Data & Knowledge Engineering 92 (2014) 77–89

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Editorial

# WB-index: A sum-of-squares based index for cluster validity

Qinpei Zhao [a,*], Pasi Fränti [b]

[a] School of Software Engineering, Tongji University, 201804, China
[b] School of Computing, University of Eastern Finland, 80110, Finland

ARTICLE INFO

*Article history:*
Received 5 September 2012
Received in revised form 2 May 2014
Accepted 11 July 2014
Available online 17 July 2014

*Keywords:*
Cluster validity
Keywords categorization
Short text mining
Clustering
Classification and association rules

ABSTRACT

Determining the number of clusters is an important part of cluster validity that has been widely studied in cluster analysis. Sum-of-squares based indices show promising properties in terms of determining the number of clusters. However, knee point detection is often required because most indices show monotonicity with increasing number of clusters. Therefore, indices with a clear minimum or maximum value are preferred. The aim of this paper is to revisit a sum-of-squares based index called the WB-index that has a minimum value as the determined number of clusters. We shed light on the relation between the WB-index and two popular indices which are the Calinski–Harabasz and the Xu-index. According to a theoretical comparison, the Calinski–Harabasz index is shown to be affected by the data size and level of data overlap. The Xu-index is close to the WB-index theoretically, however, it does not work well when the dimension of the data is greater than two. Here, we conduct a more thorough comparison of 12 internal indices and provide a summary of the experimental performance of different indices. Furthermore, we introduce the sum-of-squares based indices into automatic keyword categorization, where the indices are specially defined for determining the number of clusters.



## 1. Introduction

Different clustering algorithms with various cost functions produce different solutions, and there is no single best clustering algorithm for all possible data sets. A primary task is therefore to select the best possible clustering for a given data set. *Cluster validity* provides a way of validating the quality of clustering algorithms and a means of discovering the natural structure of data sets. For most clustering algorithms, the number of clusters is a main parameter. However, the setting of the parameter is not always known and hence determining the number of clusters is essential. Cluster validity measures can be used for determining the number of clusters. In fact, the clustering procedure and cluster validity are much like the chicken-and-the-egg problem where knowing how to define a good cluster validity index requires understanding the data and the clustering algorithm, but the clustering algorithm is one of the principal tools used to understand the data [1] without a priori information. Therefore, the study of the cluster validity is as important as that of the clustering algorithms.

The internal validity index, as one of the categories of cluster validity, evaluates the results of a clustering algorithm in terms of quantities that involve the vectors of the data set themselves. A good clustering algorithm generates clusters with intra-cluster homogeneity, inter-cluster separation and connectedness. One category of internal validity indices is based on these properties. Another category is based on whether the internal indices are applied to hard (crisp) or soft (fuzzy) clustering. A review of fuzzy cluster validity indices is available in [2].

* Corresponding author.
*E-mail address:* qinpeizhao@tongji.edu.cn (Q. Zhao).

http://dx.doi.org/10.1016/j.datak.2014.07.008
0169-023X/

There are many examples of internal validity indices based on the compactness within a cluster and the separation between clusters. The sum-of-squares based indices are founded on *sum-of-squares within cluster* (SSW) and/or *sum-of-squares between clusters* (SSB) values, for example, *Ball and Hall* [3], *Hartigan* [4], Calinski and Harabasz (CH) [5] and *Xu* [6]. *WB-index* [7] is a sum-of-squares based index where a minimum value can be attained as the number of clusters. Examples of other popular indices are given in [8–12]. *Dunn*-type indices [8] are based on the inter-cluster distance and diameter of a cluster hypersphere. A Dunn index is sensitive to outliers, whereas the *Davies and Bouldin* index is defined by the average of cluster evaluation measures for all the clusters. *Xie–Beni* [10] adopts the minimum distance between any pair of clusters and the global average of distances between each data object and clusters as inter- and intra-cluster distances, respectively. *S_Dbw* [11] replaces the total separation with the density of data objects in the middle of two clusters and omits the weighting factor. A model selection method called the *Bayesian information criterion* (BIC) [12] has been applied in model-based clustering, but it can be adapted to partition-based clustering [13], too. To reduce computation on distance calculations in the *silhouette coefficient* (SC), an approach is proposed in [14] to compute the SC quickly by decreasing the number of addition operations.

Milligan and Cooper [15] have presented a comparison of over 30 internal validity indices as stopping criteria in hierarchical clustering algorithms. Dimitriadou et al. [16] conducted a comparison study of 15 validity indices on binary data. An organized study of 16 external validation measures for k-means clustering is given in [17,18]. A survey of cluster validation indices in the analysis of post-genomic and other data familiarizes researchers with some of the fundamental concepts behind cluster validation techniques [19].

Existing studies of the comparisons of internal validity indices are primarily based on one certain clustering algorithm, for example, hierarchical clustering algorithm. In this paper, we study 12 indices on two clustering algorithms, the k-means and *Random Swap* (RS) clustering algorithms [20,21]. Conventional k-means is known to have local maxima problem because of its initialization. The RS algorithm improves on the conventional k-means, which has been shown to obtain a more stable result. We chose these two clustering algorithms to study the performance of the indices depending on if the clustering results are stable or not.

Besides a thorough comparison among the indices, we also revisit the WB-index. We attempt to analyze the difference and relation between the WB-index and two similar indices (CH and Xu-index) theoretically. For two-dimensional data, the three indices seem to work similarly. However, according to the theoretical analysis, it is shown that the CH index is more affected by the data size $N$ and more sensitive to the degree of overlapping of data sets than the WB-index. For low dimensional data (i.e., $D \leq 2$), the Xu-index is remarkably close to the WB-index. However, it rarely detects a clear minimum value for high dimensional data (i.e., $D > 2$).

Cluster validity indices are discussed in many applications, such as image segmentation [22], transactional data [23], post-genomic data analysis [19] and text clustering [24,25]. With more and more short text snippets being generated, such as search query and results, microblogs, tweets and other types of comments in social network, there is a growing interest in mining the short text by clustering techniques [26–28]. Short texts typically have limited length, pervasive abbreviations and coined acronyms. In most cases, there is no contextual information. Therefore, traditional clustering techniques used for mining documents are not suitable for short text clustering. Automatic keywords categorization is to cluster short texts of keywords with a determined number of clusters.

For automatic keywords categorization, the primary issues are the semantic similarity measure for keywords and the cluster validity in the clustering method. Since short text snippets lack contextual information, external knowledge is introduced to enrich the information contained within short texts (for example, Wikipedia data [29,30], Google search results [31] and lexical databases [32]). The number of clusters obtained through cluster validity measures is a parameter showing the match between the clustering results and users' expectation. There is very little research on cluster validity measures employing in automatic keywords categorization. We introduce a procedure for automatic keywords categorization. Based on the compactness within a class and separation between classes, variants of the sum-of-squares based indices that are the CH index, the Xu-index and the WB-index are introduced and compared in this paper.

## 2. Background

### 2.1. Preliminaries

Determining the number of clusters ($M$) relies on the cluster validity indices. Meanwhile, the number of clusters determined is used to validate a validity index. Thus, the problem of determining the number of clusters is commonly studied as a key problem in cluster validity. In order to determine the optimal number of clusters $M^*$, other parameters are fixed and the parameter $M$ is optimized by the validity criteria. A procedure for determining the optimal number of clusters is given as follows. Given the data set $X$, a specific clustering algorithm and a fixed range of number of clusters $[M_{min}, M_{max}]$, the basic procedure involves:

1. Repeat a clustering algorithm successively for the number of clusters $M$ from predefined values of $M_{min}$ to $M_{max}$.
2. Obtain the clustering results (partitions $P$ and centroids $C$) and calculate the index value for each.
3. Select the $M$ as $M^*$ for which the partition provides the best result according to some criteria (minimum, maximum or knee point).
4. Compare the detected number of clusters ($M^*$) with external information if available.

A summary of commonly used internal validity indices appears in Table 1. Symbol $X = \{x_1,\ldots,x_N\}$ represents the data set with $N$ $D$-dimensional points, and $\overline{X} = \sum_{i=1}^{N} x_i/N$ is the center of the entire data set. The centroids of clusters are $C = \{c_1,\ldots,c_M\}$, where $c_i$ is the $i$th cluster and $M$ is the number of clusters. The log-likelihood in the BIC is defined as $L$, and in Xie–Beni $u_{ik}$ denotes the membership of the $i$th point to the $k$th cluster.

**Table 1**
Formulas for internal indices.

| Name | Formula |
|---|---|
| SSW | $SSW = \sum_{i=1}^{N} \|x_i - C_{p_i}\|^2$ |
| SSB | $SSB = \sum_{i=1}^{M} n_i \|c_i - \overline{X}\|^2$ |
| Calinski–Harabasz [5] | $CH = \frac{SSB/(M-1)}{SSW/(N-M)}$ |
| Hartigan [4] | $H_M = \left(\frac{SSW_M}{SSW_{M+1}} - 1\right)(N-M-1)$ or: $H_M = log(SSB_M/SSW_M)$ |
| Krzanowski–Lai [33] | $diff_M = (M-1)^{2/D} SSW_{M-1} - M^{2/D} SSW_M$<br>$KL_M = \|diff_M / diff_{M+1}\|$ |
| Ball & Hall [3] | $BH_M = SSW_M / M$ |
| Xu-index [6] | $Xu = Dlog\left(\sqrt{SSW_M/\left(DN^2\right)}\right) + logM$ |
| Dunn's index [8] | $d(c_i, c_j) = min_{x \in c_i, y \in c_j} \|x-y\|^2$<br>$diam(c_k) = max_{x,y \in c_k} \|x-y\|^2$<br>$Dunn = \frac{min_{i=1}^{M} min_{j=i+1}^{M} d(c_i, c_j)}{max_{k=1}^{M} diam(c_k)}$ |
| Davies&Bouldin [9] | $R_{ij} = \frac{S_i + S_j}{d_{ij}}, i \neq j$ where: $d_{ij} = \|c_i - c_j\|^2$, $S_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \|x_j - c_i\|^2$ and, $R_i = \max_{j=1,\ldots,M} R_{ij}, i = 1, \ldots, M$<br>$DBI = \frac{1}{M} \sum_{i=1}^{M} R_i$ |
| SC [14] | $a(x_i) = \frac{1}{n_m - 1} \sum_{j=1, j \neq i}^{n_m} \|x_i - x_j\|^2_{x_i, x_j \in c_m}$<br>$b(x_i) = min\left\{ \sum_{t \neq m} \|c_t - c_m\|^2 \right\}_{x_i \notin C_t}$<br>$s(x_i) = \frac{b(x_i) - a(x_i)}{max(a(x_i), b(x_i))}$<br>$SC = \frac{1}{N} \sum_{i=1}^{N} s(x_i)$ |
| BIC [34] | $BIC = L * N - \frac{1}{2} M(D+1) \sum_{i=1}^{M} log(n_i)$ |
| Xie–Beni [10] | $XB = \frac{\sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik}^2 \|x_i - C_k\|^2}{N \min_{t \neq s} \left\{ \|C_t - C_s\|^2 \right\}}$ |

### 2.2. Knee point detection

The knee point potentially indicates the optimal number of clusters, but the act of locating the knee point in the validity index curve has not been well studied. The maximum and minimum values are the most straightforward knee points. Indices with clear maximum and minimum value are preferred. Some indices, for example SSW and log-likelihood, are monotonous and as such there is no clear knee point. Other indices might have several local maximum or minimum values due to the selection of $M_{max}$.

The second successive difference between index values (Fig. 1) can be used for knee point detection, although this approach only reflects local information. Other methods, such as the L-method [35], have been proposed to find the knee point of the curve by examining the boundary between the pair of straight lines that most closely fit the curve in hierarchical/segmentation clustering. More general methods should be used based on the global trend of the curve.

## 3. WB-index versus the Calinski–Harabasz index and the Xu-index

A SSW cluster is a commonly used measure of compactness, while a SSB cluster is a measure of separation. Sum-of-squares-based indices (see Table 1) are mainly functions of $M$, $N$, $D$, $SSW$ and $SSB$, and they usually have a so-called elbow phenomenon (see Fig. 1), where knee point detection is required. The second successive difference is commonly used for knee point detection.

The WB-index [7] is defined as:

$$WB(M) = M \cdot SSW / SSB \tag{1}$$

We revisit the WB-index to determine the difference among the WB-index and the CH index and the Xu-index based on theoretical analyses. The CH index and Xu-index are two commonly used sum-of-squares based indices, which work similarly as the WB-index.

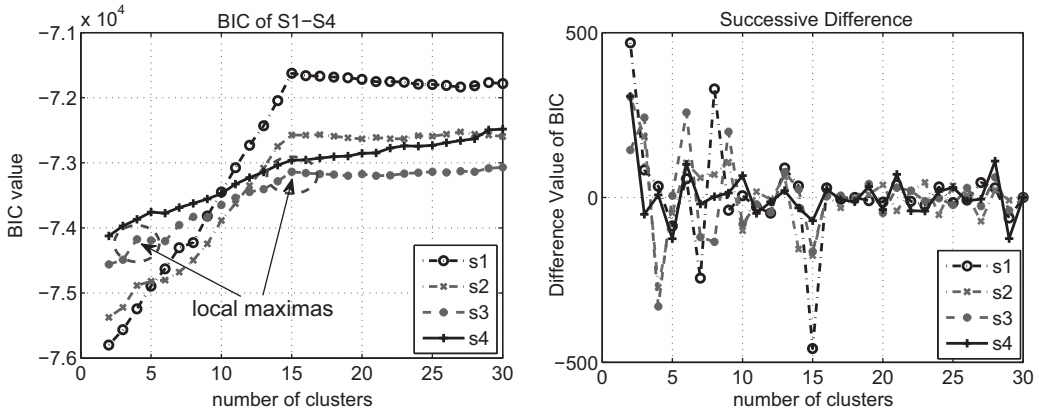


**Fig. 1.** Number of clusters versus the BIC on S1–S4 (see Section 5) and their second successive differences.

Let us assume that cluster $i$ has $n_i$ points and there is a ground-truth average point for cluster $i$, which is $x_i$. (see Fig. 2). The within-cluster variance for cluster $i$ ($W_i$) can then be reformulated as:

$$\begin{aligned} W_i &= n_i \|x_i - c_i\|^2, i \in [1, M] \\ B_i &= n_i \|\overline{X} - c_i\|^2, i \in [1, M] \\ \frac{SSW}{SSB}(M) &= \frac{\sum_{i=1}^{M} W_i}{\sum_{i=1}^{M} B_i} = \frac{W_1 + W_2 + \ldots + W_M}{B_1 + B_2 + \ldots + B_M} > 0 \end{aligned} \tag{2}$$

With an increment of one cluster, the difference of *SSW*/*SSB* can be written as:

$$\begin{aligned} \Delta\frac{SSW}{SSB}(M) &= \frac{SSW}{SSB}(M-1) - \frac{SSW}{SSB}(M) \\ &= \frac{\sum_{i}^{M-1} W_i}{\sum_{i}^{M-1} B_i} - \frac{\sum_{i}^{M-1} W_i + W_M}{\sum_{i}^{M-1} B_i + B_M} \\ &= \frac{\left(\sum_{i}^{M-1} W_i\right)\left(\sum_{i}^{M-1} B_i + B_M\right) - \left(\sum_{i}^{M-1} B_i\right)\left(\sum_{i}^{M-1} W_i + W_M\right)}{\left(\sum_{i}^{M-1} B_i\right)\left(B_M + \sum_{i}^{M-1} B_i\right)} \\ &= B_M \frac{\sum_{i}^{M-1} W_i - W_M \sum_{i}^{M-1} B_i}{\left(\sum_{i}^{M-1} B_i\right)\left(B_M + \sum_{i}^{M-1} B_i\right)} \end{aligned} \tag{3}$$

Since $\sum_i^{M-1} W_i$ is monotonically decreasing and $\sum_i^{M-1} B_i$ is monotonically increasing with respect to increasing $M$, $\Delta SSW/SSB(M)$ then monotonously decreases as well.

This result indicates that the decrement of *SSW*/*SSB* from cluster size *M-1* to *M* is larger than that from *M* to $M+1$. i.e., the decrement decreases with increasing *M* (see Fig. 3). When the decrement degree of $\Delta(SSW/SSB)$ is larger than the linear increment of *M* at the beginning, *WB* decreases until $M^* \geq M_{min}$. A special case is that *WB* is increasing for all *M* when $M^* \geq M_{min}$. Thus, there exists a value $M^*$ such that $WB(M) \geq WB(M^*)$ for $M \leq M^*$ and $WB(M) < WB(M^*)$ for $M > M^*$. The optimal number of clusters is determined by the minimum value of the WB-index. The result of the WB-index for data S1–S4 (see Section 5 for the data) is shown in Fig. 4. Although SSW decreases monotonically with increasing *M*, WB-index has a U-shape with clear minima at $M = 15$.

Based on the facts presented above, we can establish the relationship of the WB-index with the CH index and the Xu-index.

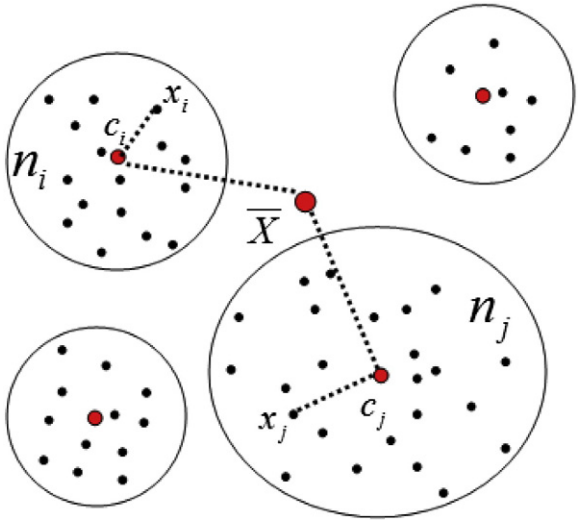


**Fig. 2.** Calculation of $W_i$ and $B_i$.

For the CH index and the WB-index, we have

$$WB \times CH = \frac{M(N-M)}{M-1} = \left(1 + \frac{1}{M-1}\right)(N-M) \tag{4}$$

Since $2 \le M \le N$, we obtain an upper bound that $WB \times CH \le 2(N-2)$. Therefore,

$$WB \le \frac{2(N-2)}{CH} \tag{5}$$

Based on Eq. (5), the WB-index shows a similar trend as the inverse of the CH index. However, the CH index is affected by the data size $N$. When $N$ is large, the factor $\frac{M-1}{N-M}$ plays a more important role than $SSW/SSB$ in the whole index. Considering data with overlapping clusters, for example S1–S4, which have increasing degree of overlapping, the CH index has difficulty dealing with highly overlapping data such as S3 and S4. For higher overlap, $SSW/SSB$ contains less information about $M^*$.

The Xu-index can be written as:

$$Xu = log\frac{M(SSW)^{D/2}}{(DN^2)^{D/2}} \tag{6}$$

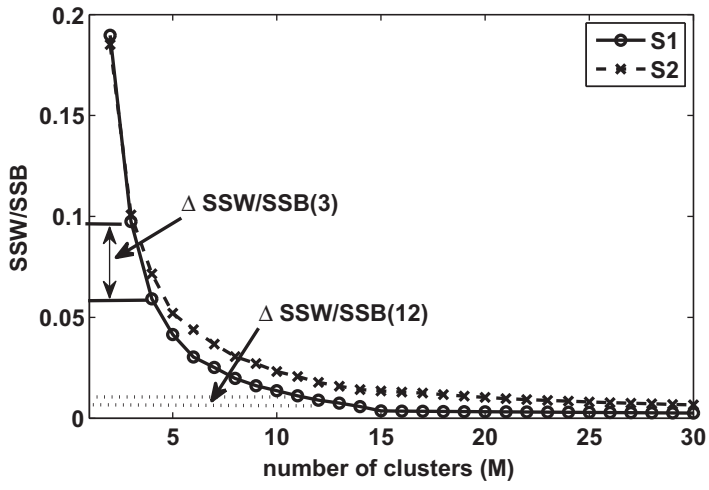


**Fig. 3.** Values of $SSW/SSB$ as a function of $M$ for clustering results on data S1 and S2. $\Delta SSW/SSB$ is decreasing when $M$ is linearly increasing.

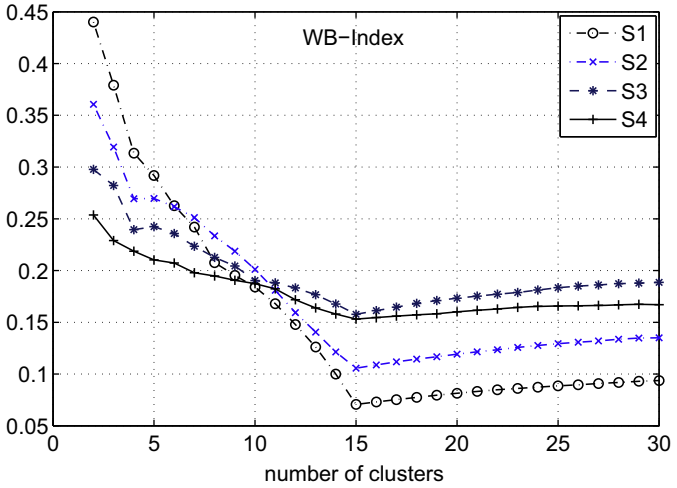


**Fig. 4.** The WB-index versus the number of clusters for S1–S4.

Then,

$$e^{Xu} = \frac{M(SSW)^{D/2}}{(DN^2)^{D/2}} = \frac{M \times SSW \times SSW^{D/2-1}}{(DN^2)^{D/2}} = \frac{WB \times SSB \times SSW^{D/2-1}}{(DN^2)^{D/2}} \approx WB \times SSB \times SSW^{D/2-1} \tag{7}$$

The relation between the Xu-index and the WB-index depends on the dimension $D$.

$$\begin{aligned} e^{Xu} &= WB \times f_1(SSB, SSW), D \le 2 \\ e^{Xu} &= WB \times f_2(SSB, SSW), D > 2 \end{aligned} \tag{8}$$

where $f_1$ is a monotonically increasing function dominated by $SSB$ and $f_2$ is dominated by $SSW$, which is monotonically decreasing. As shown in Fig. 5, $SSB$ has smaller effect than $SSW$. The information about $M^*$ originates mainly from $M \times SSW$. Therefore, the Xu-index is close to the WB-index when $D \le 2$, but for $D > 2$, the Xu-index is dominated by $SSW$, which rarely finds a clear minimum value.

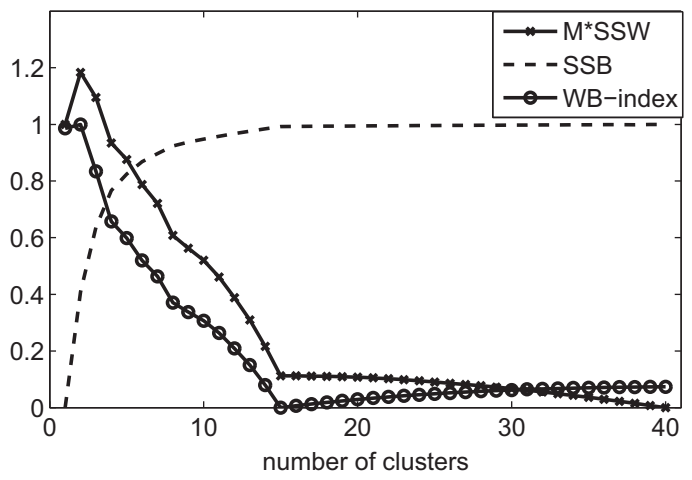


**Fig. 5.** Plot of $M \cdot SSW$ and $SSB$ and the WB-index for data S1. $SSW$, $SSB$ and the WB-index are normalized to one.

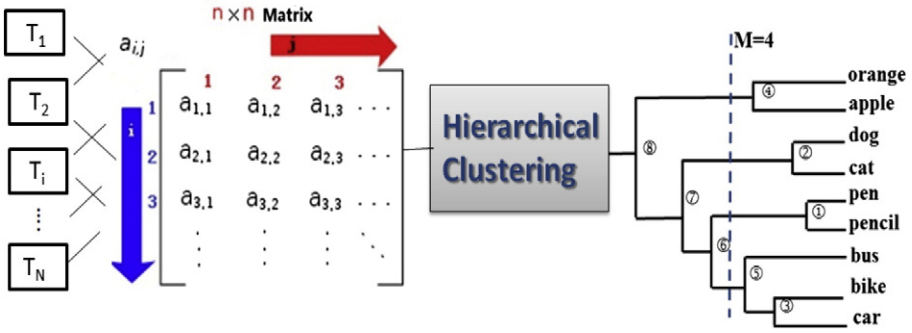


**Fig. 6.** A procedure for automatic keywords categorization.

## 4. Automatic keywords categorization

Given a list of $N$ keywords $T = \{T_1, T_2, \ldots, T_N\}$, keywords categorization aims at clustering the keywords into groups. The keywords can be obtained from several sources such as movies, services, and tags. The generated clusters are defined as $C = \{c_1, c_2, \ldots, c_M\}$. A procedure for the automatic keywords categorization is introduced in Fig. 6. There are two main aspects to consider in the procedure. The first aspect is the similarity measure between two keywords to obtain the similarity matrix. With the similarity matrix, clustering algorithms such as hierarchical or spectral clustering can be employed to obtain the clusters. The second aspect is how to determine the number of clusters.

Calculating the semantic similarity between two keywords directly is an open question. We use a lexical database *WordNet* in this paper. WordNet provides a hierarchy for thesaurus and a part of the WordNet hierarchy by a web interface *WordVis*[1] is shown in Fig. 7. The similarity score between two keywords can then be derived from the hierarchy. For example, the similarity of *eatery* and *grill* from *Jiang and Conrath* [36] is 0.36.

We defined a modified WB-index in [37] for determining the number of groups in keywords categorization. In this paper, we expand our work on the CH index and the Xu-index also. The basic elements of the sum-of-squares based indices are defined as:

$$\begin{aligned} SSW(M) &= \max_t \left\{ \max_{i,j} JC\left(T_i, T_j\right)_{T_i \neq T_j \in c_t} \right\} + \sum_{|c_t|=1} 1 \\ SSB(M) &= \sum_{t=1}^{M} \sum_{s>t}^{M} \min_{i,j} JC\left(T_i, T_j\right)_{T_i \in c_t, T_j \in c_s} \end{aligned} \tag{9}$$

In $SSW(M)$, $T_i$ and $T_j$ are the $i$th and $j$th keywords in cluster $c_i$. Since it is not possible to calculate the similarity with only one keyword in a cluster, we sum up the number of clusters with single a keyword according to $\sum_{|c_t|=1} 1$. Similarly, $T_i$ and $T_j$ in $SSB(M)$ are the $i$th and $j$th keywords in cluster $c_t$ and cluster $c_s$ respectively, $M$ is the number of clusters.

Therefore, the three sum-of-squares based are defined:

$$\begin{aligned} WB{-}index &= M\frac{SSW(M)}{SSB(M)} \\ CH &= \frac{SSB/(M-1)}{SSW/(N-M)} \\ Xu{-}index &= log\sqrt{SSW/N^2} + logM \end{aligned} \tag{10}$$

## 5. Experiments

The data[2] in the experiment include shaped, Gaussian-like and real data. The properties of the data, such as name, data size, dimension and the number of clusters for reference are summarized in Table 2. The internal indices based on k-means and RS [20] are implemented in the *C* programming language.[3]

### 5.1. Comparisons of indices

The validity indices are tested with the k-means and the RS with repetitions. The results are examined in light of the performance of the indices with different clustering algorithms using both artificial and real data, and $M_{min} = 2$ and $M_{max} = \sqrt{N}$. The values of 12 indices in the test are computed for all clusters in $[M_{min}, M_{max}]$. The determined number of clusters corresponds to the minimum (Krzanowski–Lai, Xu, Wb, DBI, Xie–Beni) or maximum value (CH, Dunn, SC, SCI and BIC) of the indices. For some of the indices (Ball & Hall, Hartigan), the minimum or maximum value of the second successive difference is used as a knee point detection method.

[1] http://wordvis.com/.
[2] http://cs.joensuu.fi/sipu/datasets/.
[3] http://cs.joensuu.fi/sipu/soft/.


**Fig. 7.** A part of the WordNet taxonomy visualized through the *WordVis* web-interface.

We plot the performance of validity indices on DBI, Xie–Beni and the WB-index with the k-means and the RS. As shown in Figs. 8 and 9, the validity indices with the k-means rarely achieve the correct number of clusters. However, there are clear minima for indices with the RS. Furthermore, the indices have higher variance with the k-means than the RS, so it is necessary to choose a stable algorithm in cluster validity. Indexes using the min or max function such as DBI and Xie–Beni have high variance among 100 repetitions. On the other hand, sum-of-squares indices such as the WB-index are more stable.

The numbers in Tables 3, 4 and 5 are the mean values of the determined number of clusters using different validity indices with the RS. The average values are obtained from 100 repetitions, except for birch1. Restricted by the running time, we calculate the indices once for birch1.

For unbalanced and shaped data (e.g., Aggregation), as well as for data with densities (e.g., Compound), almost all of the indices fail (see in Table 3 that the indices obtain 0% correctly determined number of clusters). It is interesting to view the performance of the indices for these data sets through the lens of density-based clustering or spectral clustering instead of RS, which can be studied in

**Table 2**
Attributes of the data sets that have been used.

| Name | Data size | Dimension | # of Clusters |
|---|---|---|---|
| *Shaped data sets* | | | |
| Touching | 73 | 2 | 2 |
| pathbased | 300 | 2 | 3 |
| Compound | 399 | 2 | 6 |
| Aggregation | 788 | 2 | 7 |
| *Gaussian-like data sets* | | | |
| S1–S4 | 5000 | 2 | 15 |
| R15 | 600 | 2 | 15 |
| D31 | 3160 | 2 | 31 |
| birch1 | 100,000 | 2 | 100 |
| *Real data sets* | | | |
| Iris | 150 | 3 | 3 |
| Wine | 178 | 13 | 3 |
| Control | 600 | 60 | 6 |
| Image | 2320 | 18 | 7 |
| Wdbc | 569 | 30 | 2 |
| Yeast | 1484 | 8 | 10 |

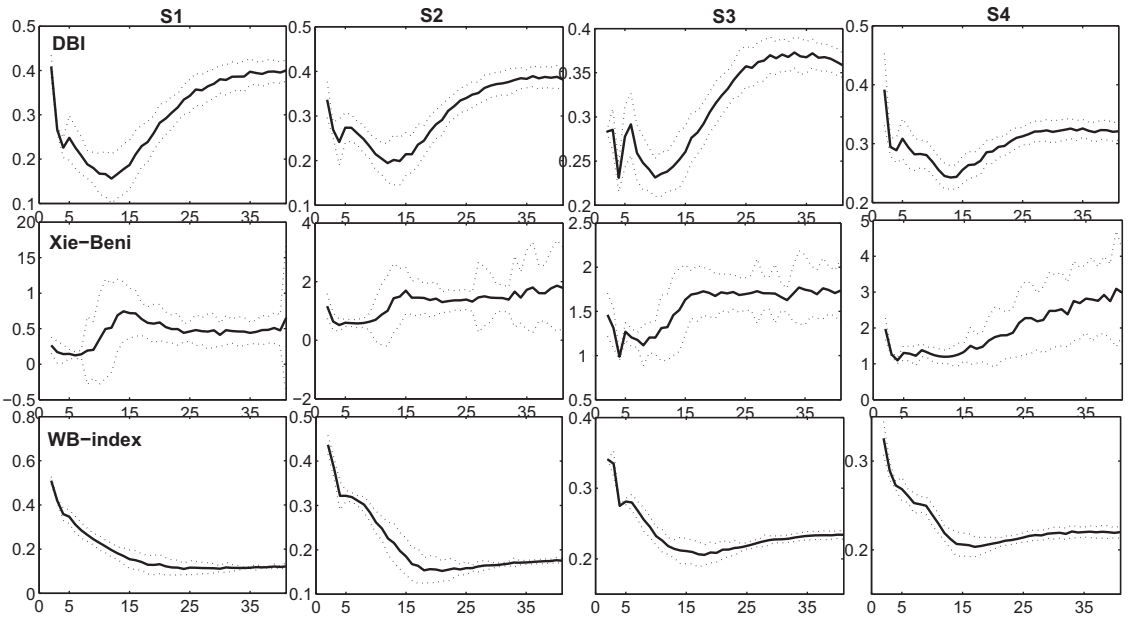


**Fig. 8.** Validity indices including DBI, Xie–Beni and WB-index with the k-means are repeated 100 times on data sets S1–S4. The solid line is the mean value and the dashed lines are the minimum and maximum values.

the future. DBI and SC work very well for both Touching and pathbased, which contain connected clusters. For this type of data sets, the performance of the WB-index is close to that of the Xu-index. However, the number of clusters determined by the WB-index is larger than those from the Xu-index. Among the three indices, the number of clusters obtained from the CH index is closest to the ground-truth value.

The clustering algorithms such as k-means and RS are suitable for Gaussian-like data in general. Therefore, the performance of indices on Gaussian-like data is much better than that on shaped data (see Table 4). For large data sets such as birch1, most of the validity indices produce cluster numbers close to 100; however, none of them gives exactly 100 as the number of clusters. The explanation may be that the data set birch1 contains the clusters in a regular grid structure. Among the sum-of-squares indices, the Xu index and the WB-index have similar performances. The Calinski–Harabsz, however, has worse results than the WB-index and the

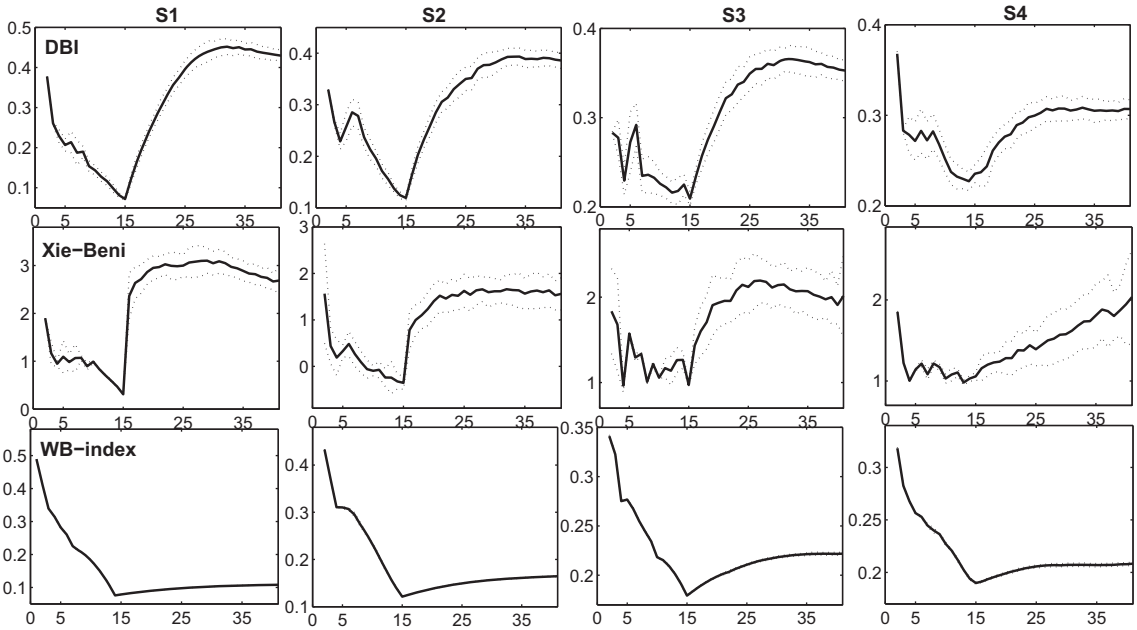


**Fig. 9.** Validity indices including DBI, Xie–Beni and WB-index with the RS are repeated 100 times on data sets S1–S4. The solid line is the mean value and the dashed lines are the minimum and maximum values.

**Table 3**
Mean value of the determined number of clusters by variants of cluster validity indices with RS on shape data.

| | Touching | pathbased | Compound | Aggregation |
|---|---|---|---|---|
| $M^*$ | 2 | 3 | 6 | 7 |
| $M_{max}$ | 8 | 17 | 19 | 28 |
| Ball & Hall | 5.0 | 8.8 | 7.7 | 4.4 |
| CH | 2.0 | 2.0 | 3.0 | 15.6 |
| Hartigan | 5.0 | 3.0 | 3.0 | 3.0 |
| Krzanowski–Lai | 2.0 | 2.3 | 5.8 | 9.6 |
| Xu-index | 5.0 | 16.9 | 18.3 | 26.3 |
| WB index | 5.9 | 16.9 | 18.4 | 26.3 |
| Dunn | 6.7 | 12.6 | 3.0 | 20.4 |
| DBI | 2.1 | 3.0 | 3.0 | 4.0 |
| SC | 2.0 | 3.0 | 3.0 | 4.0 |
| SCI | 3.0 | 3.0 | 3.0 | 4.0 |
| Xie–Beni | 4.2 | 2.7 | 3.8 | 5.0 |
| BIC | 5.7 | 3.0 | 3.0 | 4.0 |

**Table 4**
Mean values of the determined number of clusters by variants of cluster validity indices with RS on Gaussian-like data.

| | R15 | D31 | S1 | S2 | S3 | S4 | birch1 |
|---|---|---|---|---|---|---|---|
| $M^*$ | 15 | 31 | 15 | 15 | 15 | 15 | 100 |
| $M_{max}$ | 24 | 56 | 70 | 70 | 70 | 70 | 316 |
| Ball & Hall | 13.7 | 4.3 | 15.0 | 4.0 | 4.0 | 6.7 | 4 |
| CH | 15.9 | 32.1 | 15.0 | 15.0 | 2.2 | 14.2 | 104.0 |
| Hartigan | 10.4 | 5.3 | 15.0 | 6.7 | 4.0 | 3.0 | 4.0 |
| Krzanowski–Lai | 13.8 | 22.1 | 52.5 | 38.7 | 36.1 | 35.5 | 81.5 |
| Xu-index | 15.2 | 31.5 | 15.0 | 15.0 | 15.0 | 15.0 | 101.5 |
| WB index | 15.2 | 31.5 | 15.0 | 15.0 | 15.0 | 15.0 | 101.5 |
| Dunn | 6.8 | 38.5 | 14.3 | 16.5 | 40.5 | 34.3 | 109.5 |
| DBI | 12.1 | 30.2 | 15.0 | 14.7 | 11.2 | 13.8 | 99.0 |
| SC | 15.1 | 31.1 | 15.0 | 15.0 | 15.0 | 15.0 | 100.0 |
| SCI | 9.3 | 30.8 | 15.0 | 15.0 | 15.1 | 19.4 | 100.0 |
| Xie–Beni | 12.7 | 30.5 | 15.0 | 14.8 | 8.3 | 12.4 | 98.0 |
| BIC | 7.8 | 16.8 | 15.0 | 8.7 | 4.0 | 15.8 | 4.0 |

Xu-index. The CH index is more sensitive to the degree of overlapping of data sets, such as S3 and S4. The Xu-index, WB-index and SC have good performance for this type of data set compared with other indices.

For real data, the indices give more diverse results; their performance depends strongly on the data sets. For instance, the WB-index is the only index working for Iris, but it does not work for wine, control, image and yeast see Table 5. The numbers of clusters determined by the WB-index for real data sets are smaller than those by the Xu-index in general. Comparing the WB-index and the CH index, the WB-index is similar to the CH index except for the Iris data.

From the experimental results, it can be concluded that the sum-of-squares based indices are mostly fit for Gaussian-like data sets. For shape-data and real data, the performance of the indices also depends on the clustering algorithm, whether the structure of data sets is well detected by the algorithm. The WB-index, compared to the CH index and the Xu-index works quite similarly as the other

**Table 5**
Mean values of the determined number of clusters by variants of cluster validity indices with RS on real data.

| | Iris | Wine | Control | Image | Wdbc | Yeast |
|---|---|---|---|---|---|---|
| $M^*$ | 3 | 3 | 6 | 7 | 2 | 10 |
| $M_{max}$ | 12 | 13 | 24 | 48 | 23 | 38 |
| Ball & Hall | 5.0 | 4.0 | 5.1 | 4.0 | 4.5 | 8.0 |
| Calinski–Harabsz | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 5.0 |
| Hartigan | 5.0 | 3.0 | 3.0 | 4.0 | 13.0 | 4.0 |
| Krzanowski–Lai | 2.0 | 4.2 | 2.6 | 2.0 | 2.0 | 28.4 |
| Xu-index | 12.0 | 13.0 | 24.0 | 48.0 | 23.0 | 38.0 |
| WB index | 3.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 |
| Dunn | 10.3 | 12.0 | 19.5 | 36.7 | 20.0 | 33.2 |
| DBI | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 5.0 |
| SC | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 5.0 |
| SCI | 2.0 | 2.0 | 2.0 | 2.4 | 2.4 | 5.0 |
| Xie–Beni | 2.0 | 2.0 | 6.0 | 2.0 | 2.0 | 5.0 |
| BIC | 5.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.2 |

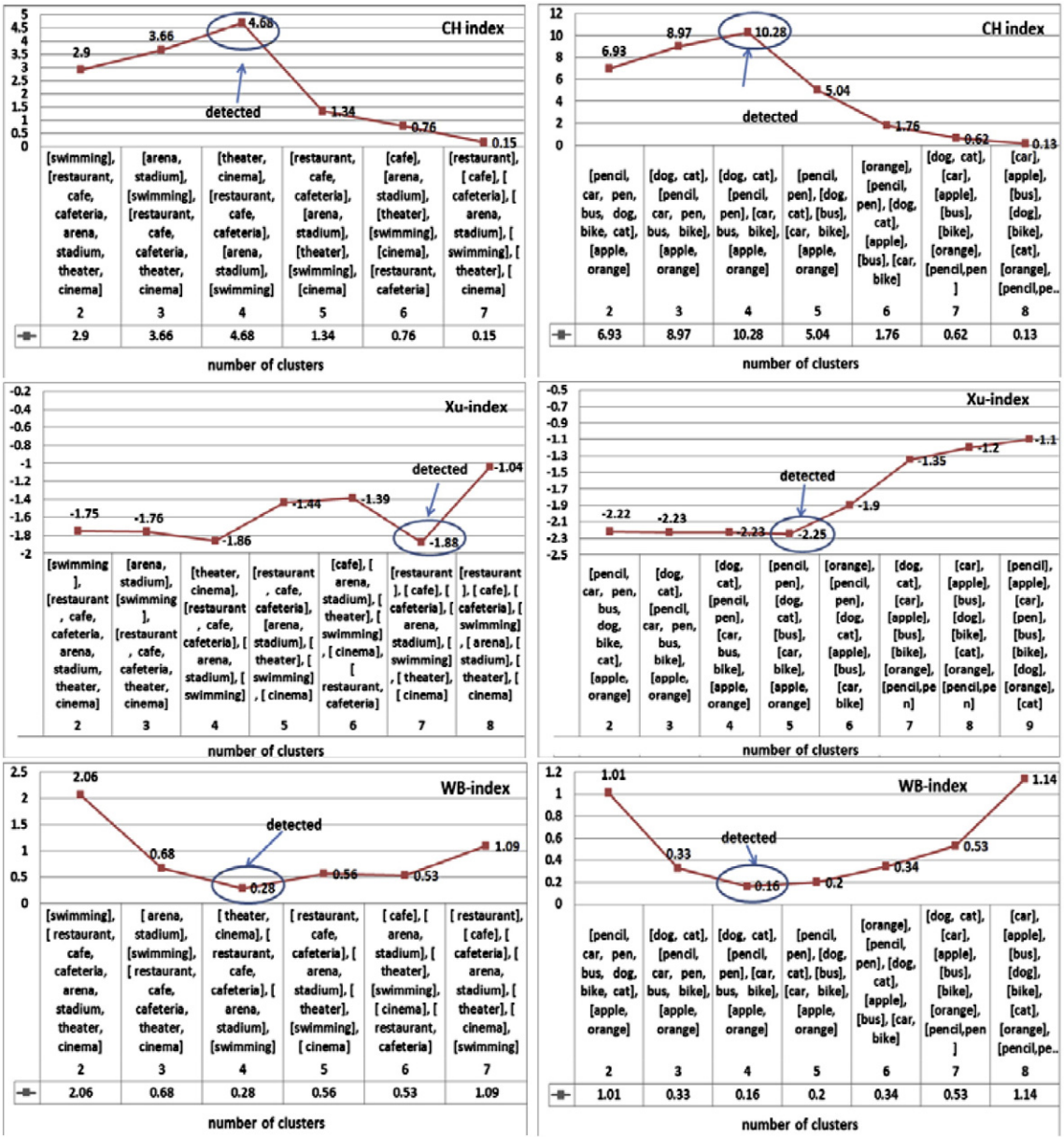


**Fig. 10.** The three sum-of-squares based indices on two manually generated data sets. The figures in the left column are the results for data 1 and the figures in the right column are the results for data 2.

two indices. However, the WB-index works slightly better among the three indices in general. It detects the correct result for eight out of 17 data sets and provides the most correctly determined number of clusters among the sum-of-squares indices and other indices (see Tables 3, 4 and 5). The CH index is more affected by the data size $N$ and is more sensitive to the degree of overlapping of data sets than the WB-index. For two-dimensional data, the Xu-index is remarkably close to the WB-index. However, it rarely detects clear minimum value for high dimensional, real-type data.

### 5.2. Comparison of indices for automatic keywords categorization

Three data sets are involved in the experiment. Two data sets (see Fig. 10) are aggregated manually, while four are humanly judged as the proper number of clusters for both data. The other data is collected from a project *MOPSI*,[4] which includes various location-tagged data such as services, photos and routes. Each service includes a set of keywords to describe what it is. In all, 378 texts were

[4] http://cs.uef.fi/mopsi.

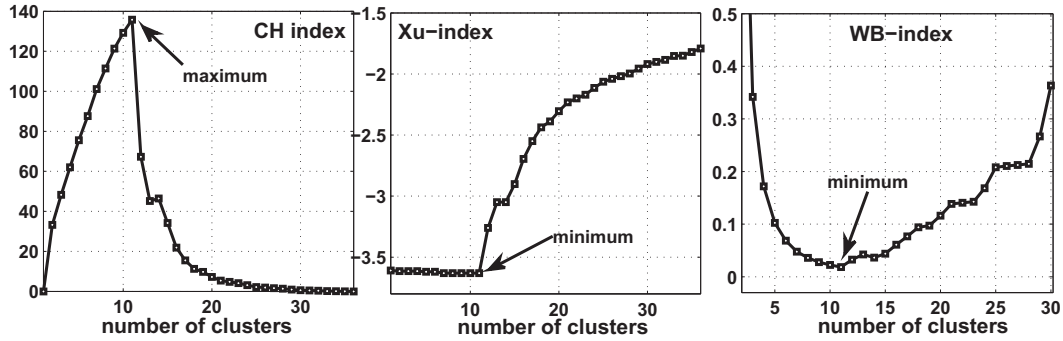


**Fig. 11.** The number of clusters V.S. the three sum-of-squares based.

collected after simple cleaning and 36 of these texts are tested in the experiments. The ground truth for the data sets are based on human judgment.

Even human judgment often differs from person to person in terms of the categorizations and the number of clusters; the clustering algorithm and validity indices can suggest a potentially appropriate categorization and provide a possible number of clusters as a guideline.

We studied the sum-of-squares based indices, CH index, Xu-index and WB-index for automatic keywords categorization on the data sets. To obtain the semantic similarity among keywords, we used JAVA API provided by WordNet 3.0. Hierarchical clustering algorithm is used in the test.

The number of clusters determined by three sum-of-squares based indices on two manually generated data sets are shown in Fig. 10. Four is determined by the CH index and the WB-index for data 1 (left column). For the Xu-index, the results four and seven are close. However, seven is detected with a minimum value. Similarly, four is determined by the CH index and WB-index on data 2 (right column). The Xu-index detects five as the number of clusters with a minimum value. The correctness of the categorization on both data sets is judged by human and we believe that the categorization of four as the number of groups mostly matches users' expectation.

The three indices are also compared with the mopsi data in Fig. 11. There is a clear maximum value of the CH index at 11. The values of the Xu-index on the number of clusters from 7 to 11 are exactly the same. Therefore, it is difficult to recognize which number is the minimum value for the potential number of groups. For the WB-index, the minimum value is detected when the number of clusters is 11. In general, the performance of the indices on the mopsi data is similar as that on the manually generated data. Comparing the result from the indices to the ground truth from human judgment, the indices cannot provide the exactly correct number of clusters; however, they are able to provide suggestions for users. In some real applications, the suggestions might be helpful to the clustering algorithms.

## 6. Conclusions

In this paper, we revisit a sum-of-squares based index, the WB-index. This index is designed to reach its minimum value when the appropriate number of clusters is achieved. There are two similar sum-of-squares based indices, which are the CH and the Xu-index. To study the difference among the three indices, we perform an analysis of the relation between the WB-index and the other two indices. It is shown that the CH index is affected by the data size ($N$) and high levels of cluster overlap. For high dimensional data where $D > 2$, the Xu-index does not work as well as for data with $D \leq 2$. Furthermore, a systematic experiment on 12 internal with two clustering algorithms and variants of data sets was conducted to study the effect of clustering algorithms and data sets on the indices. The number of clusters is used for validating the result.

According to the experimental result, a good and stable clustering algorithm should be selected for obtaining correct number of clusters from the indices. Sum-of-squares based indices work well for Gaussian-type data, although most of them cannot provide a global minimum or maximum point for the correct number of clusters. The WB-index works slightly better than the other two indices. The sum-of-squares indices (e.g., the WB index) are also more stable than indices employing min–max functions (e.g., DBI).

In addition to the comparisons of indices, we introduce a procedure for performing automatic keywords categorization, where texts with multiple keywords are considered and extensions of three sum-of-squares based indices are employed for determining the number of groups. The indices are compared in the experiments and they are shown to be valid for automatic keywords categorization.

## Acknowledgments

One of the authors is sponsored by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation Committee of China under contract no. 61202382.

## References


[1] R. Caruana, M. Elhawary, N. Nguyen, C. Smith, Meta clustering, Sixth Int. Conf. on Data Mining (ICDM'06), 2006, pp. 107–118.
[2] W. Wang, Y. Zhang, On fuzzy cluster validity indices, Fuzzy Sets Syst. 158 (19) (2007) 2095–2117.
[3] G. Ball, D. Hall, ISODATA, a novel method of data analysis and pattern classification, Menlo Park, Calif, Stanford Research Institute, 1965.
[4] J. Hartigan, Clustering algorithms, John Wiley & Sons, Inc., New York, NY, USA, 1975.
[5] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat. 3 (1974) 1–27.
[6] L. Xu, Bayesian ying-yang machine, clustering and number of clusters, Pattern Recogn. Lett. 18 (1997) 1167–1178.
[7] Q. Zhao, M. Xu, P. Fränti, Sum-of-square based cluster validity index and significance analysis, Proc. of the 17th Int. Conf. on Adaptive and Natural Computing Algorithms, 2009, pp. 313–322.
[8] J. Dunn, Well separated clusters and optimal fuzzy partitions, J. Cybern. 4 (1974) 95–104.
[9] D. Davies, D. Bouldin, Cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (2) (1979) 95–104.
[10] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 13 (8) (1991) 841–847.
[11] M. Halkidi, M. Vazirgiannis, Clustering validity assessment: finding the optimal partitioning of a data set, Proc. of the 2001 IEEE Int. Conf. on Data Mining (ICDM'01), 2001, pp. 187–194.
[12] S. Still, W. Bialek, How many clusters? An information theoretic perspective, Neural Comput. 16 (12) (2004) 2483–2506.
[13] D. Pelleg, A. Moore, X-means: extending K-means with efficient estimation of the number of clusters, Proc. of the 17th Int. Conf. on Machine Learning, 2000, pp. 727–734.
[14] M. Zoubi, M. Rawi, An efficient approach for computing silhouette coefficients, J. Comput. Sci. 4 (3) (2008) 252–255.
[15] G. Milligan, M. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika 50 (1985) 159–179.
[16] E. Dimitriadou, S. Dolnicar, A. Weingassel, An examination of indexes for determining the number of clusters in binary data sets, Psychometrika 67 (1) (2002) 137–160.
[17] J. Wu, H. Xiong, J. Chen, Adapting the right measures for k-means clustering, 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'09), 2009, pp. 877–886.
[18] J. Wu, J. Chen, H. Xiong, M. Xie, External validation measures for k-means clustering: a data distribution perspective, Expert Syst. Appl. 36 (3) (2009) 6050–6061.
[19] J. Handl, J. Knowles, D. Kell, Computational cluster validation in post-genomic data analysis, Bioinformatics 21 (15) (2005) 3201–3212.
[20] P. Fränti, J. Kivijärvi, Randomised local search algorithm for the clustering problem, Pattern. Anal. Applic. 3 (4) (2000) 358–369.
[21] P. Fränti, M. Tuononen, O. Virmajoki, Deterministic and randomized local search algorithms for clustering, IEEE Int. Conf. on Multimedia and Expo, 2008. 837–840.
[22] H.L. Capitaine, C. Frelicot, On selecting an optimal number of clusters for color image segmentation, Int. Conf. on, Pattern Recognition (ICPR'10), 2010, pp. 3388–3391.
[23] H. Yan, K. Chen, L. Liu, J. Bae, Determining the best k for clusteirng transactional datasets: a coverage density-based approach, Data Knowl. Eng. 68 (2009) 28–48.
[24] G. Mecca, S. Raunich, A. Pappalardo, A new algorithm for clustering search results, Data Knowl. Eng. 62 (2007) 504–522.
[25] D. Ingaramo, D. Pinto, P. Rosso, M. Errecalde, Evaluation of internal validity measures in short-text corpora, CICLing'08, 2008. 555–567.
[26] X. Ni, X. Quan, Z. Lu, W. Liu, B. Hua, Short text clustering by finding core terms, Knowl. Inf. Syst. 27 (3) (2011) 345–365.
[27] X. Yan, J. Guo, S. Liu, X. Cheng, Y. Wang, Clustering short text using Ncut-weighted non-negative matrix factorization, Proc. of the 21st ACM Intl. Conf. on Information and Knowledge Management, 2012, pp. 2259–2262.
[28] P. Shrestha, C. Jacquin, B. Daille, Clustering short text and its evaluation, Computational Linguistics and Intelligent Text Processing, 2012, pp. 169–180.
[29] S. Banerjee, K. Ramanathan, A. Gupta, Clustering short texts using Wikipedia, Proc. of the 30th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2007, pp. 787–788.
[30] X. Hu, X. Zhang, C. Lu, E. Park, X. Zhou, Exploiting Wikipedia as external knowledge for document clustering, Proc. of the 15th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2009, pp. 389–396.
[31] D. Bollegala, Y. Matsuo, M. Ishizuka, Measuring semantic similarity between words using web search engines, Proc. of the 16th Intl. Conf. on World Wide Web, 2007, pp. 757–766.
[32] C. Chen, F. Tseng, T. Liang, An integration of WordNet and fuzzy association rule mining for multi-label document clustering, Data Knowl. Eng. 69 (2010) 1208–1226.
[33] W. Krzanowski, Y. Lai, A criterion for determining the number of groups in a data set using sum-of-squares clustering, Biometrics 44 (1) (1988) 23–34.
[34] Q. Zhao, M. Xu, P. Fränti, Knee Point Detection on Bayesian Information, Criterion, ICTAI'08, 2008. 431–438.
[35] S. Salvador, P. Chan, Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms, Proc. of the 16th IEEE Int. Conf. on Tools with Artificial Intelligence, 2004, pp. 576–584.
[36] J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, Proc. of Int. Conf. on Research in Computational Linguistics, 1997, pp. 19–33.
[37] Q. Zhao, M. Rezaei, H. Chen, P. Fränti, Keyword clustering for automatic categorization, IEEE Int. Conf. on Pattern Recognition, 2012, pp. 2845–2848.


**Qinpei Zhao** received the B.Sc. degree in Automation Technology from Xi'dian University, Xi'an, China in 2004. She received the M.Sc. degree in Pattern Recognition and Image Processing in Shanghai Jiaotong University, Shanghai, China in 2007. She obtained the Ph.D. degree in Computer Science at the University of Eastern Finland in 2012. Her current research is focused on clustering algorithm and multimedia processing.

**Pasi Fränti** received his MSc and Ph.D. degrees from the University of Turku, 1991 and 1994 in Science. Since 2000, he has been a professor of Computer Science at the University of Eastern Finland. He has published 64 journals and 142 peer review conference papers, including 13 IEEE transaction papers. His current research interests include clustering algorithms and location-aware applications. He has supervised 19 Ph.D. theses and is the head of the East Finland doctoral program in Computer Science & Engineering (ECSE).