [4] G. Tang and Y. Xia, "Document Representation with Statistical Word Senses in Cross-Lingual Document Clustering", International Journal of Pattern Recognition and Artificial Intelligence, vol. 29, no. 2, 2015.
- [5] C. Guan, K. K. F. Yuen and F. Coenen, "Towards An Intuitionistic Fuzzy Agglomerative Hierarchical Clustering Algorithm for Music Recommendation in Folksonomy", IEEE International Conference on Systems, Man, and Cybernetics, pp. 2039-2042, 2015.
- [6] L. Fasheng and L. Xiong, "Survey on Text Clustering Algorithm", IEEE, pp. 901-904, 2011.
- [7] Bing Liu, Web Data Mining, Department of Computer Science, University of Illinois, USA, 2007.
- [8] R. Navigli, "Word Sense Disambiguation: A Survey", ACM Computing Surveys, vol. 41, no. 2, 2009.
- [9] J. Whissel, "Information Retrieval using Lucene and WordNet", Master of Science, The Graduate Faculty of the University of Akron, December, 2009.
- [10] D. Hladek, J. Stas, M. Pleva and S. Ondas, "Survey of the Word Sense Disambiguation and Challenges for the Slovak Language", 17th IEEE International Symposium on Computational Intelligence and Informatics, November, 2016.
- [11] M. Kuhkan, "A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm", International Journal of Computer Engineering and Information Technology, vol. 8, no. 6, pp. 90-95, 2016.
- [12] R. Pandit and S. K. Naskar, "A Memory Based Approach to Word Sense Disambiguation in Bengali Using K-NN Method", 2nd International Conference on Recent Trends in Information Systems (ReTIS), pp. 383-386, IEEE, 2015.