Application of the SpecHybrid Algorithm to Text Document Clustering Problem

Zekeriya Uykan[1] and Murat C. Ganiz[2]
[1]Electronics and Communications Engineering Dept.
[2] Computer Engineering Dept.
Doğuş University
Acıbadem, Kadıköy, 34722, Istanbul, Turkey
{zuykan, mcganiz}@dogus.edu.tr

Abstract—**In this paper, we present a relaxed version of the SpecHybrid Algorithm originally proposed for wireless cellular systems, and apply it to text document clustering problem. We conduct several experiments on two different datasets; a widely used benchmark dataset in English, and a Turkish textual dataset commonly used in text classification. Our results show that the proposed algorithm gives superior performance in text document clustering as compared to the standard *k*-means algorithm for any number of clusters while giving a comparable or better performance as compared to the standard EM algorithm for relatively large number of clusters depending on the similarity matrices used.**

Keywords- textual data mining, document clustering, Turkish document clustering, max cut, spectral clustering.

I. INTRODUCTION

Large amounts of textual data is accumulated in organizations with the rapid developments in information technology and internet usage. Merrill Lynch estimates that more than 85 percent of all business information consist of unstructured data which exists in the form of e-mails, memos, notes from call centers and support operations, news, user groups, chats, reports, letters, surveys, white papers, marketing material, research, presentations and web pages [1]. This mainly textual content includes valuable information that can be used for improving business communications, enhancing customer satisfaction and maintaining a competitive edge. Textual data mining (a.k.a. text mining) is the process of extracting this valuable information from large amount of textual data. Textual data mining has several application areas such as customer relationship management, analysis of customer feedbacks and comments, email management, knowledge management in banks, detection of crime patterns and connections [2]. Given the amount of textual data accumulated in organizations, document clustering is one of the most important techniques for organizing documents in an unsupervised manner [3]. In document clustering documents are automatically grouped into predefined number of clusters based on their similarity. In general similarity is determined in vector space either using a distance metric such as Euclidean distance or Manhattan distance or a similarity metric such as Cosine similarity. There are various types of document clustering methods such as partitioning, hierarchical, density-based, grid-based and model-based methods. For more information, a survey on document clustering algorithms can be found in [4].

In this paper, we introduce a relaxed version of the SpecHybrid Algorithm which is originally proposed for wireless cellular systems in [5], and apply it to text document clustering problem. To be precise, the SpecHybrid algorithm is originally proposed for finding (near) optimum solution for the channel/frequency allocation problem in wireless cellular systems, which is known to be NP-complete (**N**ondeterministic **P**olynomial time - complete) (see e.g. [6], [18], among others). The algorithm presented in this paper being adopted from the SpecHybrid falls in the area of clustering by graph partitioning. Specifically, the SpecHybrid algorithm [5] can be viewed as a graph *maxCut* algorithm. The reasons why we adopt the SpecHybrid algorithm for text clustering include its speed of convergence, and its ability to produce high quality clusters.

We conduct several experiments on two different datasets; one of them is a widely used benchmark dataset in English, and another one is a Turkish textual dataset commonly used in text classification. Our results show that the proposed algorithm gives remarkably superior performance in text document clustering as compared to the standard *k*-means for any number of clusters while giving a relatively or remarkably better performance as compared to the standard EM algorithm for relatively large number of clusters.

The rest of the paper is organized as follows: Section II gives the formulation of the problem. The adopted spectral based solution for text clustering is presented in section III. The experiments setup and results are presented in section IV and V, respectively, followed by the conclusions and remarks on future work in section VI.

II. FORMULATION

As a high level definition, document clustering can be defined as partition of N documents into a predetermined number of L subsets so that documents assigned to each subset are more similar to each other than the documents assigned to different subsets (e.g. [7]). In general clustering algorithms first generate a document-by-document similarity matrix by using a similarity metric and operate on this matrix. This is also called distance matrix if a distance metric is used instead of a similarity metric. Among these measures, the Euclidean distance is one of the most commonly used similarity metric in clustering problems. It is also the default distance metric of the

TABLE I

SPECHYBRID ALGORITM IN [1] ADOPTED FOR TEXT MINING FOR $L = 2^q$

Phase 1: Centrally find a rough estimate.

1. Establish the dissimilarity (distance) matrix.
2. Repeat for *n=1:L*
   - Determine the nodes of $\left(N/2^L\right)$ to be spectrally clustered for *L=2*.
   - Perform the clustering with respect to the sign of the maximum eigenvector of the corresponding $\left(N/2^L \times N/2^L\right)$ dimensional unnormalized Laplacian matrix in (9).
3. Finally, determine the *L* groups of documents according to the signs of the eigenvectors.

Phase 2: Distributively tune the solution by the standard "minimum-interference-channel allocation" algorithm asynchronously (e.g. by GADIA [8]).

4. Take the result of phase 1 as initial condition, and repeat until there is no change in document-cluster assignments after epoch.
   - Repeat for *n=1:L*
     - Determine the cluster *j* in which the document *n* has the minimum sum of intra-cluster distances (weights).
     - Assign document *n* to cluster *j*.

data mining tool we use in our experiments. Because of these reasons and due to the time constraints, in this study, we start with the Euclidean metric in our experiments. In our near-future studies, we will include the other metrics like cosine similarity, which is reported to perform better than the Euclidean distance (see e.g. [9], [12]).

In document clustering, distance between two documents which are represented by their term vectors can be calculated using the formula in eq.(1). In here *k* is the dictionary size which is the number of distinct terms in our dataset. If we use term frequencies (tf), then x_{ik} is the frequency of k^{th} term in document *i* and x_{jk} is the frequency of same term in document *j*.

$$w_{ij} = \sqrt{\sum_k \left(x_{ik} - x_{jk}\right)^2} \tag{1}$$

Using Euclidean distance we can form a distance matrix. Let's define the distance matrix as follows

$$\mathbf{W} = \left[w_{ij}\right]_{N \times N} \tag{2}$$

where w_{ij} represents the distance between the document *i* and *j* (and $w_{ij} = w_{ji} > 0$, and $w_{ii} = 0$).

Let the number documents be *N* and the number of clusters be *L* (where typically *N>> L*). So, we cluster *N* documents into *L* groups. We aim to minimize the sum of intra-cluster weights, denoted as I_{tot}, which is given by

$$I_{tot} = \sum_{s=1}^{L} I_S = \sum_{s=1}^{L} \sum_{j \in C_S}^{N_S} \sum_{\substack{i \in C_S \\ (i \neq j)}}^{N_S} w_{ij}, \qquad s = 1, \cdots, L \tag{3}$$

where C_S represents the set of text documents assigned to cluster *s*; N_S is the number of documents in cluster *s*, and I_S is the sum of the intra-cluster weights in cluster *s*, and $\sum_{s=1}^{L} N_S = N$. From (3), we formulate the text document clustering problem as determining the sets C_S of documents (*s=1, ..., L*) which minimizes the sum of intra-cluster weights I_{tot} in (3), i.e.:

$$\min_{\text{determine } C_S,\ (s=1,\cdots,L)} I_{tot}^{ntw} \tag{4}$$

and this is known to be an NP-complete problem (e.g.[6], [18]).

III. APPROACH

In [5], a new spectral based algorithm, called SpecHybrid, is proposed for frequency/channel allocation in wireless cellular systems for $L = 2^q$, where *q* is a positive integer. The analysis for *L=2* case is presented in [5]. In this study, we relax the condition on *L* and adopt the SpecHybrid in order to apply to text document clustering problem for arbitrary *L*.

The volume (i.e., entrywise 1-norm) of the distance matrix $\mathbf{W}$ is equal to

$$vol(\mathbf{W}) = \|\mathbf{W}\|_1 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \tag{5}$$

Then, considering the grouping of the *N* documents into *L* clusters C_1 to C_L, we write

$$vol(\mathbf{W}) = \sum_{l=1}^{L} \left\{ \sum_{i \in C_l}^{N_l} \sum_{j \in C_l}^{N_l} w_{ij} + \sum_{i \in C_l}^{N_l} \sum_{j \in \overline{C}_l}^{N-N_l} w_{ij} \right\} \tag{6}$$

where C_l and $\overline{C}_l$ represent the cluster *l* and all other clusters, respectively. Eq.(6) can be written as

$$vol(\mathbf{W}) = \text{constant} = \sum_{l=1}^{L} vol\left(C_l\right) + \sum_{l=1}^{L} cut\left(C_l, \overline{C}_l\right) \tag{7}$$

where $vol\left(C_l\right) = \sum_{i \in C_l}^{N_l} \sum_{j \in C_l}^{N_l} w_{ij}$ is the sum of *intra-cluster weights* for cluster *l*, and $cut\left(C_l, \overline{C}_l\right) = \sum_{i \in C_l}^{N_l} \sum_{j \in \overline{C}_l}^{N-N_l} w_{ij}$ represents

the sum of the *inter-cluster weights* between cluster l and all other clusters. From (7)

$$\min_{\{C_l\}_{l=1}^{L}} \left\{ \sum_{l=1}^{L} vol(C_l) \right\} \equiv \max_{\{C_l\}_{l=1}^{L}} \left\{ \sum_{l=1}^{L} cut(C_l, \overline{C}_l) \right\} \quad \textbf{(8)}$$

From eq.(8), minimizing the total intra-cluster weights is equal to well-known weighted *maxCut* problem in graph theory (e.g. [10]). The unnormalized Laplacian matrix (e.g. [10], [11]) is given as

$$\mathbf{U} = \mathbf{D} - \mathbf{W} \quad \textbf{(9)}$$

where diagonal matrix $\mathbf{D} = [d_{mn}] = \begin{cases} \sum_{j=1,\,(j\neq i)}^{N} w_{ij}, & \text{if } m = n \\ 0, & \text{otherwise} \end{cases}$.

It's known that (see e.g. [11]), the *maxCut* problem can be formulized for general L as follows

$$\text{maxCut}_{C_1,\ldots,C_k}\{\mathbf{W}\} = \max\{Tr(\mathbf{H}^T \mathbf{U} \mathbf{H})\} \quad \textbf{(10)}$$

where $Tr(\cdot)$ represents trace operation, and where

$$\mathbf{H} = [h_{ij}] = \begin{cases} 1/\sqrt{L}, & \text{if } i\text{'th document is in } C_j \\ 0, & \text{otherwise} \end{cases} \quad \textbf{(11)}$$

$$i = 1,\cdots,N, \qquad j = 1,\cdots,L$$

TABLE II
RELAXED SPECHYBRID ALGORITM IN [1] FOR ARBITRARY L.

Phase 1: Central rough estimate
Run the standard spectral clustering for L clusters as follows:

- Find the greatest L eigenvectors of the Laplacian matrix $\mathbf{U}_{N\times N}$ in (9).
- Establish a matrix $\mathbf{Y}_{N\times L}$ whose columns are the greatest L eigenvectors.
- Run *k*-means algorithm to the row vectors $\{\mathbf{y}_i\}_{i=1}^{N} \in \Re^L$ of $\mathbf{Y}_{N\times L}$, and determine the clusters $C_1,\cdots,C_L$.

Phase 2: Distributively tuned solution
Take the result of phase 1 as initial condition, and run the distributed algorithm summarized in phase 2 of Table I.

Eq.(10) is standard trace maximization problem, and relaxing the discrete-solution constraint of (11), it is well-known that the optimum solution is given by choosing **H** as the matrix which contains the greatest k eigenvectors of matrix **U** as columns (see e.g.[11]). Thus, a relaxation of the SpecHybrid algorithm [5] adopted here for text document clustering is given in Table I and Table II for $L = 2^q$ (q positive integer) and for arbitrary L, respectively.

IV. EXPERIMENT SETUP

In order to demonstrate the efficiency of our algorithm, we run it on two different textual datasets in different languages and compare the evaluation results with two of the most well known clustering algorithms; *k*-means and Expectation Maximization (EM).

The first dataset used is the mini-newsgroups dataset which is a subset of well-known 20-Newsgroups dataset [1] and it includes 100 articles from each group. This is a labeled (supervised) dataset which consist of 20 classes. 1150haber dataset consist of Turkish newspaper articles in five categories namely economy, magazine, health, political, and sports. There are 230 documents in each category [13].

We apply several preprocessing methods to these datasets. For both datasets, infrequent terms that exist in less than 3 documents are filtered. We apply stop word filtering and lovins stemmer algorithm [14] to Mini-newsgroups dataset resulting dictionary size of 10621 terms. On the other hand, for 1150haber dataset we apply stop word filtering with a Turkish stop word list adopted from [15] and zemberek stemmer [16]. Dictionary of this dataset consist of 5092 terms.

In order to observe the performance of clustering algorithms in different conditions we created three versions of each dataset using two different term weighting schema and reducing the dimensionality of the feature space by feature selection. As the term weighting schema we created two different versions, in the first one terms are represented by their frequencies in the documents. This is called term frequency (tf) weighting. In the second version we use a common normalization which is called tf.idf. This is basically the term frequencies divided by document frequencies of these terms. A third version of each dataset is created by selecting the best 2000 features using a widely used feature selection method: Information Gain (IG).

We use *k*-means and EM algorithms that are implemented in WEKA data mining software [17]. Default similarity metric or distance metric in our experiments is Euclidean distance.

In general, evaluation of clustering algorithms is not easy because of their unsupervised nature and may be subjective. However, both of the datasets we use are supervised datasets meaning that each document is labeled by a category (a.k.a class) label. Using supervised datasets in evaluation of clustering algorithms is common e.g. [9], [2], and allows us to do more objective evaluation by using several different evaluation metrics. We use three different evaluation metrics; purity, entropy and percentage of incorrectly clustered instances. First two are commonly used metrics in document clustering (e.g. [9]). While calculating these metrics, we cluster the dataset into the same number of clusters with the number of categories and label the clusters with category labels in order to apply a wide range of evaluation metrics. If a particular category instances are majority in the cluster, it is labeled with this particular category. After this we now attempt to compare how well the clustering results match the category labels of documents.

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/

The first metric of this kind is Purity. Purity measure calculates the degree of a cluster that contains documents from a single category. An ideal cluster will have instances from only a single category and its purity value will be 1 (e.g. [9]).

$$P(C_i) = \frac{1}{n_i} \max(n_i^c) \quad \textbf{(12)}$$

In the purity formula above, $\max_c(n_i^c)$ is the number of documents belongs to majority category in cluster i.

Second metric is Entropy. Entropy evaluates the quality of a clustering solution by measuring how different categories of documents distributed within each cluster (see e.g. [9]). Entropy of a cluster C_i can be calculated by following formula

$$E(C_i) = -\frac{1}{\log L} \sum_{c=1}^{k} \frac{n_i^c}{n_i} \log(\frac{n_i^c}{n_i}) \quad \textbf{(13)}$$

where L is the total number of category labels in the dataset and n_i^c is the number of documents from category c in cluster C_i . These metrics are specific to each cluster. In order to evaluate the overall clustering solution we calculate average purity and weighted average entropy. Average entropy of overall solution is calculated by summing entropy value of each cluster that is weighted by the number of documents in each cluster. When evaluating a clustering solution by these metrics, in general, quality of clustering is better if entropy is small and purity is high (see e.g. [9]). We should note that these metrics do not require the number of clusters to be the same as the number of classes.

Our third metric is the percentage of incorrectly clustered documents. This metric show how well the clustering solution performs if it is used as a classification system and applied to the same dataset. We use the class labels assigned to clusters by majority vote. All the other documents belong to other classes in a cluster are considered as misclassification. We count all the instances in clusters that belong to minority classes, and divide it to the total number of instances. Combination of these three evaluation metrics provides a broad perspective for evaluating the clustering solution.

V. Experiment Results

The results of our experiments described in section IV are given in the Figures 1 to 4 below where the proposed algorithm is denoted as RSH (Relaxed SpecHybrid).

Our main criterion of clustering quality in this study is the "percentage of incorrectly clustered documents" which is also the default evaluation metric in WEKA data mining software [17]. This metric is indicated as "misclustering %" in the figures. It has an inverse relationship with the quality of the clustering solution: The lower the misclassification rate, the higher the quality. This is also valid for entropy metric. For purity it is the opposite. Higher purity values represent better clusters. However, we observe that purity is not a good metric to measure the clustering quality since a cluster with only a single document will have the maximum purity value of one. We also observe that this is actually the case with *k*-means algorithm which creates clusters with only a one instance and consequently evaluate to very high purity values although other metrics indicates lower clustering quality.

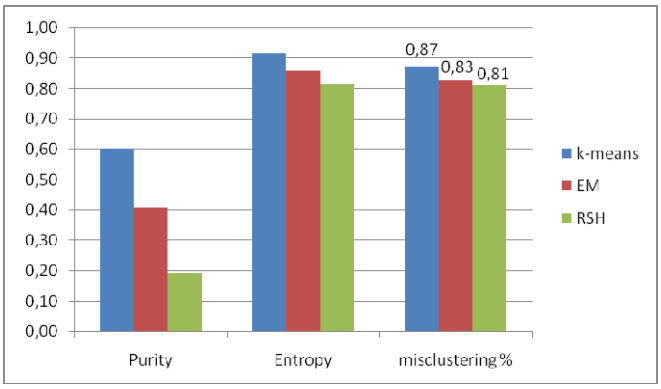

Figure 1. Evaluation of clustering algorithms on mini-newsgroups dataset with tf*idf term weigting.

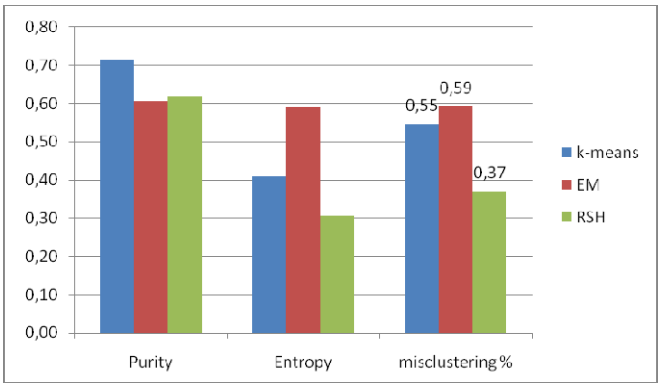

Figure 2. Evaluation of clustering algorithms on mini-newsgroups dataset with tf*idf weighting and feature selection (IG, 2000 terms)

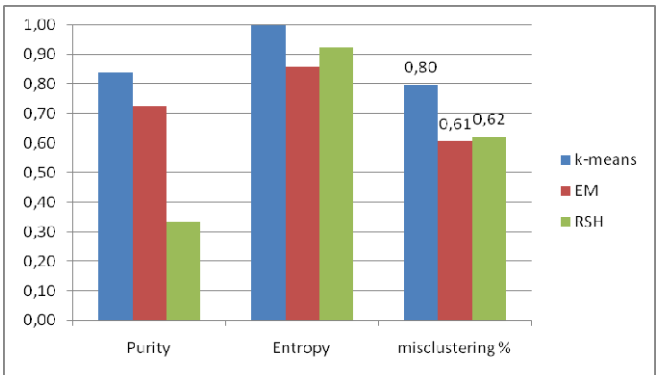

Figure 3. Evaluation of clustering algorithms on 1150haber dataset with tf*idf weighting.

Figures 1 to 4 show that *i)* the proposed algorithm Relaxed SpecHybrid (RSH), gives always better performance as compared to the standard *k*-means. *ii)* the RSH gives comparable results for the 1150haber (1150TurkishNews) with 5 clusters (relatively small number of clusters case), and comparable or remarkably better results for well-known mini-newsgroups dataset with 20 clusters (relatively high number of

clusters case) as compared to the EM depending on the similarity matrices chosen.

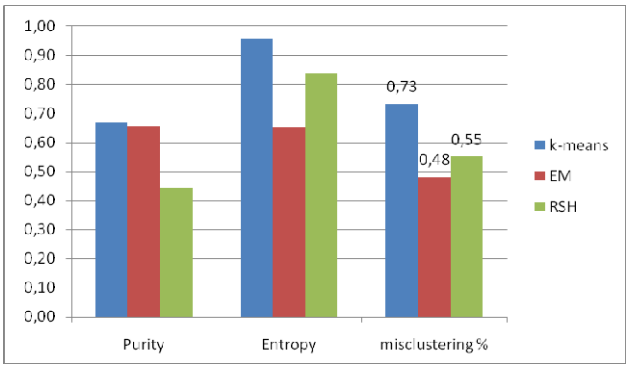

Figure 4. Evaluation of clustering algorithms on 1150haber dataset with tf*idf weighting and feature selection (IG, 2000 terms)

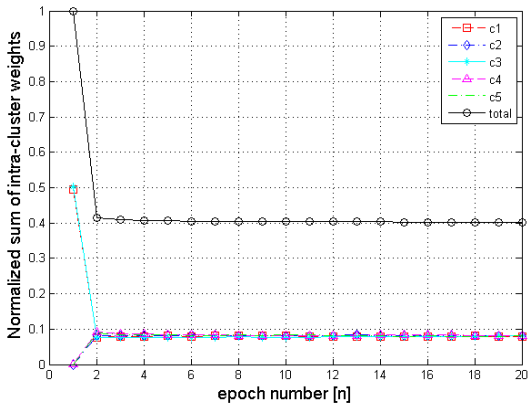

Figure 5. Evaluation of the second phase of the RSH algorithm for mini-newsgroup dataset with IG.

The Figure 5 shows the evaluation of the normalized sum of the intra-cluster distances in the second phase of the RSH algorithm for mini-newsgroup dataset with IG. As seen from the figure, the second phase converges very fast. This observation is in parallel to that of [5] in the case of wireless cellular systems as well.

VI. Conclusions and Future Work

In this paper, we present a relaxed version of the SpecHybrid Algorithm originally proposed for wireless cellular systems in [5], and apply it to text document clustering problem.

Experiment results suggest that the RSH gives better performance as compared to the *k*-means. Furthermore, the RSH gives comparable results for the 1150haber (1150TurkishNews) with 5 clusters (relatively small number of clusters); and comparable or better results for well-known mini-newsgroups dataset with 20 clusters (relatively high number of clusters) as compared to the EM algorithm depending on the similarity matrices used.

In this paper, we have chosen Euclidean metric because it was the only readily available metric in the data mining software tool WEKA at the times when we've carried out the experiments. However, the implementation of the other similarity metrics like cosine similarity and several other evaluation metrics such as Normalized Mutual Information, Rand Index and F-Measure into our experiment setup is about to be finished, and we are going to include those metrics in our near-future research. Furthermore, we are planning to examine the RSH algorithm in document classification problem as well.

References

[1] R. Blumberg, S. Atre, "The Problem with Unstructured Data", Information Management Magazine, February 2003

[2] J. Han, M. Kamber, "Data Mining: Concepts and Techniques", 2nd ed., Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6

[3] A. Zanasi, "Text Mining and its Applications to Intelligence", Crm and Knowledge Management (Advances in Management Information). WIT Press. (2005).

[4] J. Jayabharathy, Dr. S. Kanmani, and A. Ayeshaa Parveen' "A Survey of Document Clustering Algorithms with Topic Discovery", Journal of Computing, vol. 3, no 2, pp. 21-27, Feb.2011.

[5] Z. Uykan, "Spectral Based Solutions for (Near) Optimum Channel/Frequency Allocation", accepted to IWSSIP 2011 (18th International Conference on Systems, Signals and Image Processing), June 2011, Sarajevo, Bosnia and Herzegovina.

[6] D.C. Cox and D.O. Reudink, "Dynamic Channel Assignment in High-Capacity Mobile Communications Systems," Bell System Technical Journal, vol. 50, no. 6, July-August 1971.

[7] Y. Zhao, G. Karpysis, "Empricial and Theoritical Comparisons of Selected Criterion Functions for Document Clustering", Machine Learning, 55, pp.311-331, 2004

[8] B. Babadi and V. Tarokh, "GADIA: A Greedy Asynchronous Distributed Interference Avoidance Algorithm", *IEEE Transactions on Information Theory,* vol. 56, no. 12, Dec. 2010.

[9] A. Huang, "Similarity Measures for Text Document Clustering", *NZCSRSC 2008*, New Zealand, April 2008.

[10] U.V. Luxburg, M. Belkin, and O. Bousquet, "Consistency of spectral clustering", *Annals of Statistics*, vol.36, no. 2, pp.555-586, 2008.

[11] U.V. Luxburg, "A Tutorial on Spectral Clustering", Technical Report TR-149, Max-Planck Institute for Biological Cybernetics, August 2006.

[12] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering", AAAI-2000, Workshop on Artificial Intelligence for Web Search, July 2000.

[13] M.F. Amasyalı, A. Beken, "Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması", SIU 2009, Antalya.

[14] J.B, Lovins, "Development of a stemmer algorithm", Mechanical Translation and Computational Linguistics, 11,1986,pp 22-31.

[15] F. Can, S. Kocberber, E. Balcik, C. Kaynak, H. C. Ocalan, O. M. Vursavas, "Information retrieval on Turkish texts", Journal of the American Society for Information Science and Technology. Vol. 59, No. 3 (February 2008), pp. 407-421.

[16] A.A. Akın , M.D. Akın, "Zemberek, an open source nlp framework for Turkic languages".

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten. "The WEKA Data Mining Software: An Update". SIGKDD Explorations, Volume 11, Issue 1. 2009, pp. 10-18

[18] I. Stojmenovic (Editor), "Handbook of Wireless Networks and Mobile Computing", John Wiley & Sons, 2002.