Position Papers of the Federated Conference on Computer Science and Information Systems pp. 3–9
DOI: 10.15439/2016F371
ACSIS, Vol. 9. ISSN 2300-5963

Clustering Validity Indices evaluation with regards to semantic homogeneity

Tomasz Dziopa
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw
Tomasz.Dziopa@students.mimuw.edu.pl

***Abstract*—Clustering validity indices are methods for examining and assessing the quality of data clustering results. Various studies provide a thorough evaluation of their performance using both synthetic and real-world datasets. In this work, we describe various approaches to the topic of evaluation of a clustering scheme. Moreover, a new solution to a problem of selecting an appropriate clustering validity index is presented. The approach is applied to a problem of selecting a suitable clustering validity index for a real-world task of clustering biomedical articles using the MeSH ontology.**

I. Introduction

THE PROBLEM of clustering is one of the fundamental problems in Machine Learning. The goal of clustering is to find the best way to divide a set of points into groups. This formulation reflects a natural process of learning—humans tend to categorize entities, like objects, people or events into clusters, which are characterized by common attributes. In this paper, we will present a comparison of clustering validity indices with regards to their applications to clustering biomedical documents from PubMed database.

Clustering validity is a common name for quantitative evaluation of the results of clustering algorithms [12]. Clustering Validity Index (CVI) can be perceived as a function which takes as arguments the dataset and clustering scheme and outputs some value which represents the quality of the clustering scheme.

Cluster validity index should provide some insight about the quality of grouping. The most intuitive notions reflected by the concept of "good clustering" are *compactness* and *separation*. The cluster is compact, when points within this cluster are possibly close to each other, whereas clusters are separated, when neighboring clusters are possibly far from each other [4].

The most common applications of cluster validity methods are:

1) Fine-tuning parameters of clustering parameters—the comparison of varying clustering schemes obtained using different parameters in order to find the best grouping.
   One of the most common parameters studied in the literature is the number of clusters in algorithms that assume fixed number of clusters *a priori*, like k-means citeHennig2014.
2) Examining *clustering stability* of a dataset—sensitivity of result of clustering algorithm to modification of algorithm's parameters
3) Examining *clustering tendency*—in some cases we do not know if the dataset has any clustering structure so that it can be grouped in a meaningful way. By applying cluster validity methods we can determine, whether the dataset has adequate grouping structure.

II. Clustering Validity Indices

Clustering Validity Indices are most commonly categorized into three main categories: internal, external and relative. In this chapter we will present the most common methods for assessing the quality of a clustering scheme.

A. External methods

Indices from this group assume that for dataset D some reference clustering $T = \{T_1, \dots, T_m\}$ is given. The idea is that these indices try to express similarity between some scheme $C = \{C_1, \dots, C_k\}$ being examined and T, sometimes referred to as *gold standard*.

Pair-counting indices

We introduce a label for a pair of points (x_a, x_b) for each $x_a, x_b \in D$:

- True Positives: x_a and x_b belong to the same partition in T and are also in the same partition in C.
- False Negatives: x_a and x_b belong to the same partition in T, but are in different partitions in C.
- False Positives: x_a and x_b do not belong to the same partition in T, but they belong to the same partition in C.
- True Negatives: x_a and x_b belong to different partitions in both T and C.

Pair-counting indices are defined as functions calculated over the sizes of the TP, TN, FP and FN sets.

1) Rand Statistic:

$$R = \frac{TP + TN}{TP + TN + FP + FN}$$

Index describes the ratio of correctly guessed pairs (clusterings C and T agree on membership of both points to either the same or different clusters). Perfect clustering will achieve $R = 1$.

2) Jaccard Coefficient:

$$J = \frac{TP}{TP + FN + FP}$$

Perfect clustering achieves $J = 1$, as there are no false negatives and no false positives. Jaccard coefficient is asymmetric in terms of true negatives and true positives, as it ignores true negatives. The influence of pairs of points belonging to the same cluster in both clusterings is amplified and the impact of pairs of points not belonging together is discounted.

3) Fowlkes and Mallows index: Let's introduce the notions of *pairwise precision* and *pairwise recall*, defined as follows:

$$prec = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

The Fowlkes-Mallows index is defined as the geometric mean of the *pairwise precision* and *pairwise recall*:

$$FM = \sqrt{prec \cdot recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$

This measure is asymmetric in terms of true positives and true negatives, because true negatives are ignored. Maximum value of FM is 1, when there are no false positives or negatives.

Matching-based measures

Matching-based measures try to match clusters from C with *gold standard* clusters T and calculate various statistics on the matching.

4) Purity: This measure tries to capture the concept of cluster being *pure* - that is, containing only points from one golden-standard partition. Purity can be defined as follows[13]:

$$purity = \frac{1}{N}\sum_i \max_j |C_i \cap T_j|$$

where T is the set of ground-truth clusters and C is the set of examined clusters.

We can distinguish following cases based on the cardinality of sets C and T:

1) when $|C| = |T|$ and purity $= 1$, then C is a perfect clustering
2) when $|C| > |T|$ purity can still achieve 1, when each cluster in C is a subset of cluster in ground-truth partitioning T
3) when $|C| < |T|$ purity is always < 1 - at least one cluster in C contains points from more than one clusters in T

5) Maximum matching: Maximum matching [13] is defined as the value of maximum matching between sets in C and T - unlike in purity, each cluster in C is assigned a unique partition from T.

More formally, given a graph $G = (V, E)$, where $V = C \cup T$ and $\forall_{i,j}(C_i, T_j) \in E$ we want to find a maximum weighted matching in G. Weights on edges are given as $w(C_i, T_j) = |C_i \cap T_j|$.

The problem of finding a maximum matching in a bipartite weighted graph, assuming $|C| \approx |T|$, can be solved in $O(|C|^2 \log |C| + |C|^3) = O(|C|^3)$ time complexity.

6) F-Measure: Let $T_{match}(i) = \arg\max_{T_j \in T} |C_i \cap T_j|$ denote the cluster in ground-truth partition T, which is represented the most in cluster C_i. We can define precision and recall for cluster C_i as follows:

$$prec_i = \frac{|C_i \cap T_{match}(i)|}{|C_i|}$$

$$recall_i = \frac{|C_i \cap T_{match}(i)|}{|T_{match}(i)|}$$

The F-measure for cluster C_i is a harmonic mean of the precision and recall for this cluster:

$$F_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{recall_i}} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i}$$

F-measure for the clustering scheme C is the mean of F-measures for all clusters:

$$F = \frac{1}{|C|}\sum_{i=1}^{|C|} F_i$$

B. Internal methods

Internal indices are designed to express some properties of the resulting clustering scheme with regards to proximity measure.

Internal methods operate on the proximity matrix, which can be defined as:

$$W = \{\delta\{x_i, x_j\}\}_{i,j=1}^{n}$$

Proximity measure δ should be non-negative, symmetrical and fulfill the triangle inequality.

1) Dunn index: Dunn index is defined as a ratio of the minimum distance between clusters to the maximum cluster's diameter. These two notions can be interpreted in various ways, resulting in various definitions of Dunn Index.

Inter-cluster distance can be defined as:

- minimum distance between points originating from different clusters,
- maximum distance between points originating from different clusters,
- distance between centroids of the clusters

Cluster's diameter can be defined as:

- maximum distance between two points within the cluster,
- mean distance between all pairs of points from the cluster,
- sum of distances of each points to the mean of the cluster

The larger the Dunn index, the better the clustering - the distance between points in different clusters is much larger than the distance between points inside the same cluster. However, Dunn index can be insensitive as inter- and intracluster distance does not capture all information about the clustering.

2) Davies-Bouldin Index: Let μ_i denote the mean of cluster C_i:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

and σ_i denote the dispersion of the points in the cluster C_i around it's mean μ_i:

$$\sigma_i = \sqrt{\frac{\sum_{x_j \in C_i} \delta(x_j, \mu_i)}{|C_i|}}$$

The Davies-Bouldin measure [6] for pair of clusters C_i, C_j is defined as follows:

$$DB_{ij} = \frac{\sigma_i + \sigma_j}{\delta(\mu_i, \mu_j)}$$

DB_{ij} measures the compactness of clusters compared to the distance between the cluster means.

$$DB = \frac{1}{|C|} \sum_{i=1}^{|C|} \max_{j \neq i} \{DB_{ij}\}$$

That is, for each cluster C_i we pick another cluster C_j which produces the largest value of DB_{ij} ratio. The smaller the DB value the better the clustering, because this means that clusters are well-separated (the distance between cluster means is large) and each cluster is compact (has a small spread).

3) Silhouette Coefficient: Silhouette coefficient [11] is a measure of both compactness and separation of clustering. Let a_i denote average dissimilarity of x_i with all other points within its cluster. a_i can be interpreted as how well x_i has been assigned to its cluster. Let b_i denote the lowest average dissimilarity of x_i to any other cluster in C, of which x_i is not a member. Assuming $x_i \in C_j$:

$$a_i = \frac{1}{|C_j|} \sum_{x_l \in C_j; x_l \neq x_i} \delta(x_i, x_l)$$

$$b_i = \min_{C_l \in C; C_l \neq C_j} \frac{1}{|C_l|} \sum_{x_k \in C_l} \delta(x_i, x_k)$$

The silhouette coefficient for data point x_i is defined as:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

s_i can obtain values in interval $[-1, 1]$. $s_i = 1$ indicates that x_i is close to points in its assigned cluster and far from other clusters, $s_i = 0$ indicates that x_i lies close to the boundary between two neighbouring clusters. $s_i = -1$ indicates that x_i is much closer to another cluster than its own cluster - the point has been misclustered.

Silhouette coefficient for clustering C is defined as:

$$SC = \frac{1}{N} \sum_{i=1}^{N} s_i$$

4) Normalized Γ: Let W be the proximity matrix of the dataset, and Y be the proximity matrix defined as follows:

$$Y = \left\{ \delta(\mu_{x_i}, \mu_{x_j}) \right\}_{i,j=1}^{n}$$

μ_{x_i} is the mean of all points that belong to the same cluster as x_i. Let $\mathbf{w}, \mathbf{y} \in \mathbb{R}$ be vectors obtained by linearizing the upper triangular elements excluding main diagonal of W and Y.

Let $\mathbf{z}_W$ and $\mathbf{z}_Y$ denote mean-centered vectors $\mathbf{w}, \mathbf{y}$. Now, the normalized Γ statistic can be defined as:

$$\Gamma_n = \frac{\mathbf{z}_W^T \mathbf{z}_Y}{||\mathbf{z}_W|| \cdot ||\mathbf{z}_Y||}$$

5) Within-Between Ratio: Within-Between Ratio is a ratio of average distance within clusters μ_{within} to average distance between clusters $\mu_{between}$.

$$\mu_{within} = \frac{\sum_{C_i \in C} \sum_{x_j, x_k \in C_i; j \neq k} \delta(x_j, x_k)}{\sum_{C_i \in C} \binom{|C_i|}{2}}$$

$$\mu_{between} = \frac{\sum_{C_i, C_j \in C; i \neq j} \delta(C_i, C_j)}{\binom{|C|}{2}}$$

$$WB = \frac{\mu_{within}}{\mu_{between}}$$

The smaller Within-Between Ratio, the better the clustering scheme.

6) Within cluster sum of squares: Within cluster sum of squares is a sum of within-cluster squared dissimilarities divided by the cluster size.

$$WCSS = \frac{1}{2} \sum_{C_i \in C} \frac{\sum_{x_j, x_k \in C; j \neq k} \delta(x_j, x_k)}{|C_i|}$$

The smaller the Within Cluster Sum of Squares, the more compact are the clusters.

7) Calinski-Harabasz Index: Given a clustering of a dataset $C = \{C_1, \ldots, C_k\}$ consisting of N points, Calinski-Harabasz index is defined as [3]:

$$CH(k) = \frac{SS_B}{SS_W} \cdot \frac{N - k}{k - 1}$$

SS_B is the overall between-cluster variance, defined as:

$$SS_B = \sum_{i=1}^{k} |C_i| \cdot ||\mu_i - \mu||^2$$

where μ_i is a mean of i-th cluster, μ is an overall mean of the sample data, and $||\mu_i - \mu||$ is the L^2 norm (Euclidean distance) between the two vectors.

SS_W is the overall within-cluster variance, defined as:

$$SS_W = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2$$

The larger the value of Calinski-Harabasz index, the better the quality of the clustering scheme - good clustering has large between-cluster variance SS_B and a small within-cluster variance SS_W.

C. Relative Validity Indices

Relative validity indices are used for comparison of clustering schemes. Indices from this group are used for deciding which clustering scheme fits the data best. This definition covers both external and internal indices which can assess the quality in relative terms, however, this notion usually refers to internal methods that are also relative.

The most common use case for relative validity indices is selecting the best clustering scheme from the set of schemes obtained using different parameters.

III. Other approaches to Clustering Validity

A. Ensembles of Clustering Validity Indices

Internal indices presented in the previous section tend to capture only particular properties of evaluated clustering scheme. The question that arises is whether a combination of internal indices would correctly judge the quality of every clustering scheme.

In [9] multiple strategies of building ensembles of clustering validity indices are evaluated. Authors have conducted an experiment on various synthetic and real-life datasets and examined the correlation of the ensembles of relative indices with regards to external validity index. In this work the quality of validity index (or ensemble) is measured as Spearman's rank correlation coefficient with external index, referred to as *effectiveness*.

The basic motivation for validation using ensembles of CVI is that multiple indices with high effectiveness and a high degree of complementarity should produce more robust results than any single index.

Another problem regarding ensembles of CVI is the choice of aggregation strategy. In [9] authors have compared various strategies of combining the results of the indices. The authors found, that score-based strategies, like mean or median of the normalized values of indices in the ensemble, appeared to give inferior results to methods based on rank aggregation. The main advantage of using rank-based aggregation is that it does not rely on the concrete values of indices, therefore it does not require value normalization.

The conclusions from the experiments are, that all examined ensembles, even those assembled from random subsets of measures and with random aggregation strategy, achieved higher effectiveness than the expected value of single validity index. Surprisingly, the strategy of aggregation of results did not affect the overall effectiveness of the ensembles.

B. Semantic Approach to Clustering Validity

We can divide the indices describe previously into two major groups: internal and external. Internal indices rely only on problem space but capture the specific characteristic of the clustering scheme. External indices rely only on explicit clustering assignment therefore limiting the expedience in real world scenarios. Another approach to clustering scheme validation is the use of partial information, which does not reflect direct cluster membership, but rather implies the semantic relationships between the points.

More formally, apart from the dataset D we are given a set $L = \{L_1, \ldots, L_n\}$ consisting of sets of labels with a one-to-one correspondence between elements from L and D.

In this approach, the clustering is produced based only on the original space, discarding the additional information. This information is used afterward to assess the quality of the output clustering.

The formulation of the clustering validation problem is motivated by the rise of popularity of datasets which are annotated using labels, *tags*, or *hashtags*.

Semantic Explorative Evaluation

Semantic Explorative Evaluation described in [10] tries to capture human reasoning of assessing clustering scheme. When an expert faces the problem of manual evaluation of clustering results, he tries to explain the contents of the clusters in his own words. The main idea of SEE is that the quality of the cluster is correlated with the measure of the complexity of expert description of the cluster.

More formally, SEE takes as an input the dataset tagged with expert tags describing the points in the dataset and the clustering. Next, for each cluster, we need to calculate the complexity of expert's description of the cluster - a model of the cluster in terms of expert tags. Any classification algorithm can provide such a model. Then the model complexity measure needs to be defined. Authors have chosen decision tree classifier as their classification algorithm, and an average depth of the resulting tree as the measure of model's complexity.

IV. Evaluation of Cluster Validity Indices

In most works evaluating CVIs, the first step is to choose a set of datasets. Usually, synthetic datasets are used forcing various characteristics of desired clustering schemes, like varying densities, compactness, overlapping, shapes, or added noise. Additionally, a number of sample real-world datasets with known number of clusters can be chosen for the experiments.

A. Optimal K criterion

Cluster validity index evaluation has been thoroughly covered in many papers in recent years. Authors of [5] evaluated multiple papers on the topic. Surprisingly, multiple works appear to use the same methodology.

The methodology requires preparation of synthetic datasets with known number of clusters. For this reason, usually, two-dimensional datasets are used, for the ease of visualization and human verification. Additionally, we need to choose a clustering algorithm for the experiment, which allows an input parameter that sets the number of clusters for the output partition, k. The most popular algorithms in the literature are agglomerative hierarchical algorithm and k-means.

Let's denote the ideal partition of a dataset as P^*. Subsequently, algorithm is run over the dataset with a set $K = \{k_1, \ldots, k_l\}$ of different values of parameter k. As a result, we obtain a set of partitions, $S = \{P_1, \ldots, P_l\}$, with one of them being a partition with a correct number of clusters for the dataset, denoted as P_N. More formally,

$$P_N = \{P_i \in P : |P^*| = |P_i|\}$$

Finally, CVI is computed for all partitions in S. The idea is, that the partition obtaining the best value for the evaluated $CVI(P_x)$ will serve to predict an actual number of clusters. Let's assume for simplicity, that function $CVI(P_x)$ assigns greater values to "better" partitions. We say the partition P_{CVI} is proposed by the cluster validity index, when

$$P_{CVI} = \underset{P_i \in S}{\arg\max}\, CVI(P_i)$$

Clustering validity index has predicted that the dataset contains $|P_{CVI}|$ clusters if it has made a successful guess so that $|P_{CVI}| = |P^*|$.

The method works under a fundamental assumption, that algorithm used for clustering works "correctly" - that is, algorithm-generated partition P_N is the one that fits the data best. Obtained results of CVI are biased if the assumption does not hold so that there exists partition P_i that captures the clustering scheme of the data better.

B. External criteria similarity

The problem of unrealistic assumption of the clustering algorithm being able to correctly partition every dataset has been addressed in [5]. The authors have proposed a modified version of the Optimal k criterion. In contrast to this method, the CVI is said to have succeeded if it has proposed the partition most similar to the optimal partitioning, instead of the partition containing the same number of clusters as an ideal partition.

Similar as in previous method, we need to provide the input dataset D, the set of potential values of parameter k, $K = \{k_1, .., k_l\}$, the set of partitions from a clustering algorithm $S = \{P_1, \dots, P_l\}$ and the *gold standard* partition P^*. Additionally, we need to provide a partition similarity measure $sim(P_i, P_j)$, for example one of external validity indices. Then, the partition obtaining highest similarity of all computed partitions can be defined as:

$$\widehat{P} = \underset{P_i \in S}{\arg\max}\, sim(P_i, P^*)$$

In the new methodology, we say clustering validity index has made a successful guess if $P_{CVI} = \widehat{P}$ - when the partition obtaining the best value of examined CVI is at the same time the most similar to the ideal partitioning.

The similarity measure is another input parameter of the methodology, so its choice can be adapted to the characteristics of the experiment. Extension of this idea is to use multiple partition similarity measures and to either aggregate their results by averaging or using a voting system.

V. Semantic Evaluation of Clustering Validity Indices

We present a new approach to selecting the best clustering validity index. We examine the problem of clustering with additional information. The intuition behind this problem formulation is that we are given a data set, which has been manually annotated with multiple labels reflecting semantic relationships between documents. The clustering is produced based only on the original space, discarding the additional information. This information is used afterward to assess the quality of the output clustering.

More formally, apart from the data set D we are given a set $L = \{L_1, \dots, L_n\}$ consisting of sets of labels with a one-to-one correspondence between elements from L and D.

The motivation behind this formulation of the clustering validation problem is, that it uses additional information about the relationships, which is not an explicit grouping of the points. Moreover, recently the number of data sets annotated using labels, *tags*, or their social-network equivalents, *hashtags* has increased.

The proposed method requires calculating the partitionings S of the dataset using only information from D into k groups, for each k in $K = \{k_1, \dots, k_l\}$. Similarly as in Section IV-B, we want to assess the quality of the clustering using external knowledge, but since we do not have reference clustering, we cannot use an external CVI. Instead, we calculate the semantic quality index $ASH(P_i)$, proposed in [8].

The ASH index uses a notion of *semantic distance*, which is defined for documents T_i, T_j with corresponding sets of assigned expert tags L_i, L_j to be a F1 score between sets L_i, L_j:

$$F_1 distance(T_i, T_j) = 1 - 2 \cdot \frac{precision(L_i, L_j) \cdot recall(L_i, L_j)}{precision(L_i, L_j) + recall(L_i, L_j)} \tag{1}$$

Precision and recall are defined as:

$$precision(L_i, L_j) = \frac{|L_i \cup L_j|}{|L_i|} \tag{2}$$

$$recall(L_i, L_j) = \frac{|L_i \cap L_j|}{|L_j|} \tag{3}$$

The *semantic distance* between two sets of documents is defined as an average of pairwise $F_1 distances$ between pairs of texts from different sets:

$$semDist(D_1, D_2) = \frac{\sum_{T_i \in D_1, T_j \in D_j} F_1 distance(T_i, T_j)}{|D_1| \cdot |D_2|} \tag{4}$$

The measure of document's semantic homogeneity is defined similarly as Silhouette Coefficient:

$$homogeneity(T_i) = \frac{B(T_i) - A(T_i)}{\max(A(T_i), B(T_i))} \tag{5}$$

$A(T_i)$ is a $semDist$ distance calculated between document T_i and all other documents within the same cluster as T_i. $A(T_i)$ can be interpreted as the measure of quality of T_i's assignment to its cluster. In case T_i forms a singleton cluster, $A(T_i) = 0$.

$B(T_i)$ is the $semDist$ measure calculated between document T_i and all documents which do not belong to the same cluster as T_i. $B(T_i)$ is the dissimilarity measure describing how far a document T_i is from all other clusters.

Finally, we define the Average Semantic Homogeneity:

$$ASH = \frac{1}{|D|} \sum_{T_i \in D} homogeneity(T_i) \qquad (6)$$

as the measure of semantic quality of clustering scheme.

Similarly as in a methodology described in Section IV-B, we could formulate the CVI's correctness criterion as:

$$\arg\max_{P_i \in S} ASH(P_i) = P_{CVI} \qquad (7)$$

Stating that the partition obtaining the highest value of Average Semantic Homogeneity is the one which is suggested by examined CVI.

However, in real world scenarios the size and dimensionality of the datasets may be too big for this criterion to select one particular clustering as the one fitting the data best.

We propose a modified CVI correctness criterion: we say that CVI is suitable for the task of evaluation the clustering schemes and preserves semantic relationships between the documents, when:

- CVI calculated on document space indicates optimal number of clusters
- CVI calculated on document space is correlated with Average Semantic Homogeneity index

The main advantage of this method is that it can be used for real-world applications since it does not require a ground-truth partitioning for the dataset. Instead, we pick a random sample from the dataset and have it manually annotated by the experts. Then, using the described method we select the good CVI for assessing the quality of the clustering of the subset of the original dataset. Finally, we state that the selected CVI is appropriate for assessing the quality of the original dataset.

VI. The Experiment

We have conducted an experiment to demonstrate the usage of Semantic Evaluation of Clustering Validity Indices in order to find the best CVI for assessing the quality of clustering text documents.

In the experiment, we used a dataset obtained from U.S. National Library of Medicine (NLM). The dataset consists of 42200 abstracts of scientific articles in English. Each article has been manually labeled by experts from NLM using concepts from MeSH ontology [1] using on average 12 concepts.

Abstracts were tokenized, stemmed and common English stopwords were removed. Documents are modeled using bag-of-words in a document-term matrix with tf-idf weighting [2]. MeSH terms are treated as labels, discarding the information about major topics and contexts in which the term appears, resulting in the total of 17169 unique labels.

The experiment has been conducted on a subset of 5054 documents from NLM dataset which were selected from the MeSH headings presented in the Table I. The selection of thematically disjoint headings is intended to strengthen the clustering tendency of the dataset.

TABLE I
CHARACTERISTICS OF DATASET USED FOR EXPERIMENT

MeSH heading	No. of documents
Bacterial proteins	772
Brain	652
Breast Neoplasms	848
Pregnancy	1033
E. Coli	574
HIV	629
Malaria	251
Diabetes	491

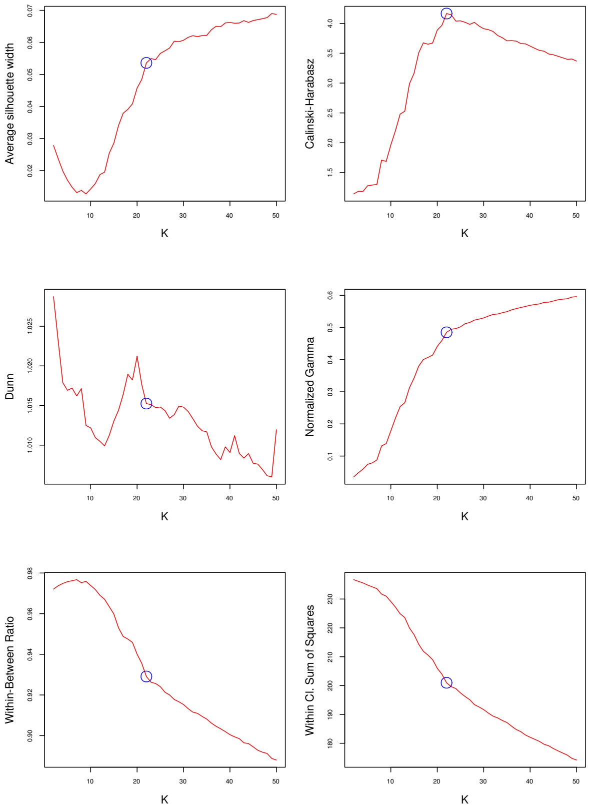

Fig. 1. Values of CVI indexes calculated by Average silhouette width, Calinski-Harabasz, Dunn, Normalized Γ, Within-Between Ratio and Within Cluster Sum of Squares methods on clusterings into $k \in [2, 50]$ groups in document space. Values for optimal number of clusters $k = 22$ have been marked with blue circle.

The values presented on Figure 1 are mean aggregate on 10 randomly chosen without replacement, equinumerous subsets ($|D| \approx 505$). The datasets have been partitioned with Agglomerative Nesting algorithm implementation using cosine distance. The values of indices were calculated using `cluster.stats` implementation from R package `fpc` [7].

Figure 2 shows the values of Average Semantic Homogeneity - the value of semantic quality of clustering. Average Semantic Homogeneity achieves higher values for smaller clusters, with a maximum value of 1 for singleton clusters. We can find the best value of parameter k using the elbow criterion, with possible values of $k = \{5, 9, 14, 22\}$.

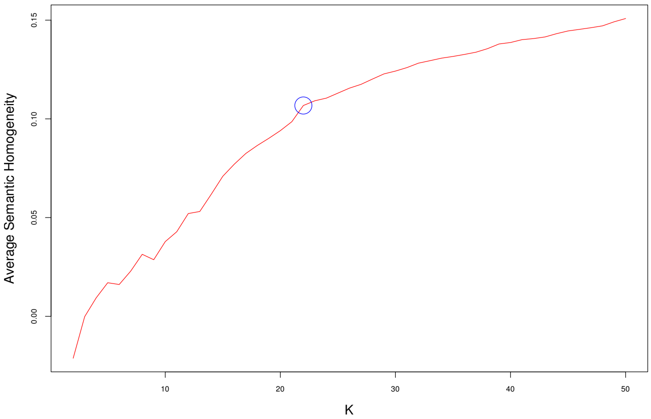

Fig. 2. Value of Average Semantic Homogeneity in experiment setup with optimal number of clusters $k = 22$ marked with blue circle

TABLE II
THE RESULTS OF COMPARISON

Index	L_2 Norm	best k
Calinski-Harabasz	1.200	22
Average Silhouette Width	1.014	22 (elb. crit.)
Dunn	2.102	20 (local max.)
Normalized Γ	0.184	18, 22 (elb. criterion)
Within-Between Ratio	1.889	-
Within Cl. Sum of Squares	1.063	-

The results of the experiment are summarized in Table II. The L_2 norm has been calculated on 0-1 normalized values of Average Semantic Homogeneity and examined clustering validity indices. Our experiment shows, that Normalized Γ shows very high correlation with ASH. Moreover, the index enables to find the best value of parameter k for evaluated dataset.

The Average Silhouette Width and Calinski-Harabasz index show relatively high correlation with ASH and both reach the optimal value of $k = 22$. Moreover, Calinski-Harabasz index achieves global maximum at $k = 22$, additionally strengthening the argument of 22 being the optimal value of k.

Dunn index has a relatively weak correlation with ASH, although it has the local optimum at $k = 20$. We might suppose, that it does not preserve semantic relationships between the documents, and suggested value is derived from other properties of the dataset.

The remaining indices, Within-Between Ratio and Within Cluster Sum of Squares do not suggest an optimal number of partitions for the dataset.

VII. Conclusions

In this work, we have shown that for the given problem, Calinski-Harabasz, Average Silhouette Width and Normalized Γ indices appear to reflect semantic relationships between the clusterings using bag-of-words and labels annotations models.

In contrast to previous studies, the method does not make any assumption on the correctness of the clustering algorithm. Moreover, this approach does not require datasets with known ground-truth partitioning. The presented methodology using the semantic measure of clustering scheme can be used in real-life problems. Additionally, the number of datasets tagged with expert labels or ontologies has increased in recent years.

One of the many possible directions for development of this method is to evaluate other measures of the semantic quality of the clustering. Additionally, we could make more extensive use of the MeSH ontology. In this thesis, we treated the concepts as labels, but we can take advantage of the tree-like structure hierarchy of MeSH terms and incorporate this knowledge into distance calculation in the label representation of data.

Furthermore, future studies should investigate the applications of the method to other types of documents. In this work we have examined the usage of the method with scientific documents from a particular domain only, whereas the overall applicability to clustering other kinds of documents should be researched. It is also worth examining the applications of the ensembles [9] of multiple CVIs pointed out as suitable for assessing the document clusters by our method.

References

[1] https://www.nlm.nih.gov/mesh/introduction.html, 2016. [Online; accessed 5.05.2016].

[2] Charu C. Aggarwal and Cheng Xiang Zhai. *Mining Text Data*. Springer Publishing Company, Incorporated, 2012.

[3] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, 3(1):1–27, 1974.

[4] Brian Everitt. Cluster analysis. *Quality and Quantity*, 14(1):75–100, 1980.

[5] Ibai Gurrutxaga, Javier Muguerza, Olatz Arbelaitz, Jesús M. Pérez, and José Ignacio Martín. Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recognition Letters*, 32(3):505–515, 2011.

[6] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145.

[7] Christian Hennig. *fpc: Flexible Procedures for Clustering*, 2015. R package version 2.1-10.

[8] Andrzej Janusz, Dominik Ślęzak, and Hung Son Nguyen. Unsupervised similarity learning from textual data. *Fundam. Inf.*, 119(3-4):319–336, August 2012.

[9] Pablo A. Jaskowiak, Davoud Moulavi, Antonio C. S. Furtado, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. On strategies for building effective ensembles of relative clustering validity criteria. *Knowledge and Information Systems*, pages 1–26, 2015.

[10] Hung Son Nguyen, Sinh Hoa Nguyen, and W. Swieboda. Semantic explorative evaluation of document clustering algorithms. In *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*, pages 115–122, Sept 2013.

[11] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[12] Sergios Theodoridis and Konstantinos Koutroumbas. Chapter 16 - cluster validity. In Sergios Theodoridis, , and Konstantinos Koutroumbas, editors, *Pattern Recognition (Fourth Edition)*, pages 863 – 913. Academic Press, Boston, fourth edition edition, 2009.

[13] Mohammed J. Zaki and Jr. Wagner Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, May 2014.