Document Clustering with Explicit Semantic Analysis (ESA)

Muhammad Adnan[1], Muhammad Rafi[2]

[1]*Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST) Karachi Pakistan*

[2]*National University of Computer & Emerging Sciences NU-FAST Karachi Pakistan*

[1]adnansiddiq@outlook.com

[2]rafi.muhammad@gmail.com

Abstract: **Document clustering recently became a vital approach as numbers of documents on web and on proprietary repositories are increased in unprecedented manner. The documents that are written in human language generally contain some context and usage of words mainly dependent upon the same context; recently researchers have attempted to enrich document representation via external knowledge base. This can facilitate the contextual information in the clustering process. An enrichment process with explicit content analysis using Wikipedia as knowledge base has been proposed. The approach is distinct in the sense that only the conceptual words from a document were used and their frequency to embed the contextual information. Hence, the approach does not over enrich the documents. A vector based representation, with cosine similarity and agglomerative hierarchical clustering is used to perform actual document clustering. The proposed method was compared with existing relevant approaches on NEWS20 dataset, with evaluation measure for clustering including F-Score, Entropy and Purity.**

Keywords **— Explicit semantic analysis, Similarity measure, Document clustering, Document representation, Hierarchical clustering**

I. INTRODUCTION

Clustering is an unsupervised machine learning technique that can be applied in different areas of computational problems. Clustering process understands the patterns to implicitly create clusters. Clustering can be defined as, the organization of objects in such a way that the objects are grouped together on the basis of similarity between them. Thus, the objective of clustering is to focus on the intrinsic grouping for a set of untagged information. The purpose of Clustering of documents is to figure out a pattern between collections of untagged documents. The document clustering process involves representation of documents, counting similarity measure among the pair of documents and nature of clustering algorithm applied for the clustering process. In order to cluster documents, the clustering process has to learn from patterns of documents and cluster them on the basis of similarity measure. For document clustering, the group of documents are subdivided into smaller and manageable collections called clusters. This subdivision of groups of documents into clusters implicitly analyse the natural grouping of characteristics among words of documents.

Traditional document clustering algorithms depend upon words, phrases and sequences from the documents to find the similarity measure between documents to cluster them. To decide about the relatedness among documents, they simply apply feature extraction techniques that depend on feature counting and frequency distribution of those features in documents. Therefore, these approaches are unable to grab the meaning behind the text; they simply perform clustering of documents independent of the context. A document's context in natural languages mostly depends on conveying the information by selection of word sequence.

The paper in hand proposes a probabilistic term weight model that uses ESA to cluster documents and to describe the computed result values. The basic idea that makes ESA more robust than other traditional document clustering techniques is to represent and compare texts as vectors in a high dimensional concept space [1]. The association between the text and the respective concepts is quantified by the entries in the concept vector. For the purpose of calculating such association and related values, each natural language concept is represented by an index document of document corpus. The articles from Wikipedia are used as index documents since Wikipedia is one of the largest databases for natural language concepts and it contain many articles, while each article is focused on specific topic. For experimental purpose, individual document from the document dataset has been read, retrieve the concepts of the document and then find the weight of repetition frequency of those concepts using Wikipedia document corpus. The concepts with most frequency of repetition are selected for the second step of clustering process. The experiments are performed on NEWS20 dataset. The proposed method was compare with existing relevant approach on NEWS20 dataset, with evaluation measure for clustering including F-Score, Entropy and Purity.

II. LITERATURE REVIEW

As discussed in introduction, clustering is an unsupervised machine learning technique that can be applied in different areas of computational problems. For document clustering, objects that have to be clustered are in the form of actual text documents. Clustering of documents is essential for information retrieval from large document databases. A substantial number of document accumulations like Internet

and propriety documents are required to be grouped on basis of likeness measure, for simple and fast recovery. The document clustering is used to group similar documents in to a single cluster, where the similarity can be on basis of type of document or contents of the document. Many documents with related content and type can be clustered in to single cluster. The important task related to document clustering is to understand the meaning of text from given dataset, the information retrieved by this process is used to find out number of classes of such groups exist in the collection. Documents clustering find the characteristic aggregating around documents, in such way that, the document in a cluster are related on basis of certain similarity measure to all other documents in same cluster. In the same way, documents from one cluster do not relate to the documents on other clusters. Therefore, clustering is a good approach for computation of searching as discussed in [2]. It provides functionality for combination of related results [3] and furthermore provides connections between the outcomes of clustering results [4].

The clustering process is grouped into two major categories particularly (i) Partition vs. Overlapping and (ii) Hierarchical vs. Flat [5], as it has been reported in [5]. The difference between Agglomerative hierarchical clustering and Partitioned based clustering is explained in [6]. The Agglomerative hierarchical clustering (AHC) is a bottom up approach for clustering where each document is initially treated as an individual cluster, while after processing the documents, the pair of documents that are merged as clusters on basis of some defined similarity measure are obtained. Another major type of clustering is Partitioned based clustering. In Partitioned based clustering, one level partitioning of documents is created as stated in [3, 6]. This type of clustering process uses K-mean algorithm to create k-documents which are further used as base level documents for the next step of clustering process. In next step, documents that are similar on the basis of some similarity measure are merged together and the base level is recalculated using the results of clustering process. This step is repeated until maximum numbers of clusters are produced from this process [7].

There are many approaches that are suggested for document clustering having pros and cons of their own. There are two new algorithms that claim to grab context of document more accurately from the conventional methodologies. Clustering dependent upon Frequent Word Sequence (CFWS) and Clustering dependent upon Frequent Word Meaning Sequence (CFWMS) are proposed in [8]. These methodologies keep up a list of special words that holds word that are frequently utilized within the documents. Take an example of database D comprise of three documents da, db and dc. Each document holds different words and the database D has unique words from all documents. Then frequent word sequence is obtained by generating a 2-word sequence for each document. A threshold for occurrence of a word sequence is maintained to control the list of words. Suppose if a word's occurrence is less than threshold value, the word will be dropped from the list of words. After completion of filtering process of unnecessary words, a final dictionary is obtained as D' = {da', db', dc'}. This filtering of unnecessary word sequences improves the results of clustering process. At last, a final dictionary is produced which contain the list of words greater than or equals to the threshold value. In Clustering dependent upon Frequent Word Sequence (CFWS), the documents that support the same continuous word succession are acknowledged to be cluster candidates. The threshold for utilized sequence of words is supposed to be between 5-15% words. K-mismatch concept is used to merge documents. The next algorithm, named 'Clustering dependent upon Frequent Word Meaning Sequence' (CFWMS) [8], uses sequence of meaning for frequent word to obtain relatedness among documents. Similar approach of document clustering based on topic maps was based on the topic map representation of the documents [9] in which the document is transformed into a compact form, then a similarity measure is proposed based upon the inferred information through topic maps data.

The bag of words (BOW) representation assures that document retrieval can be improved by breaking lengthy documents into shorter passages [10]. The extraction of features similar to words, phrases and sequences of words are used in traditional document clustering approaches [11-12]. Using these features, similarity among documents is calculated; however, this approach does not guarantee the similarity of documents on basis of exact contextual thought. The bag of word model is effective, but only when the connection among the documents is required. For instance, synonymy (e.g. the expression "chemistry" and the phrase "chemical sciences" are semantically identical as defined by WorldNet) between documents is disregarded, it decreases the effectiveness of requisitions utilizing a standard text document.

The method proposed in this paper is based on Explicit Semantic Analysis (ESA) for document clustering. ESA is a variation of the generalized vector space model [13]. In traditional general vector space model (GVSM), a document is represented as a vector, based on a weighted combination of n-dimensional term vectors [14]. ESA resolves the mismatching of vocabulary problem [15] that existed in the traditional bag-of-word (BOW) model. ESA uses the weighted mixture of a predetermined set of natural concepts that exists in the form of Wikipedia. It index documents with respect to Wikipedia article and indicate the relation of a word or the whole document with specific Wikipedia article [16-17]. Due to association with similar Wikipedia articles, in the ESA model, two documents can be semantically related in spite of not having any word in common [1]. After generating concept vector, each word appearing in the Wikipedia corpus can be seen as triggering one of the concepts it points, with the attached weight representing the degree of association between that word and the concept, which later on used to

make clusters of documents. The hierarchical agglomerative clustering is selected as the clustering approach for experiments of this study. Hierarchical agglomerative clustering is a bottom up approach for clustering, which means each individual document is treated as a separate cluster at the startup but eventually pair of documents are combined of form clusters on the basis of similarity between them [5,18].

III. Research Objective

The objectives are mentioned below:

- The purpose of this research is to find out the impact of ESA for document clustering.
- Finding results after conducting experiments on test data sets and comparing them with Clustering dependent upon Frequent Word Sequence (CFWS) algorithm.
- Structure the conclusion dependents upon the outcomes acquired and propose that for the given sample data for which the calculations were performed ideally gave best accuracy for clustering.

IV. Research Methodology

The below mentioned methodologies are followed for this research.

A. Experimental Research

Experimental Research is the research for which trials are carried out to achieve a result. For this research, the dataset of NEWS20 to text my approach has been used. A simple computer application is created to test the process. After parsing documents, some pre-processing tasks are performed on documents. This process includes removal of stop words from documents and performing lemmatization on documents.

B. Quantitative Research

For this research, the proposed approach of using phenomena of Explicit Semantic Analysis has been implemented for document clustering. The similarity measures of documents are calculated and finally, clusters are created using Hierarchical Agglomerative Clustering algorithm. The quantitative results are calculated for a selected dataset of NEWS20.

V. Experiment

A. Experimental Setup

The method that has been proposed in this paper is tested by performing experiments. A setup is created for testing proposed method. The experiment is conducted on Acer Aspire 5741 machine with a core i3 Intel Processor, 4GB of RAM and 320GB of hard drive. As ESA approach is used for clustering, it required access to document corpus of Wikipedia. Therefore internet connection is required for the test setup. The internet connection of 1mb is used for the experimental process. The Stanford CoreNLP library for lemmatization is used for the process of clustering. The proposed approach is implemented by a java application that was created as a software requirement for experimental setup.

B. Data Set

Sample datasets from NEWS20 is used for experimental purpose. The dataset of 20Newsgroups is an accumulation of 20,000 newsgroup papers, apportioned (about) equally over 20 separate newsgroups. The 20 newsgroups data collection is well-known dataset for experiments in text processing applications particularly document clustering.

Table 1. Details of Selected Documents

Dataset	Number of documents	Number of Classes
N20a	50	5
N20b	100	8
N20c	200	21
N20d	400	32

A dataset of 50 documents, 100, 200 and a 400 is used for testing as mentioned in table 1. The experiments were conducted on these datasets and the results were obtained for the comparison of proposed approach with one of the best traditional approaches used for document clustering. These datasets are popular among text mining community because of their free availability. They are widely used for document text processing experiments.

C. Document Processing

Initially, a document was selected from the dataset for processing. Then, for the purpose of filtering unnecessary words from document, stop words were removed from the document. Stop words are the words that do not represent the theme of the text and must be filtering out prior to processing text documents for better results. After removing the stop words, lemmatization process was applied to make the document representation in proper form of word which means the inflected forms of words are removed and therefore, they can be analysed as unique single item. After the process, a list of distinct words was obtained, in such a way that if a word is repeating in the list, its frequency is increased each time. The content was searched for each word in list on Wikipedia and getting wiki document for that word. Once the related documents were obtained for that word, the process of stop word removal and lemmatization were performed on the wiki document. The wiki document ended up as a list of words. Each concept from the list of concepts of the document is checked for the list of words for Wikipedia document, if the word exists in the list, its frequency was increased every time. At last, the document ended up with list of concepts along with their frequency of repetition. A threshold was defined for the value of repetition of concepts, to restrict the concepts that are repeating frequently remained in the list, while others are filtered out by the process. After performing these operations on all documents of the dataset, the Hierarchical Agglomerative Clustering was performed on the resulted documents. The Agglomerative hierarchical clustering (AHC) is a bottom up approach in which each document is initially

treated as an individual cluster, after completing the clustering process, the pair of documents that are merged to be as an clusters on basis of some defined similarity measure was obtained.

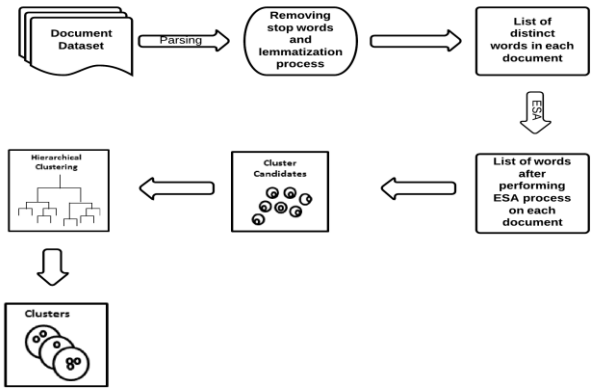

Fig. (1). Document Clustering using ESA

As illustrated in fig: 1, a document is chosen from the dataset for processing. The document was then passed through different pre-processes for removal of stop words and lemmatization for making the words appears in same form. Through this process, words in the documents got ready to be processed with ESA. After performing ESA on the documents, the clustering candidates were ready for clustering process. Hierarchical Agglomerative Clustering was performed on the clustering candidates by whom final clusters of the documents were obtained.

D. Evaluation

The method proposed is effective for clustering and it is justified by the use of quality measures for clustering like F-Measure, purity and entropy.

E. F-Measure

F - Measure is used commonly for estimating the efficiency of a clustering algorithm or document classification algorithm. F-measure technique can be used to evaluate the effectiveness of the process that is selected for the document clustering. F-measure use mixture of *precision* and *recall* properties of clusters. Let's assume, a base for document D which is comprised of N number of documents, the clustering process will produce C={Ca , Cb,… Ck} cluster for document base D. Actual clusters that document base have are C*={C*a, C* b,.. C* c}. *Recall* of cluster *j* with respect to class *i*, *rec (i,j)* will be given by | Cj ∩C* i| / | C* i|, and the *precision* of cluster *j* with respect to class *i*, *prec(i,j)* will be given by | Cj ∩C*i| / | Cj|. F-Measure is the combination of both precision and recall with the following formula:

$$F(i,j) = \frac{2 * prec(i,j) * rec(i,j)}{prec(i,j) + rec(i,j)}$$

The overall quality of cluster C is given by the following formula for F-measure:

$$F = \sum_{i=1}^{k} \frac{|C_i^*|}{N} * \max_{i=1,2,..k} \{F(i,j)\}$$

F. Purity

Purity could be characterized as the maximal accuracy esteem for each class j. The purity of cluster shows the rate of the prevailing class parts in the given cluster. Overall purity for cluster C could be process as the weighted normal purity by the following equation:

$$Purity = \sum_{j=1}^{k} \frac{|C_j|}{N} * \max_{i=1,2,..k} \{Prec(i,j)\}$$

G. Entropy

It is the measurement of homogenous property for each cluster. The entropy is calculated by this formula:

$$Ei = -\sum_{j \in L} precision(i,j) * \log(precision(i,j))$$

The total entropy for a set of cluster is calculated as the sum of entropies for each cluster weighted by the size of each cluster:

$$Entropy_c = \sum_{i \in c} ((\frac{Ni}{N}) * Ei)$$

The purity is required to be maximized while the entropy of the clusters should be minimized to achieve high quality of clustering.

VI. RESULTS

The results of proposed approach are compared with a recently proposed algorithm CFWS. The F-measure on four selected datasets is calculated. For CFWS, the algorithm was executed several times on all datasets. The average of best three scores is reported for the purpose of comparison. It is evident that CFWS performs equally for all datasets. Similarly, the proposed method was executed several times on the datasets and average result of best three scores is considered for comparison. Results of F-measure are presented in the table 2 below.

Table 2. Results of F-Measure For Cfws And Esa

Dataset	CFWS	ESA
N20a	0.646	0.645
N20b	0.641	0.642
N20c	0.645	0.646
N20d	0.644	0.643

These results can be described graphically in figure 2 as:

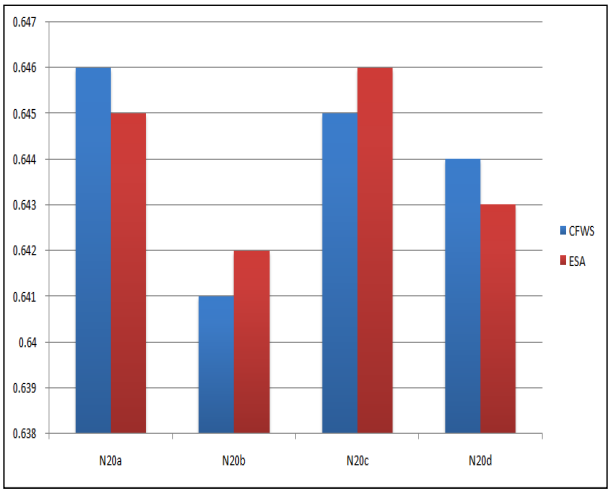

Fig. (2). Graphical Representation of Results for F-measure

Purity is also evaluated for this experiment. The purity value of the proposed approach is shown in table 3 below. The purity results for the ESA approach for document clustering is worthy and comparative with the traditional approaches. The high number of purity is indicating the effectiveness of clustering process.

Table 3. Results of Purity For Cfws And Esa

Dataset	CFWS	ESA
N20a	0.738	0.739
N20b	0.737	0.736
N20c	0.738	0.737
N20d	0.735	0.736

These results can be described graphically in figure 3 as:

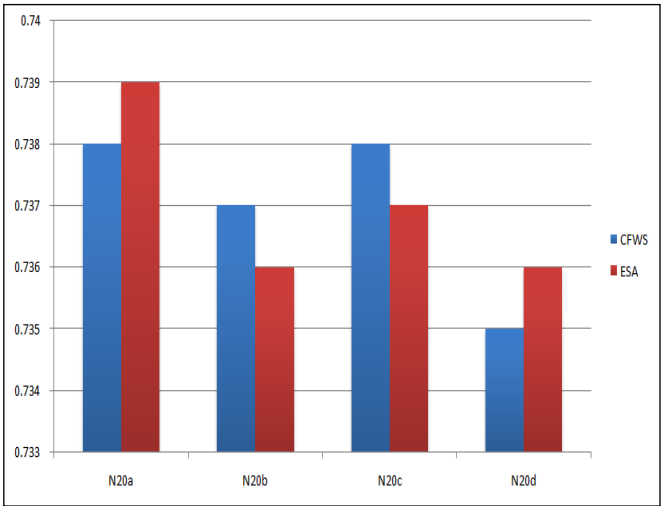

Fig. (3). Graphical Representation of Results for Purity

Similarly, the entropy is also calculated. The results comparison for the entropy between the proposed approach and CFWS are described in table 4. The entropy results for the ESA approach for document clustering is comparative with the traditional approaches. The low results of entropy are indicating the effectiveness of clustering process.

Table 4. Results of entropy for cfws and esa

Dataset	CFWS	ESA
N20a	0.21	0.18
N20b	0.19	0.2
N20c	0.22	0.19
N20d	0.22	0.21

These results can be described graphically in figure 4 as:

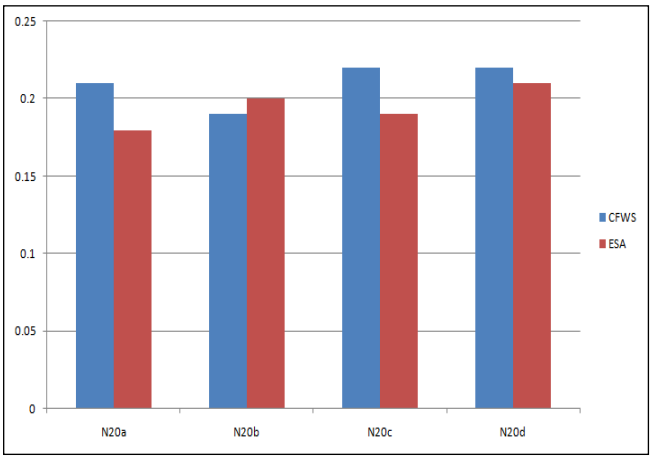

Fig. (4). Graphical Representation of Results for Entropy

VII. Conclusion

The main purpose of document clustering is to cluster the documents on the basis of certain similarity measure. In this paper, the approach of using ESA for the purpose of document clustering is proposed. In this approach, documents are clustered by actual concepts that represent them. This approach has truly helped to grab the theme of the documents and therefore the grouping of documents on the basis of similarity measure becomes more accurate. The similarity between documents is calculated by using list of concepts that present in the documents. The actual lists of concepts are grabbed by using ESA approach. This provides lexical semantic for similarity computation. The hierarchical agglomerative clustering is used to produce final clusters. The proposed algorithm filters the document and captures the actual concepts that reflect the theme of the documents; hence, it optimizes the accuracy of the clusters. The proposed approach is compared with the results of clustering the same documents with Frequent Word Sequence (CFWS) algorithm by f-measure, purity and entropy. The results of experiment reflect that the approach is scoring comparatively equal results.

VIII. FUTURE WORK

There are several ways by which this approach could be improved. In this study, only Wikipedia document corpus was used to retrieve concepts of the document, but in future, more than one document corpus can be used to extract more filtered document concepts for clustering.

REFERENCES

[1] P. Sorg and P. Cimiano, "An Experimental Comparison of Explicit Semantic Analysis Implementations for Cross-Language Retrieval", Institute AIFB, University of Karlsruhe & Web Information Systems Group, Delft University of Technology.

[2] A. Campi and S. Ronchi, "The Role of Clustering in Search Computing", In *Proceedings of 20th International Workshop on Databases and Expert Systems Application*, 2009, pp: 432- 436.

[3] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections," In *Proceedings of Fifteenth Annual International ACM SIGIR Conference on Research and development in information retrieval*, 1992, pp: 318-329.

[4] M. A. Hearst and J. O. Pedersen, "Re-examining the cluster hypothesis: scatter/gather on retrieval results," In *Proceedings of 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 1996, pp: 74-84.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clutering: a review," *ACM Computing Survey*, vol. 31, no. 3, pp. 264-323, 1999.

[6] I. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis: John Wiley & Sons, 1990.

[7] M. Steianbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," In *Proceedings of KDD-Workshop on Text Mining*, 2000.

[8] Y. Li, S. M. Chung, and J. D. Holt, "Text document clustering based on frequent word meaning sequences", *Data and Knowledge Engineering*, vol. 64, no. 1, pp: 381-404, January 2008.

[9] M. Rafi, S. M. Shaikh and A. Farooq. "Document Clustering based on Topic Maps". *International Journal of Computer Applications*, vol. 12, no. 1, pp: 32–36, December 2010.

[10] X. Liu, and W. B. Croft, "Passage retrieval based on language models". In *Proceedings of the eleventh International Conference on Information and Knowledge Management*. 2002, pp: 375–382.

[11] K. M. Hammouda, and M. S. Kamel,"Efficient Phrase-Based Document Indexing for Web Document Clustering," *IEEE Transaction on Knowledge and Data Engineering*, vol. 16, no. 10, pp: 1279-1296, 2004.

[12] C. Hung and D. Xiaotie, "Efficient Phrase-Based Document Similarity for Clustering," *IEEE Transaction on Knowledge and Data Engineering*, vol. 20, no. 9, pp: 1217-1229, September 2008.

[13] T. Gottron, M. Anderka, and B. Stein "Insights into Explicit Semantic Analysis" In *Proceedings of 20th ACM international conference on Information and knowledge management (CIKM '11)*, 2011, pp: 1961-1964.

[14] S. K. M. Wong, W. Ziarko, and P. C. N. Wong, "Generalized vector spaces model in information retrieval". In *Proceedings of the 8th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 85)*, 1985, pp: 18–25.

[15] G. Furnas, T. Landauer, L. Gomez and S. Dumais. "The vocabulary problem in human-system communication", *Communications of the ACM*, vol. 30, no. 1, pp: 964–971, 1987.

[16] E. Gabrilovich and S. Markovitch, "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis", In *Proceedings of 20th international joint conference on Artifical intelligence (IJCAI'07)*, 2007, pp: 1606-1611

[17] O. Egozi, S. Markovitch and E. Gabrilovich. "Concept-Based Information Retrieval using Explicit Semantic Analysis", *ACM Transactions on Information Systems*, vol.29, no. 2, 2011.

[18] P. Zezula, G. Amato, V. Dohnal and M. Batko, Similarity Search-The Metric Space Approach: Springer, 2006.