Pendekatan Clustering untuk Ekstraksi Pengetahuan pada Pembangunan Sistem Manajemen Pengetahuan

Dwinta Rahmallah Pulukadang[a,*], Mustafid[b], Farikhin[c]

[a] Mahasiswa Magister Sistem Informasi Universitas Diponegoro
[b] Jurusan Statistik, Fakultas Sains dan Matematika, Universitas Diponegoro
[c] Jurusan Matematika, Fakultas Sains dan Matematika Universitas Diponegoro

Naskah Diterima : 13 Mei 2015; Diterima Publikasi : 13 Juli 2015

Abstract

The importance of knowledge management in any organization encourages the development of a knowledge management system with features that can facilitate knowledge management processes such as storing, organizing, filtering, searching, and most important is the transfer of knowledge. The purpose of this researchis to develop a knowledge management system with a clustering approach for knowledge extraction by using a knowledge of publication writing. This study uses clustering k-means method which is used for cluster knowledge feature where at the same time can help the process of organizing, filtering, browsing and searching knowledge. The results of this research showed that the clustering k-means can be used for knowledge management system with the best value of *purity*= 0,8454 which is found by using k = 20. Clustering approach in the system again can help the process for knowledge searching based on knowledge cluster. This can be proved by carried out 15 times experiments which result in average level of accuracy (precision) about 89.13% and the average rate of completeness (recall)about 85.73 %.

Keywords **:** Knowledge Management System; Extraction; K-means; Clustering

Abstrak

The importance of knowledge management in any organization encourages the development of a knowledge management system with features that can facilitate knowledge management processes such as storing, organizing, filtering, searching, and most important is the transfer of knowledge. The purpose of this researchis to develop a knowledge management system with a clustering approach for knowledge extraction by using a knowledge of publication writing. This study uses clustering k-means method which is used for cluster knowledge feature where at the same time can help the process of organizing, filtering, browsing and searching knowledge. The results of this research showed that the clustering k-means can be used for knowledge management system with the best value of *purity*= 0,8454 which is found by using k = 20. Clustering approach in the system again can help the process for knowledge searching based on knowledge cluster. This can be proved by carried out 15 times experiments which result in average level of accuracy (precision) about 89.13% and the average rate of completeness (recall)about 85.73 %.

Keywords **:** Knowledge Management System; Extraction; K-means; Clustering

1. Pendahuluan

Suatu pengetahuan di dalam organisasi yang termasuk dalam aset tak berwujud (*tacit*), apabila dikelola dan dimanfaatkan dengan baik maka organisasi tersebut dapat mempunyai sumber pengetahuan untuk daya saing bagi organisasi itu sendiri. Namun, pengetahuan tersebut sering hanya dimiliki oleh satu orang. Hal inilah yang menyebabkan sering adanya ketergantungan pengetahuan dari setiap individu apabila tidak adanya suatu referensi atau pengetahuan yang memadai (Kurniawan, 2012). Oleh karena itu, suatu pengetahuan hendaknya perlu didokumentasikan dan tersedia, agar tidak adanya ketergantungan pada individu yang lain apabila individu tersebut sudah tidak berada pada organisasi.

Salah satu sistem untuk mendukung pengelolaan pengetahuan dari suatu organisasi adalah *Sistem Manajemen Pengetahuan* (SMP). Proses yang dilakukan dalam pengelolaan pengetahuan terdiri dari penciptaan, penyimpanan, penelusuran, pemindahan, dan penerapan (Butler *et al*., 2008). Adapun peran Sistem Informasi untuk SMP ini adalah untuk proses penyimpanan, penelusuran dan pemindahan. Hasil dari SMP dapat menjawab pertanyaan 'apa' dan 'bagaimana' (Boahene *et al.*, 2003) yang dapat mendukung proses pembelajaran dan meningkatkan efektivitas organisasi (Maier, 2007).

*) Penulis korespondensi: dwintarahmap@yahoo.com

Publikasi pada universitas merupakan suatu hasil yang dapat membantu universitas mendapatkan keunggulan kompetitif. Perkembangan gagasan dari para pakar (pengetahuan tacit) yang berhubungan dengan penulisan publikasi harusnya dapat dikonversikan ke bentuk eksplisit, terdokumentasi dan dikelola bersama pengetahuan eksplisit yang berbentuk dokumen/arsip. Pengetahuan tersebut juga sebaiknya tersebar secara merata agar dapat menjadi bahan diskusi untuk mendapatkan standar prosedur dan strategi dalam penulisan publikasi ilmiah universitas. Pengetahuan yang terdokumentasi akan terus berkembang dan jumlahnya akan semakin bertambah. Oleh sebab itu, proses penelusuran pengetahuan diperlukan untuk membantu menelusuri setiap pengetahuan yang dibutuhkan.

Pada penelitian ini, dibangun suatu SMP yang didasari dengan penggunaan *data mining* untuk ekstraksi pengetahuan. Salah satu metode data mining yang digunakan adalah *clustering*. Pemilihan *clustering* digunakan untuk proses ekstraksi pengetahuan dan sekaligus sebagai metode yang dipilih karena dapat mendukung fitur dari SMP seperti *knowledge cluster*.

Hasil dari prototipe SMP ini adalah sistem yang dapat mendukung proses pengelolaan pengetahuan seperti penyimpanan, penelusuran, penyebaran untuk mendukung pengambilan keputusan, dan hasil *cluster* pengetahuan dapat mendukung pemberi keputusan dalam hal menentukan pakar untuk setiap *cluster* yang dihasilkan SMP atau mendukung aktifitas lain yang dapat membantu proses perkembangan universitas dalam penulisan publikasi.

2. Kerangka Teori

2.1. Sistem Manajemen Pengetahuan

Pengetahuan terdiri dari pengharapan bersifat kognitif, yang merupakan pengamatan yang mempunyai makna, telah teroganisir, terakumulasi dan tertanam dalam konteks yang diperoleh melalui kesimpulan dari pengalaman dan komunikasi (Maier, 2007). Sistem manajemen pengetahuan (SMP) berperan untuk menggabungkan dan mengintegrasikan fungsi untuk penanganan kontekstual dari pengetahuan eksplisit dan pengetahuan tacit pada seluruh atau bagian dari organisasi yang ditargetkan untuk dilakukan pengelolaan pengetahuan (Maier, 2007). Tujuan utama SMP untuk mendukung dinamika organisasi pembelajaran dan efektivitas organisasi (Maier, 2007).

Terdapat empat proses inti yang harus tersirat dalam Sistem Manajemen Pengetahuan, yang terdiri dari (Babu *et al.,* 2012) :

1. *Capturing* (Menangkap)
   Fase memperoleh data pengetahuan yang berasal dari e-mail, file audio, file text, file digital dan sejenisnya.
2. *Organizining* (Mengatur)
   Data atau informasi yang diperoleh dapat diambil dan digunakan untuk menghasilkan pengetahuan yang berguna. Untuk melakukan proses organisasi ini dapat menggunakan pengindeksan, *Information Retrieval* dan metode lainnya.
3. *Refining* (Menyempurnakan)
   Pada fase ini, *data mining* bisa diimplementasikan.
4. *Trasferring* (Mentransfer)
   Pengetahuan harus disebarkan dan ditransfer dengan menjadikan pengetahuan dapat tersedia untuk pengguna.

2.2. Peran Data Mining dalam Sistem Manajemen Pengetahuan

Data mining bertujuan untuk memenuhi kebutuhan keluaran SMP untuk pengguna. Kebutuhan informasi dinyatakan dalam konteks *data mining* memiliki tingkat yang lebih tinggi mengandung ketidakjelasan dan ketidaklengkapan dari suatu informasi maka perlu diungkapkan dalam konteks *information retrieval* (Baets, 2005).

2.3. Analisis Prapemrosesan

Tahapan prapemrosesan untuk membantu proses penelusuran dalam membangun SMP ini terdiri dari proses *tokenization, stopword removal,* dan *document indexing* (*term weighting*).

A. Tokenization

Tokenization digunakan untuk memecah setiap kalimat dari seluruh dokumen pengetahuan ke dalam kata-kata (*term*) dengan menggunakan pembatas *tab* dan karakter spasi (Darawaty *et al.,* 2010).

B. Stopword Removal

Stopword removal bertugas dalam penghapusan *stopword* yang merupakan langkah penting dalam prapemrosesan. *Stopword* merupakan kata – kata spesifik atau fungsional yang biasanya tidak berisi suatu informasi atau yang tidak diperlukan dalam suatu penelusuran (biasanya adalah kata ganti, kata depan, konjungsi) (Srividhya *et al.,* 2010)

C. Term Weighting

Tiga komponen utama yang mempengaruhi pentingnya *term* dalam dokumen adalah *term frequency* (TF), *Inverse Document Frequency* (IDF) dan normalisasi panjang dokumen. TF dan IDF adalah bobot yang tergantung dalam distribusi setiap kata dalam dokumen. Hal tersebut menerangkan pentingnya suatu kata dalam dokumen. TF.IDF adalah suatu teknik yang menggunakan TF dan IDF untuk menentukan bobot dari istilah pada dokumen. Hasil dari TF.IDF adalah vektor dengan berbagai istilah bersama dengan pembobotan dari istilah tersebut (Srividhya *et al.,*2010). Adapun persamaan

TF.IDF adalah sebagai berikut (Srividhya *et al.*, 2010):

$$w_{m,i} = tfreq_{m,i} \times log_2\left(\frac{N}{n_m}\right) \quad (1)$$

dimana,

$w_{m,i}$ = Bobot term *m* terhadap *dokumeni*

m = *Termm*

i =Dokumen *i*

$tfreq_{m,i}$ = Frekuensi kemunculan suatu *term m* di dalam suatu dokumen *i* dibandingkan dengan frekuensi *term m* yang sering muncul pada dokumen.

N = Jumlah seluruhdokumen.

n_m = Jumlah dokumen *i* yang mengandung *m.*

2.4.Algoritma Clustering K-Means

K-Means dianggap sebagai algoritma yang efektif untuk mengelompokkan suatu data (Larose, 2005). Tahapan algoritmanya adalah sebagai berikut (Larose, 2005) :

Tahapan algoritma adalah sebagai berikut:

1. Menginisialisasi nilai *k* sebagai jumlah *cluster*. Jumlah *k* disesuaikan dengan kebutuhan.
2. Menentukan secara acak bobot pada dokumen yang akan menjadi pusat *cluster* sebanyak jumlah *k* yang sesuai dengan tahap 1.
3. Menentukan jarak antara bobotsetiap *term* padamasing - masing dokumen yang bukan pusat *cluster* dengan bobot setiap *term* pada masing - masing dokumenpusat *cluster* menggunakan jarak *Euclidean (d)*.

$$d_i = \sqrt{\sum_{i=1}^{N}(x_{m,i} - y_{m,i})^2} \quad (2)$$

dimana, d_m = jarak dari setiap *dokumen*,
i = setiap dokumen,
N = jumlah dokumen,
x_i =bobot pada *dokumen* terhadap yang termasuk pusat cluster
y_i =bobot pada *dokumen* terhadap yang bukan pusat cluster

4. Setelah mendapatkan jarak antar bobot dokumen dengan pusat cluster, maka tentukan jarakyang bernilai minimum untuk menjadi anggota *cluster*.
5. Menentukan pusat *cluster* (*centroid*) baru

$$centroidvalue = \sum \frac{a_i}{c} \quad (3)$$

dimana, a_i = jumlah bobot *m* terhadap setiap *i* yang terpilih menjadi anggota cluster *c*,
c = jumlah anggota *cluster* pada setiap *c* yang terbentuk.

6. Mengulangi tahap 3- 5 sampai nilai centroid atau anggota *cluster* sudah tidak berubah

2.4.Validasi Proses Clustering

Validasi *Clustering* diperlukan untuk menghindari adanya pola pada *noise,* membandingkan algoritma *cluster*, dan membandingkan dua *setcluster* atau lebih (Kumar *et al.*, 2004). Salah satu metode yang digunakan dalam validasi ini adalah *purity* (Manning *et al.*, 2009). *Purity* adalah salah satu metode pengukuran validasi *clustering* untuk mengukur kemurnian dari setiap atau keseluruhan *cluster* dengan dihubungkan dengan label kelas yang telah diberikan (Xiong *et al.*, 2009). Semakin besar nilai dari *purity* maka semakin baik solusi *clustering* yang dihasilkan (Xiong *et al.*, 2009).Adapun persamaan dari *purity* adalah sebagai berikut (Wibisono, 2011) :

$$Purity(\Omega, K) = \frac{1}{N}\sum_k max_j \left|\omega_k \cap K_j\right| \quad (4)$$

Nilai $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ adalah kumpulan anggota dari setiap *cluster*. Nilai *N* = Jumlah Objek data yang *dicluster*. Nilai $K = \{K_1, K_2, \ldots K_{j,j}$adalah kumpulan anggota dari setiap *cluster* yang ideal.

3. Metodologi

Target pengetahuan pada pembangunan SMP ini berfokus pengetahuan yang terdiri dari strategi, kebijakan tentang publikasi, keahlian, pengalaman, *knowledge know-how,* prosedur dan arsip tentang penulisan publikasi ilmiah.

Pada bagian ekstraksi dilakukan pendekatan *clustering k-means*. Pada bagian ini hal pertama yang dilakukan adalah menentukan variabel input kemudian dilakukan prapemrosesan terkait proses *tokenization, stopword removal* dan *term weighting* menggunakan TF-IDF. Hal tersebut dilakukan untuk mendapatkan bobot dari setiap *term* pada setiap dokumen pengetahuan. Dari prapemrosesan dilanjutkan dengan melakukan uji skenario dimana dilakukannya proses *clustering k-means* dengan skenario pertama menggunakan jumlah *cluster* (k) hasil dari persamaan *Rule of thumb* (7) (Mardia *et al.*, 1979) dan skenario selanjutnya menggunakan persamaan menemukan nilai *k* pada basis data teks (8) (Can dan Ozkarahan, 1990).

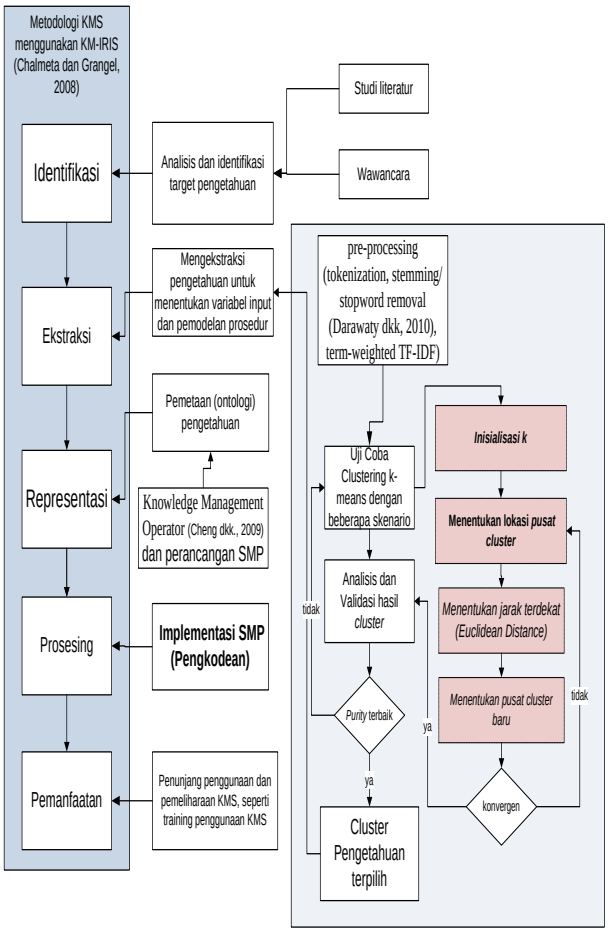
$$k \approx \sqrt{n/2} \quad (7)$$

Nilai *n* adalah jumlah objek yang *dicluster*. Nilai *k* adalah jumlah *cluster*.

$$k \approx \frac{m \times n}{t} \quad (8)$$

Dalam menentukan nilai *k* diperlukan jumlah objek/dokumen (*n*), jumlah *term* (*m*), dan jumlah *record*(*t*)yang berisi nilai lebih dari 0 dari matriks jumlah dokumen × jumlah *term*.

Hasil dari proses *clustering* tersebut divalidasi dengan melihat *purity* dari masing – masing hasil. Hasil yang menghasilkan *purity* terbaik yang akan digunakan SMP khususnya untuk proses *retrieval* pada penelusuran pengetahuan. Gambar 1 menerangkan ilustrasi kerangka penelitian yang mengacu pada proses pembangunan SMP menggunakan metodologi KM-IRIS.



Gambar 1. Kerangka Penelitian

4. Hasil dan Pembahasan

Penelitian ini bertujuan untuk menghasilkan suatu SMP dimana terdapat pendekatan *clustering* sebagai metode*data mining* untuk ekstraksi pengetahuan yang sekaligus menyempurnakan proses keluaran SMP. Adapun target pengetahuan dalam penelitian ini menggunakan 110 dokumen pengetahuan penulisan publikasi berbahasa Indonesia diperoleh dari masukan langsung melalui SMP yangberasal dari hasil wawancara, pedoman penulisan dikti, buku, jurnal dan sumber lain dari artikel – artikel di Internet yang terkait dengan penulisan publikasi.

Dokumen pengetahuan yang telah diperoleh dilakukan prapemrosesan. Prapemrosesan diperlukan untuk mendapatkan bobot dari setiap *term* pada setiap dokumen. Jumlah *term* yang dihasilkan dari prapemrosesan setelah dilakukan *tokenization, stopword removal* dan TF-IDF adalah sebanyak 153 *term* serta banyaknya bobot dari setiap *term* pada setiap dokumen yang nilainya lebih besar dari 0 adalah sebanyak 853.

Setelah dilakukan prapemrosesan maka dilakukan *clustering k-means*. Tahapan pertama algoritma *k-means* adalah menentukan nilai *k.* Untuk melakukan tahapan ini dilakukan dua kali percobaan. Percobaan pertama menggunakan persamaan *rule of thumb* (7).

$$k \approx \sqrt{n/2} \approx \sqrt{110/2} \approx 7 \qquad 9$$

Dari percobaan pertama diperoleh nilai k = 7. Kemudian menentukan nilai k menggunakan percobaan kedua dengan persamaan menemukan nilai k pada basis data teks (8) dan diperoleh nilai k = 20.

$$k \approx \frac{m \times n}{t} \approx \frac{153 \times 110}{853} \approx 20 \qquad 10$$

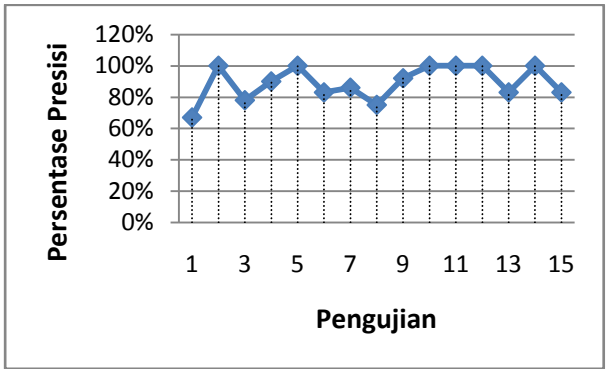
Setelah memperoleh nilai k maka dilanjutkan dengan semua tahapan kmeans sampai anggota dari setiap hasil *cluster* sudah tidak berubah untuk masing-masing percobaan. Untuk menentukan hasil *cluster* yang digunakan dalam pembangunan SMP dilakukan dengan cara menghitung masing-masing *purity* dari setiap hasil *cluster* yang diperoleh. Perbandingan hasil *purity* pada setiap percobaan terlihat pada Tabel 1.

Tabel 1. Perbandingan *Purity* dari setiap percobaan

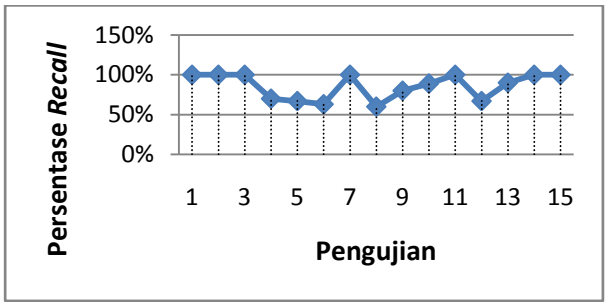
Percobaan	Nilai *k*	*Purity*
1	7	0,7909
2	20	0,8454

Berdasarkan Tabel 1, terlihat bahwa yang mempunyai *purity* terbesar atau semakin mendekati angka 1 adalah percobaan kedua dengan k = 20 dan nilai *puritynya* 0,8454. Berdasarkan hal tersebut, maka hasil percobaan kedua yang akan diekstrak sebagai acuan pada penentuan kata kunci dan keluaran dari proses *retrieval* pada proses penelusuran SMP.

Setelah SMP dibangun dilakukan beberapa pengujian untuk validasi hasil pendekatan *clustering* pada proses *retrieval* untuk penelusuran pengetahuan. Pengujian dilakukan sebanyak 15 kali dengan memilih kata kunci untuk penelusuran pengetahuan. Grafik Validasi pengukuran presisi dan *recall* pada proses penelusuran pengetahuan setelah SMP dibangun terlihat pada Gambar 2 dan Gambar 3.



Gambar 2. Grafik tingkatan persentase presisi pada proses penelusuran SMP



Gambar 3. Grafik tingkatan persentase *recall* pada proses penelusuran SMP

5. Kesimpulan

Berdasarkan penelitian yang dilakukan, maka dapat diambil kesimpulan bahwa pendekatan *clustering* dapat digunakan untuk ekstraksi pengetahuan pada pembangunan SMP. *Clustering* juga dapat digunakan untuk mendukung salah satu fitur utama SMP yaitu *knowledge cluster* dan dapat menyempurnakan keluaran dari SMP. Pendekatan *clustering* secara khusus dapat membantu dalam penelusuran pengetahuan dengan kata kunci berbentuk *cluster* pengetahuan yang dibuktikan dengan pengukuran rata-rata nilai presisi dan *recall* di atas 80%.

Ucapan Terima Kasih

Terima kasih diucapkan pada Ketua Lembaga Penelitian UNIMA yang telah memberikan kesediaan dalam proses wawancara guna keperluan data dalam penelitian ini.

Daftar Pustaka

Babu, K.V.S.N.J., Harshavardhan, T. and Kumar, A.J.S., 2012. The Role of information retrieval in knowledge management. International Journal of Social Science & Interdisciplinary Research 1(10), 212-226.

Baets, W., 2005. Knowledge Management and Management Learning. Springer, USA.

Boahene, M., Ditsa, G., 2003. Conceptual Confusions in Knowledge Management and Knowledge Management Systems: Clarifications for Better KMS Development. IRM PRESS, Australia.

Butler, T., Feller, J., Pope, A., Emerson B., Murphy C., 2008. Designing a core IT artefact for knowledge management systems using participatory action research in a government and a non-government organisation, Journal of Strategic Information Systems 17 (4), 249 – 267.

Can, F., E. A., 1990. Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. ACM Transactions on Database Systems 15 (4), 483 - 517.

Darawaty, I.S., Syarah, S., Nugroho, A.H., Ayuningtyas, F., Istianto, Y., Prasetyo, B., Uliniansyah, M., Gunawan, M., Ani, D., Jarin, A., Handoko, D., 2010. Intelligent Searching Using Association Analysis for law Documents of Indonesian Government. *Second International Conference on Advances in Computing, Control, and Telecommunication Technologies*, Jakarta, December 2-3, 122-124.

Kumar, V., Steinbach, M., Tan, P., 2004. Introduction to Data Mining. Addison-Wesley.

Kurniawan, C., 2012.Perancangan Knowledge Management System PT. XYZ. *Master Thesis*, Graduate Program-Binus University, Jakarta.

Larose, D.T., 2005. Discovering Knowledge in Data : An Introduction to Data Mining. Wiley-Interscience, Canada.

Maier, R., 2007. Knowledge Management Systems - Information and Communication Technologies for Knowledge Management : Third edition.Springer, New York.

Manning, C.D., Raghavan P, dan Schütze H., 2009. Introduction to Information Retrieval. Cambridge University Press.

Mardia. K., Kent., Bibby., 1979. Multivariate Analysis, Academic Press.

Srividhya, R., Anitha, R., 2010. Evaluating preprocessing techniques in text categorization. international Journal of Computer Science and Application Issue, 49 -51.

Wibisono, Y., 2011. Perbandingan Partition Around Medoids (PAM) dan K-means Clustering untuk Tweets, *Prosiding Konferensi Nasional Sistem Informasi*. Medan, Februari 25-26, 483 – 486.

Xiong, H., Wu, J., dan Chen, J., 2009. K-Means clustering versus validation measures: a data distribution perspective. IEEE Transactions on Systems, Man, and Cybernetics – Part B 39 (2), 318-331.