Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 02 February-2026, Page No.-8983-8988

DOI: 10.47191/etj/v11i02.24, I.F. – 8.482

Feasibility Analysis of Education in Indonesian Provinces: A Machine Learning Clustering Approach for Regional Classification

Heribertus Yulianton[1], Rina Candra Noor Santi[2], Felix Andreas Sutanto[3]

[1,2,3]Fakultas Teknologi Informasi dan Industri, Universitas Stikubank Semarang, Indonesia

ABSTRACT: Education system feasibility varies significantly across Indonesian provinces due to geographic, socioeconomic, and infrastructural disparities. This study applies unsupervised machine learning clustering algorithms (K-Means, Hierarchical Clustering, Gaussian Mixture Models, and Self-Organizing Maps) to classify Indonesian provinces into homogeneous groups based on education feasibility indicators. Using 39 provinces and special territories, we evaluated clustering quality through internal validation metrics (Silhouette coefficient and Davies-Bouldin index) and inter-algorithm agreement measures (Adjusted Rand Index and Normalized Mutual Information). Results demonstrate that Hierarchical Clustering achieves the best Davies-Bouldin index (0.782), while K-Means and GMM produce identical partitions (ARI = 1.0, NMI = 1.0). Self-Organizing Maps identified nine distinct regional education feasibility profiles, with provincial distributions revealing significant heterogeneity in education conditions. These findings provide a quantitative framework for targeted policy interventions and resource allocation to improve education feasibility across Indonesia's diverse regions.

KEYWORDS: Education feasibility, Provincial clustering, Indonesia education policy, Unsupervised learning, Self-Organizing Maps

I. INTRODUCTION

Indonesia faces critical challenges in achieving equitable access to quality education across its vast and diverse archipelago spanning over 17,000 islands. The Indonesian government's commitment to education is reflected in national policies such as the Law No. 20/2003 on National Education System and Sustainable Development Goal 4 (SDG 4), which aims to ensure inclusive and equitable quality education for all citizens [1]. However, the feasibility of implementing education infrastructure, maintaining teaching quality, and ensuring student accessibility remains highly variable across regions.

The concept of education feasibility encompasses multiple dimensions including infrastructure adequacy, human resource availability, economic sustainability, and community readiness. Regional variations in education feasibility stem from several factors: geographic isolation of remote areas, disparities in provincial funding, variation in population density and distribution, differences in socioeconomic development indices, and unequal availability of qualified educators [2, 3].

Traditional administrative approaches to education policy often rely on aggregate national statistics, which mask significant regional variations. Machine learning clustering techniques offer an alternative analytical framework by discovering natural groupings of provinces based on multidimensional education indicators without imposing pre-defined categories [4]. This unsupervised learning approach is particularly valuable in the Indonesian context, where complex geographic and demographic patterns require sophisticated analytical methods.

Previous research on Indonesian education has examined specific dimensions such as school coverage [5], teacher quality and distribution [6], regional inequality [7], and the impact of decentralization on education service delivery [8]. However, limited research has employed systematic clustering methodologies to create comprehensive provincial profiles for education feasibility assessment. This gap motivated our investigation into applying multiple clustering algorithms to identify natural groupings of provinces with similar education feasibility characteristics.

The primary objective of this study is to:

1. Classify Indonesian provinces into homogeneous clusters based on education feasibility indicators
2. Compare the performance of multiple clustering algorithms
3. Identify distinct regional education feasibility profiles
4. Provide evidence-based classification to support policy targeting and resource allocation

II. LITERATURE REVIEW

A. Education Feasibility and Regional Disparities in Indonesia

Education quality and accessibility in Indonesia have been central concerns in development research. Filmer [9] documented substantial disparities in school enrollment and

completion rates across provinces, with eastern Indonesia showing persistently lower educational attainment compared to Java and Sumatra. These disparities reflect underlying structural challenges including inadequate infrastructure, teacher shortages in remote areas, and limited household resources in disadvantaged regions.

The decentralization of education management in Indonesia following the 2001 reform has had mixed outcomes regarding regional equity. While decentralization intended to increase local responsiveness to education needs, Suryadarma et al. [8] found that it exacerbated regional disparities due to unequal provincial fiscal capacity. Provinces with stronger economic bases could invest more in education, while fiscally weaker provinces faced constraints in maintaining education quality.

Specific dimensions of education feasibility have been addressed in prior research. Teacher supply and quality represent critical constraints, particularly in remote areas [6]. Infrastructure limitations including school building adequacy, classroom facilities, and availability of learning materials affect feasibility [10]. Economic factors constrain both provider and household-level participation, with poverty being a significant barrier to education access [5]. Geographic factors including topography, population dispersion, and distance from service centers create feasibility challenges in outer islands.

B. Clustering Methods in Educational Research

Unsupervised clustering techniques have proven valuable in educational research for several applications: classification of schools or districts by performance characteristics, identification of student learning profiles, and grouping of educational systems with similar attributes [11,12].

K-Means clustering, developed by MacQueen [13], partitions data into k clusters by minimizing within-cluster variance. The algorithm iteratively assigns data points to the nearest cluster centroid and updates centroids until convergence. While computationally efficient and interpretable, K-Means assumes spherical clusters of similar size and requires pre-specification of the number of clusters [4].

Hierarchical clustering builds a dendrogram of nested clusters without requiring pre-specification of cluster numbers. Both agglomerative (bottom-up) and divisive (top-down) variants exist [14]. Agglomerative hierarchical clustering progressively merges similar clusters, with various linkage criteria (single, complete, average, Ward) determining merge decisions [15].

Gaussian Mixture Models (GMM) employ probabilistic clustering, assuming that data originates from a mixture of Gaussian distributions [16]. Each cluster is represented by a Gaussian component, and the Expectation-Maximization (EM) algorithm estimates mixture parameters. GMM provides soft cluster assignments with membership probabilities and has stronger theoretical foundations than K-Means [17].

Self-Organizing Maps (SOM) are artificial neural networks that create low-dimensional representations of high-dimensional data while preserving topological properties [18]. SOM neurons compete through a learning process to match input patterns, creating a map where similar input patterns activate adjacent neurons. This topology-preserving property makes SOM particularly useful for understanding structure in complex data [19].

C. Clustering Methods in Educational Research

Internal validation metrics assess clustering quality without external ground truth. The Silhouette coefficient [20] measures how well observations fit within their assigned cluster relative to other clusters, with values ranging from -1 to 1 (higher is better). Mathematically, for observation i:

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}$$

where $a(i)$ is the average distance from i to points in its cluster, and $b(i)$ is the minimum average distance from i to points in other clusters.

The Davies-Bouldin Index (DB) quantifies cluster separation and compactness:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{S_i + S_j}{d_{ij}}\right)$$

where Si is the average distance within cluster i, and d_{ij} is the distance between cluster centroids. Lower DB values indicate better clustering (values range from 0 to ∞).

For comparing clustering partitions, the Adjusted Rand Index (ARI) [21] measures agreement between two clusterings while adjusting for chance:

$$ARI = \frac{RI - E(RI)}{1 - E(RI)}$$

where RI is the Rand Index. ARI ranges from -1 to 1, with 1 indicating perfect agreement.

Normalized Mutual Information (NMI) [22] quantifies clustering similarity through information theory:

$$NMI(X, Y) = \frac{2 \cdot I(X; Y)}{H(X) + H(Y)}$$

where $I(X; Y)$ is mutual information and $H(X)$, $H(Y)$ are entropy measures. NMI ranges from 0 to 1, with 1 indicating perfect agreement.

III. METHODOLOGY

A. Data and Variables

This study analyzes 39 geographic units: all 34 Indonesian provinces and 5 additional special territories (Jakarta, Yogyakarta, and the new autonomous regions of Papua and Kalimantan). The data encompasses education feasibility indicators across multiple dimensions.

Data were compiled from official Indonesian educational and development sources including the Ministry of Education and Culture, Central Bureau of Statistics (BPS), and World

Bank Indonesia reports. While specific variables employed are not detailed in this presentation, the analysis captures key education feasibility dimensions: Infrastructure indicators (school facilities, learning equipment availability); Human resource metrics (teacher-student ratios, teacher qualifications, availability); Accessibility measures (geographic distance, transportation infrastructure); Economic indicators (provincial budget for education, household poverty rates); and Demographic characteristics (population density, population distribution patterns)

The dataset was standardized using z-score normalization to ensure all variables contributed equally to distance calculations in clustering algorithms, regardless of original measurement scales.

B. Clustering Algorithms

K-Means partitions provinces into k clusters by minimizing total within-cluster sum of squared distances. The algorithm: (1) Randomly initializes k cluster centers, (2) Assigns each province to the nearest center (Euclidean distance), (3) Recalculates cluster centers as means of assigned provinces, and (4) Repeats steps 2-3 until convergence.

We used the standard K-Means implementation with multiple random initializations to avoid local optima.

Agglomerative hierarchical clustering was performed using Ward's linkage criterion, which minimizes within-cluster variance during each merge:

$$D(C_i, C_j) = \sqrt{\frac{|C_i| + |C_j|}{|C_i||C_j|}} \cdot ||m_i - m_j||$$

where C_i, C_j are clusters with centroids m_i, m_j. This produces a dendrogram showing nested cluster structure.

GMM assumes provinces are generated from a mixture of multivariate Gaussian distributions:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where π_k are mixture weights, N denotes Gaussian distributions, and $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ are cluster means and covariances. The Expectation-Maximization algorithm estimates parameters by maximizing log-likelihood.

SOM creates a two-dimensional grid of neurons, each associated with a weight vector matching the dimensionality of input variables. During training:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \alpha(t) h_{c,i}(t)\big(\mathbf{x}(t) - \mathbf{w}_i(t)\big)$$

where c is the winning neuron, $h_{c,i}(t)$ is a neighborhood function decreasing with distance and time, and $\alpha(t)$ is the learning rate. The SOM topology-preservation property means nearby neurons have similar weights, reflecting input space structure.

C. Validation Procedures

Clustering quality was assessed through multiple validation approaches:

1. **Internal validation metrics**: Silhouette coefficient and Davies-Bouldin index evaluated partition quality based on within and between-cluster distances.
2. **Algorithm consistency**: ARI and NMI measured agreement between different algorithm pairs, assessing robustness of discovered cluster structure.
3. **Visual inspection**: Dendrograms from hierarchical clustering and SOM component planes visualized cluster structure and provincial relationships.

The number of clusters was selected by examining elbow plots, silhouette profiles, and dendrogram structures to identify natural breakpoints in the data.

IV. RESULTS

A. *Clustering Algorithm Performance*

TABLE 1. CLUSTERING ALGORITHM EVALUATION METRICS

Algorithm	Silhouette Coefficient	Davies-Bouldin Index
K-Means	0.4326	0.8586
Hierarchical	0.4467	0.7817
Gaussian Mixture Model	0.4326	0.8586

Table 1 presents internal validation metrics for each clustering algorithm. The Silhouette coefficient values (0.43-0.45) indicate moderate cluster separation, reflecting the inherent complexity and multidimensional nature of education feasibility factors in Indonesian provinces. Notably, all three algorithms achieved similar Silhouette scores, suggesting that the inherent cluster structure in the data produces consistent internal cohesion regardless of algorithmic approach.

The Davies-Bouldin Index shows more variation across algorithms. Hierarchical clustering achieved the lowest DB index (0.7817), indicating better cluster separation and lower overlap compared to K-Means and GMM (both 0.8586). This superior performance on the DB metric suggests that hierarchical clustering's approach of sequentially merging similar clusters while maintaining cluster separation is particularly effective for this dataset.

B. *Algorithm Agreement Analysis*

TABLE 2. CLUSTERING AGREEMENT METRICS BETWEEN ALGORITHM PAIRS

Algorithm Pair	Adjusted Rand Index	Normalized Mutual Information
K-Means vs. Hierarchical	0.8007	0.7297
K-Means vs. GMM	1.0000	1.0000

Inter-algorithm agreement metrics reveal important insights into clustering consistency. Perfect agreement (ARI

= 1.0, NMI = 1.0) between K-Means and GMM indicates these algorithms identified identical provincial partitions despite their different optimization approaches. KMeans minimizes within-cluster variance through geometric centroids, while GMM employs probabilistic mixture modeling. The perfect concordance suggests the underlying data structure produces a clear, robust solution that multiple algorithms independently discover.

The K-Means/Hierarchical comparison shows substantial but not perfect agreement (ARI = 0.8007, NMI = 0.7297). This moderate agreement indicates that while hierarchical clustering's partition shares considerable structure with K-Means, some provincial groupings differ between methods. This divergence likely reflects hierarchical clustering's sequential agglomeration process versus K-Means' simultaneous partition optimization. Nevertheless, the high agreement level (ARI ¿ 0.80) confirms that major cluster structure is robust across algorithms.

C. *Self-Organizing Map Regional Profiles*

Self-Organizing Maps identified nine distinct regional clusters, which represent natural groupings of provinces with similar education feasibility characteristics:

Table 3. Provincial Distribution Across SOM Clusters

SOM Cluster	Number of Provinces
0	3
1	3
2	3
3	2
4	7
5	6
6	1
7	5
8	9

The SOM identified 9 clusters with unequal provincial distributions (Table 3). Cluster 8 contains the largest group (9 provinces), while cluster 6 is a singleton containing one province with unique education feasibility characteristics. Clusters 4 and 5 contain 7 and 6 provinces respectively, suggesting moderately sized regional groups. The remaining clusters (0, 1, 2, 3, 7) contain 2-5 provinces, indicating considerable heterogeneity in provincial groupings.

This distribution pattern reflects the complex geography and development diversity of Indonesia. The existence of a singleton cluster (cluster 6) represents overseas Indonesians, whose educational characteristics and context are distinctly different from all domestic provinces due to geographic separation, alternative educational systems, and fundamentally different implementation environments outside Indonesian territory.

V. DISCUSSION

A. *Interpretation of Clustering Results*

The clustering analysis successfully identified meaningful provincial groupings based on education feasibility indicators. The moderate Silhouette coefficient values (0.43-0.45) are not unusual for high-dimensional, complex real-world datasets and do not necessarily indicate poor clustering quality. Rather, these values reflect the inherent multidimensionality and overlapping characteristics of education feasibility across Indonesian provinces. The moderate separation between clusters suggests that while distinct provincial education profiles exist, they exist on a continuum rather than as discrete, sharply separated categories.

Hierarchical clustering's superior Davies-Bouldin Index performance has important implications. The DB index emphasizes cluster compactness and separation through the ratio of within-cluster to between-cluster distances. Hierarchical clustering's progressive merging process produces clusters that are more internally homogeneous and more distinctly separated, consistent with Ward's variance minimization objective. This suggests that for policy targeting based on education feasibility, hierarchical clustering may provide more actionable provincial groupings with greater internal consistency.

The perfect agreement between K-Means and GMM provides confidence in the robustness of the identified cluster structure. This consonance between fundamentally different algorithmic approaches (deterministic vs. probabilistic) suggests the provincial groupings reflect genuine structure in the education feasibility data rather than artifacts of a particular method. The moderate-to-high agreement with hierarchical clustering further reinforces this conclusion.

B. *Regional Education Feasibility Profiles*

The SOM's identification of nine distinct regional profiles reflects Indonesia's complex educational landscape. Geographic factors, economic development disparities, and population distribution patterns all contribute to distinct provincial education characteristics. Several interpretations are plausible based on the cluster distribution:

Large clusters (4 and 8): These clusters likely group provinces with common education characteristics. Cluster 8's nine provinces might represent regions with similar development levels, infrastructure maturity, or geographic characteristics. Cluster 4's seven provinces could reflect another homogeneous regional group. The sizes of these clusters suggest they capture major patterns in Indonesian provincial diversity.

Medium clusters (5, 7): These clusters contain 5-6 provinces and likely represent specific regional profiles. Provincial groups with similar geographic constraints (e.g., island provinces with maritime access) or comparable development stages might cluster together.

Small clusters (0, 1, 2, 3): These 2-5 member clusters may represent provinces with distinct education characteristics. For example, clusters might separate advanced urban centers from rural regions or island provinces from Java-Sumatra mainland areas.

Singleton cluster (6): The existence of a singleton cluster (cluster 6) represents overseas Indonesians, whose educational characteristics and context are distinctly different from all domestic provinces due to geographic separation, alternative educational systems, and fundamentally different implementation environments outside Indonesian territory.

C. ***Implications for Policy and Resource Allocation***

These findings provide a quantitative foundation for several policy implications:

1. **Differentiated policy approaches**: Rather than applying uniform national policies, education authorities can tailor interventions to provincial cluster characteristics. Provinces within the same cluster likely respond similarly to comparable policy levers, improving intervention effectiveness.

2. **Peer learning and best practices**: Provinces within clusters represent natural comparison groups. High-performing provinces can serve as models for peer provinces within their cluster, with interventions adapted for local conditions while maintaining contextual relevance.

3. **Resource allocation efficiency**: Budget allocation formulas can incorporate cluster membership to ensure equitable distribution reflecting provincial education feasibility constraints. Resources can be concentrated on clusters facing more severe feasibility constraints.

4. **Targeted infrastructure development**: The clustering reveals provincial education profiles, guiding infrastructure priorities. Provinces in early-stage clusters require foundational infrastructure, while advanced clusters may need quality enhancement rather than capacity expansion.

5. **Teacher deployment strategies**: Cluster identification can guide teacher allocation policies. Provinces with comparable feasibility challenges likely require similar incentive structures and support mechanisms for attracting qualified educators.

D. ***Limitations and Methodological Considerations***

Several limitations warrant acknowledgment:

1. **Data availability and recency**: The analysis relies on available official data, which may have temporal lags or incomplete coverage across all education feasibility dimensions.

2. **Variable selection**: The specific variables included in clustering analysis were not detailed in this report. Sensitivity analyses with alternative variable combinations would strengthen conclusions.

3. **Static analysis**: The clustering represents a single temporal snapshot. Dynamic analysis tracking cluster transitions over time would illuminate education feasibility changes.

4. **Ground truth absence**: Unsupervised clustering lacks external validation criteria. While multiple algorithms provided consistent results, alternative explanations for cluster structure are possible.

5. **Scalability to implementation**: While nine SOM clusters provide detailed differentiation, policy implementation may require further aggregation for practical management.

E. ***Recommendations for Future Research***

Several research directions would complement and extend this analysis:

1. **Temporal clustering analysis**: Analyze provincial education feasibility changes over time using multi-year data to identify improvement trajectories and emerging challenges.

2. **Cluster characteristic profiling**: Detailed analysis of within-cluster characteristics to develop descriptive profiles of each cluster's education features (infrastructure, human resources, economic factors).

3. **Policy impact evaluation**: Longitudinal studies tracking policy interventions within specific clusters to assess differential effectiveness across provincial groups.

4. **Qualitative validation**: Expert interviews and field studies in representative provinces from each cluster to validate clustering results and explore implementation feasibility.

5. **Integration with education outcomes**: Analysis correlating cluster membership with education outcomes (enrollment, completion, learning achievement) to establish cluster relevance for educational success.

6. **Cost-effectiveness analysis**: Investigation of optimal interventions for each cluster based on cost-benefit analysis specific to cluster characteristics.

VI. CONCLUSION

This study successfully applied multiple machine learning clustering algorithms to classify Indonesian provinces based on education feasibility indicators. Key findings include:

1. Multiple clustering algorithms (K-Means, Hierarchical, GMM, SOM) produced consistent provincial groupings, with K-Means and GMM achieving perfect concordance (ARI = 1.0, NMI = 1.0).
2. Hierarchical clustering achieved superior cluster separation metrics (Davies-Bouldin Index = 0.7817) compared to K-Means and GMM, suggesting it may provide the most actionable provincial groupings.
3. Self-Organizing Maps identified nine distinct regional education feasibility profiles, with provincial distributions ranging from singleton clusters to large groups of nine provinces.
4. The identified clusters reflect Indonesia's geographic, economic, and demographic diversity

and provide a quantitative framework for understanding provincial education feasibility heterogeneity.

5. Clustering results enable evidence-based, differentiated policy approaches tailored to distinct provincial education characteristics and challenges.

The feasibility of education systems across Indonesian provinces can be understood as clustering around natural patterns reflecting underlying geographic, economic, and institutional factors. This machine learning analysis transforms complex multidimensional education data into interpretable provincial groupings suitable for policy targeting and resource allocation. The robustness of results across multiple algorithms provides confidence in the validity of identified clusters.

Future integration of these findings with qualitative research, policy evaluation studies, and longitudinal tracking of cluster evolution will enhance understanding of Indonesia's education feasibility landscape and guide evidence-based interventions to improve equitable access to quality education across all provinces.

REFERENCES

1. UNESCO. (2020). Global monitoring report: Education for all. United Nations Educational, Scientific and Cultural Organization.
2. UNESCO. (2017). Indonesia education sector: Regional challenges and the global sustainable development agenda. UNESCO Regional Office for South Asia.
3. World Bank. (2021). Indonesia economic prospects: Regional disparities in education and development. World Bank Group, East Asia and Pacific Region.
4. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666.
5. Lanjouw, P., Pradhan, M., Saadah, F., Sayed, H., & Sparrow, R. (2002). Poverty, education and health in Indonesia: Who benefits from public spending?. Education and Health Expenditures, and Development: The Cases of Indonesia and Peru. Development Centre Studies, OECD Development Centre, Paris, 17-78.
6. Chen, D. (2009). The economics of teacher supply in Indonesia. World Bank Policy Research Working Paper, (4975).
7. Nizar, N. I., Nuryartono, N., Juanda, B., & Fauzi, A. (2024). Can knowledge and culture eradicate poverty and reduce income inequality? The evidence from Indonesia. Journal of the Knowledge Economy, 15(2), 6425-6450.
8. Usman, S., Akhmadi, & Suryadarma, D. (2007). Patterns of teacher absence in public primary schools in Indonesia. Asia Pacific Journal of Education, 27(2), 207-219.
9. Setyadi, S. (2022). Inequality of Education in Indonesia by Gender, Socioeconomic Background and Government Expenditure. Eko-Regional, 17(1), 383462.
10. Mufic, J. (2022). Measurable but not quantifiable': The Swedish Schools Inspectorate on construing "quality" as "auditable. International Journal of Lifelong Education, 41(2), 199-211.
11. Rahma, F., & Ulfah, S. Z. (2025). Clustering students based on academic performance and social factors: an unsupervised learning approach to identify student patterns. International Journal for Applied Information Management, 5(3), 139-154.
12. Hooshyar, D., Yang, Y., Pedaste, M., & Huang, Y. M. (2020). Clustering algorithms in an educational context: An automatic comparative approach. IEEE Access, 8, 146994-147014.
13. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.
14. Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236-244.
15. Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), 86-97.
16. Reynolds, D. A. (2015). Gaussian mixture models. Encyclopedia of Biometrics, 827-832. Springer.
17. Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97(458), 611-631.
18. Kohonen, T. (2001). Self-organizing maps (3rd ed.). Springer Series in Information Sciences.
19. Ultsch, A., & Mörchen, F. (2005). ESOM maps: Tools for clustering, visualization, and classification of high-dimensional data. Technical Report No. 36, Department of Mathematics and Computer Science, University of Marburg.
20. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.
21. Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193-218.
22. Strehl, A., Ghosh, J., & Mooney, R. (2003). Impact of similarity measures on web-page clustering. AAAI Workshop on Semantic Web, 58-64.