RESEARCH ARTICLE OPEN ACCESS

Community Analysis of Fashion Coordination Using a Distance of Categorical Data Sets

Akira Otsuki*
**Nihon, University*

ABSTRACT
While various clustering methods have been proposed for categorical data, few studies exist on the clustering of the categorical variables that are represented as item sets (groups), e.g. fashion coordination data. In order to fill this gap, this study focuses on the patterns of similarities that are found between coordinated fashion items, so as to define a new set of score value rows for calculating similarity indexes from the patterns. These similarity indexes are then inverted into distances, based on which the clustering method proposed in this study (hereinafter "the Proposed Method") is achieved. Then, it was do two evaluated experiment using about 150,000 coordination data obtained from the wear's web site. Firstly, a silhouette analysis confirms that the Proposed Method offers better cluster partitioning in comparison to previous studies. And secondly, an efficacy evaluation reveals seasonal patterns in fashion coordination, as well as trends for brands and colors of individual fashion items.

***Keywords* -** Fashion Coordinate, Categorical Data, Clustering, Data Mining

I. FORWARD

The community analysis in this study is an example of a network analysis based on graph theory. Graph networks may consist of various elements (data) such as people, organizations, products, a corpus, or academic papers. Graph networks can be created by conceptualizing these elements as networks of relationships within or between communities. Similarities between the network data are then calculated in order to cluster them, so that similar communities can be identified within a huge graph network. While various clustering methods have been proposed for categorical data, few studies exist on the clustering of the categorical variables that are represented as item sets (groups), e.g. fashion coordination data. In order to fill this gap, this study focuses on the patterns of similarities that are found between coordinated fashion items, so as to define a new set of score value rows for calculating similarity indexes from the patterns. These similarity indexes are then inverted into distances, based on which the clustering method proposed in this study (hereinafter "the Proposed Method") is achieved. The Proposed Method is validated in two ways, each using actual fashion coordination data available from the Wear website [1]. Firstly, a silhouette analysis confirms that the Proposed Method offers better cluster partitioning in comparison to previous studies. And secondly, an efficacy evaluation reveals seasonal patterns in fashion coordination, as well as trends for brands and colors of individual fashion items.

II. PREVIOUS STUDIES AND RELATED STUDIES

Ⅱ-1 Community analysis of fashion coordination

Sekozawa et al. [2] proposed a system for analyzing shopping baskets through the formation of clusters, by studying customer preferences and identifying their correlations with fashion. Also, Kawasaki [3] proposed a model based on a simple regression formula $Y=\alpha+\beta t$ (t: year, α and β: coefficients, Y: estimated regression value) that can forecast the next year's trends by analyzing data about the three previous years' fashion trends. Like these, many of the previous studies relate to systems that recommend a certain degree of fashion coordination. Outside Japan, there have also been research studies on the fashion brand trends in different global regions [4]. However, there have been few studies that try to the cluster fashion coordination data directly, as this study attempts to do.

Kanamitsu [5] created a tripartite graph model of fashion brands (based on data for lifestyle, consumers, and brands) for the purpose of a brand analysis. Here, a correlation coefficient is used to calculate the distances between data; then clusters are identified based on the number of each cluster relative to the entire field of clusters, and on the smallness of the standard deviations. While his approach is relatively similar to this study, many more methods for similarity calculations—an integral part of clustering—have been proposed, as we will see in the next section. We will examine these previous studies before discussing the approach used in this study.

Ⅱ-2 Methods for calculating cluster similarities

There are many metrics used for measuring the distances or similarities between (data) nodes. The typical distance metrics used for planar graphs include: the Euclidian distance [6], which is based on the Pythagorean theorem and is used for measuring the distance between two points of known coordinates; the Minkowski distance [7], which is a generalized Euclidian distance and can have different weights for extreme distances; the Manhattan distance [8], which is known to have the same length regardless of its route, as may be the case when moving through a block-pattern city like Manhattan; the Canberra distance [9], which is a relativized variation of the Manhattan distance; and the Mahalanobis distance [10] that is used for two correlated points.

On the other hand, the typical methods used for measuring similarities include: the cosine similarity approach [11], the co-occurrence feature approach [12], and the evaluation of similarities in academic paper citations [13].

It should be noted that all of the popular methods for calculating distances and similarities shown above are used for measuring nodes (data) with vectored data, i.e. data with a quantitative vector and the same data length; while this study focuses on fashion coordination data that is qualitative (or categorical) and requires item sets (or item groups) in order to calculate the similarities. The existing methods for calculating distances are therefore unsuitable for establishing the similarities in this study.

In the next section, we will look at some of the many existing methods used for clustering categorical data.

Ⅱ-3 Methods for clustering categorical data

While there are many previous studies of the methods used for clustering categorical data, they mostly aim to extend the methods that are valid only for quantitative data to categorical data. Huang et al. [14] derived dissimilarity measurements from a data combination of quantitative and categorical variables, and then proposed the k-prototypes method that uses this dissimilarity as a similarity factor for the purpose of clustering. They also proposed the fuzzy k-modes model [15] for the clustering of only categorical variables. Ahmad et al. [16] proposed a method to calculate similarities from the co-occurrence of categorical data, and applied this to the k-means method. These models use the centroids of clusters as their representative points for calculating the similarities, and so can derive the similarity directly from the number of matching categorical values without using the distance.

The "Robust Clustering using linKs," or ROCK, method [17] is another categorical clustering approach that uses the link concept instead of using the distance. For example, when using the Jaccard coefficient, a similarity is established between two targets when there are more co-occurring links between them than the Jaccard coefficient threshold. Both the k-modes and ROCK models rely directly on the number of matching categorical values instead of the distances. The k-medoids method [18], on the other hand, uses the medoids instead of the centroids of clusters as the representative points for similarity calculations. A medoid is a point in a cluster where the total sum of the dissimilarity within the cluster is minimal. In other words, these methods are different from the k-means approach in that they directly apply a distance matrix, and have a similarity with the method used in this study. Although the Kanamitsu method also calculates the similarity by distance calculations, it only uses the correlation coefficients for this distance calculation; therefore, the k-medoids approach is more relevant in the context of this study.

The k-modes, k-medoids, and ROCK models all calculate similarities based on representative points (i.e. the centroids and medoids), whereas the Proposed Method's model is based on the graph network communities themselves. Later, in the section that describes the evaluation methods, we will discuss which approach is suitable for clustering categorical variables, e.g. fashion coordination data that is actually data sets of fashion items.

III. Proposed Method

The target of this study, fashion coordination data, consists of sets of fashion items. Therefore, the clustering similarity is defined as "a combination (group) of similar fashion items" and is applied in calculating the similarities between the fashion coordination data by using the score value rows that are described later. The proposed clustering model is then achieved based on the distances which are derived by inverting these similarities.

Ⅲ-1 Fashion coordination data: definition

The fashion coordination data used for the analysis was retrieved from the Wear website through the method of web scraping. On the website, each user may have more than one piece of fashion coordination data. Each piece of coordination data, in turn, may have one or more items. Each item has the following three attributes, which are shown in Formula (1) below: 1) item type (e.g. shirt, pants, skirt, shoes…); 2) brand; and 3) color. Other attributes are available for some but not all of the Wear data, which is the reason why this study has limited its scope to these three.

$$1Coordinate = \{(itemType1, brand1, color1), (itemType2, brand2, color2), \dots\} \qquad (1)$$

The fashion coordination data described by Formula (1) are converted into Table 1: Multivariate coordination data (see the next page). The scores in the table are the results of the similarity calculation described in the next section.

Table.1: Multivariate coordination data

		Co1			Co2			Co3			Co4		
		$I1_{b1,c1}$	$I2_{b2,c2}$	$I6_{b3,c3}$	$I3_{b2,c2}$	$I4_{b2,c2}$	$I5_{b3,c1}$	$I3_{b4,c2}$	$I2_{b4,c1}$	$I5_{b3,c1}$	$I1_{b1,c1}$	$I4_{b1,c2}$	$I5_{b3,c1}$
Co1	$I1_{b1,c1}$				0.00 (No matching items)			1.00 (Only I2 matched）			1.99 ($I1_{b1,c1}$ matched）		
	$I2_{b2,c2}$												
	$I6_{b3,c3}$												
Co2	$I3_{b2,c2}$	0.00 (No matching items)						2.74 ($I3_{c2}$ and $I5_{b3,c1}$ matched)			2.74 ($I4_{c2}$ and $I5_{b3c1}$ matched)		
	$I4_{b2,c2}$												
	$I5_{b3,c1}$												
Co3	$I3_{b4,c2}$	1.00 (Only I2 matched)			2.74 ($I3_{c2}$ and $I5_{b3,c1}$ matched)						1.99 ($I5_{b3,c1}$ matched)		
	$I2_{b4,c1}$												
	$I5_{b3,c1}$												
Co4	$I1_{b1,c1}$	1.99 ($I1_{b1,c1}$ matched)			2.74 ($I4_{c2}$ and $I5_{b3,c1}$ matched)			1.99 ($I5_{b3,c1}$ matched）					

Note 1) Co_n: Fashion coordination data (Coordination 1, Coordination 2…).
Note 2) $I_{n:}$ Item type (I1: shoes, I2: shirt…).
Note 3) $b_{n:}$ Brand. c_n: Color.

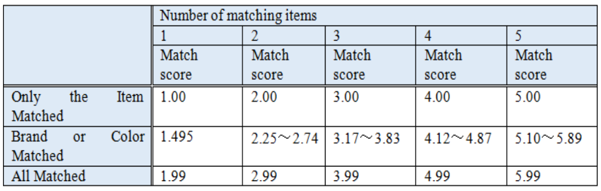
Ⅲ-2 Calculating the similarities between the fashion coordination data

CPM_{ij}(Coordination Patterns Muched) a model for calculating to what extent the fashion coordination data ***i*** and ***j*** matched, is defined by Formula (2). This model is based on the assumption that one or more of the item types match. Table 2 displays the score value rows based on this model.

$$CPM_{ij} = M_{ij} + \frac{S_{ij}}{2 * M_{ij}} \times 0.99 \qquad (2)$$

M_{ij} represents the number of item types that match between the fashion coordinate data ***i*** and ***j***. S_{ij} represents the number of brands or colors that match between the item types. The denominator M_{ij} is multiplied by 2 to accommodate cases where both the brand and the color matches. The last coefficient 0.99 is applied for two purposes: to use the decimal part of the calculation result to represent the number of item type matches, and the fractional part to represent the percentage of matching brands or colors; and to avoid overlaps between the cases where only the item types match. For example, in Table 2, the score for the "All Matched" case when only one item matches is 1.99. This does not overlap with the score 2.00 in the "Only the Item Matched" row where two items match.

Table.2: Score value rows for calculating the fashion coordination data match levels



	Number of matching items				
	1	2	3	4	5
	Match score	Match score	Match score	Match score	Match score
Only the Item Matched	1.00	2.00	3.00	4.00	5.00
Brand or Color Matched	1.495	2.25～2.74	3.17～3.83	4.12～4.87	5.10～5.89
All Matched	1.99	2.99	3.99	4.99	5.99

Note 1) Only the Item Matched: Only the item type matched.
Note 2) Brand or Color Matched: Either the brand or color matched, together with the item type.
Note 3) All Matched: The item type, brand and color matched.

As is shown in Table 2, when neither the brand nor the color matches but the item type does, the relevant score can be found in the "Only the Item Matched" row, in the column for the number of item types that matched. The calculation of the scores in the "Brand or Color Matched" and "All Matched" rows is described based on Table 3.

Table.3: Score calculation example

	Brand	Color
Coordination 1: Cap	○	○◎△
Coordination 2: Cap		
Coordination 1: Shoes	○◎	○◎△□
Coordination 2: Shoes		

Table 3 shows the cases where two item types, the cap and shoes, match in the coordinations 1 and 2. The score for the "○" case, where only the brand matches, which can be found in the "Brand or Color Matched" row in Table 2, can be calculated as follows by applying Formula (2): "2+(1/4)*0.99=2.248." In a similar style, when both the brands and colors match as in the "○◎△□" case, the scores will be in line with those suggested in the "All Matched" row of Table 2, with the actual score being "2+(4/4)*0.99=2.99." Note that while Table 2 mentions only up to five matching items, the same algorithm is applicable to cases where there are six or more item matches. These score value rows enable the similarities between the categorical data sets to be calculated. The primary focus of this study is to propose a clustering method (model) based on the distances derived by inverting these similarities.

The score value rows also have the following characteristics: with regard to the relationship between Co2 and Co3 in Table 1, they have two matching items, i.e. “I3c2” and “I5b3, c1.” The sequence of these matches is irrelevant: no matter which of “I3c2” or “I5b3, c1” matches first, the match score will always be 2.74. This means that the score value row is not affected by the order of the item type matches and always returns the same total score for the same match combination. This characteristic makes the Proposed Model suitable for calculating matches among data sets with disparate match patterns, e.g. fashion coordination data.

Ⅲ-3 Cluster partitioning model of the fashion coordination data

This section describes a clustering model based on CPM_{ij}. The following Similarity Matrix (SM) is derived from Table 1.

$$\text{SM} = \begin{bmatrix} 0.00 & 0.00 & 1.00 & 1.99 \\ 0.00 & 0.00 & 2.74 & 2.74 \\ 1.00 & 2.74 & 0.00 & 1.99 \\ 1.99 & 2.74 & 1.99 & 0.00 \end{bmatrix} \quad (3)$$

This data sets can also be represented as an undirected graph of the fashion coordination data, CoG, which is represented by the pairs of the node group *V* and the edge group *E*.

$$CoG = V, E$$
$$V = \{c1, c2, c3, c4\},$$
$$E = \{c1, c3\}^{1.00}, \{c1, c4\}^{1.99}, \{c2, c3\}^{2.74}, \{c2, c4\}^{2.74}, \{c3, c4\}^{1.99} \quad (4)$$

The “n” in $\{\ \}^{n}$ represents CPM_{ij}. Formula (4) can be developed into a graph network, as is shown as Figure (1). The similarities calculated by Formula (3) are inverted into distances (by subtracting the individual CPM_{ij} values from the maximum CPM_{ij} value).

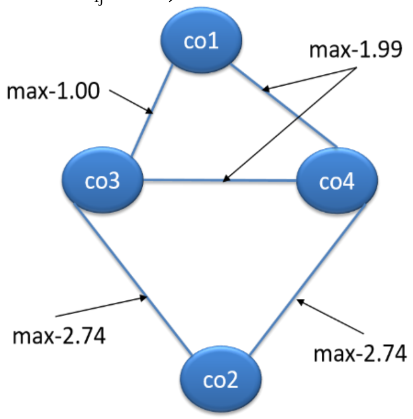


Figure.1: Example of a CoG graph network (co_n is the fashion coordination data. The distances are calculated by subtracting the individual CPM_{ij} values from the maximum CPM_{ij} value.)

IV. Evaluation Process

Ⅳ-1 Overview of the evaluation process

Approximately 150,000 pieces of fashion coordination data, available from the Wear web site between January 2013 and December 2014, were retrieved by the technique of web scraping. Figure 2 shows these pieces of data by each quarter. As the amount of data significantly increased from July 2014, this study focused on 2014 and analyzed each quarter’s data. Based on the amount of data retrieved in the first quarter (1,600), the same number of pieces of data was sampled for each of the other three quarters as well.

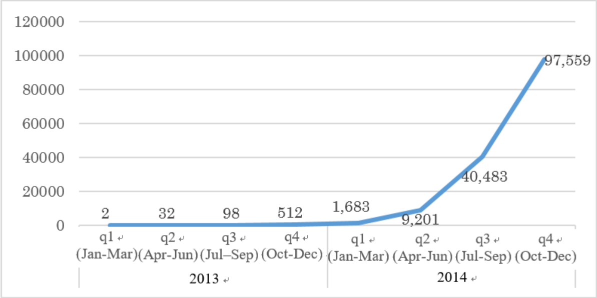


Figure.2: Quarterly transition graph of the fashion coordination data (q1~4 are each quarter)

Each of the 1,600 piece data samples had their similarities calculated by the method described in Sections 3.1 and 3.2. These similarities were then inverted into distances, based on which an undirected graph network was created using the method described in Section 3.3. The remainder of this Chapter will discuss the two evaluation methods that were conducted using this graph network. First, in Section 4.2, we will evaluate the adequacy of cluster partitioning by a silhouette analysis. Then, in Section 4.3, the Proposed Method will be applied to cluster the fashion coordination data retrieved from the Wear website. Each of the resulting clusters will then have their characteristics analyzed to find out the trending item types, brands and colors.

Ⅳ-2 Silhouette Analysis Evaluation

As was discussed in Chapter 2, we will evaluate the advantages of the Proposed Method over the k-medoids, k-modes, and ROCK methods. The evaluation process in this section used 1,600 pieces of data from the first quarter (January–March) of 2014. Furthermore, the Proposed Method for the similarity evaluation (the similarity calculation method described in Section 3.2) was applied to the following clustering methods to determine the clustering method that was best suited for the Proposed Method.

- Edge betweenness method [19]
- Walktrap method [20]
- Infomap method [21]

The adequacy of the cluster partitioning was evaluated by a silhouette analysis [22]. The silhouette index shows how similar each point is to other points within the same cluster. The next formula shows how the index Si for the point *i* is calculated.

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)} \quad (5)$$

a_i is the average distance between the point *i* and other points within the same cluster. b_i is the average minimal distance between the point i and the other points within the same cluster. When the silhouette index S_i is closer to 1.0, a cluster has more nodes with the same distance and is further from other clusters. Figures 4 to 9 show the results of the silhouette analysis. Figure 10 shows how the average silhouette widths in Figs. 4 to 9 compare to each other. In order to find out the optimum number of clusters, *k*, for the k-medoids, k-modes and ROCK algorithms, all of which require this number to be fixed before the simulation, different scenarios had to be explored (Fig. 3). The result was that the k-medoids achieved its highest score (highest average silhouette width) when k=23. This number was k=20 for k-modes; and k=2 for ROCK. These scores were used for the evaluation references. On the other hand, the clustering based on the Proposed Method for the similarity evaluation will always have clusters that are partitioned in the optimum style and there is no need to look for the highest scoring cluster number *k*.

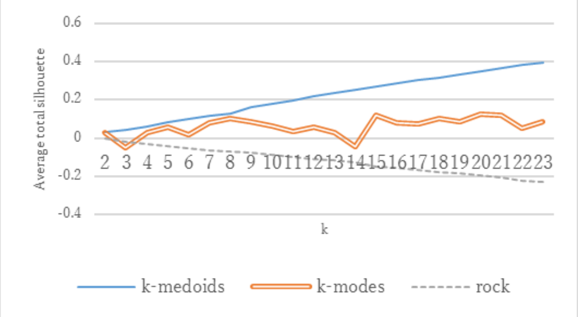


Figure.3: Average silhouette widths of k-medoids, k-modes, and ROCK for all cluster number *k*s

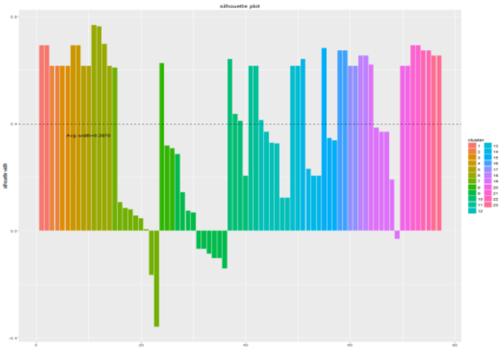


Figure.4: K-medoids silhouette (Average total silhouette width: 0.397038)

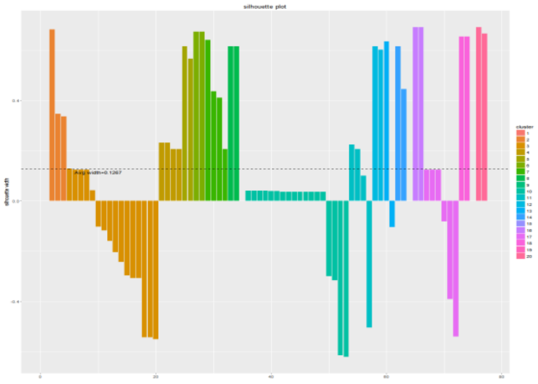


Figure.5: K-modes silhouette (Average total silhouette width: 0.126742)

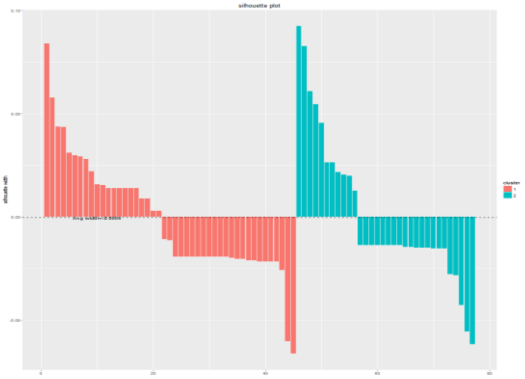


Figure.6: ROCK silhouette (Average total silhouette width: -0.000398)

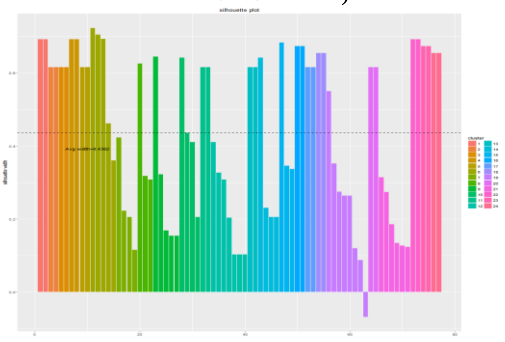


Figure.7: Edge betweenness silhouette (Average total silhouette width: 0.436190)

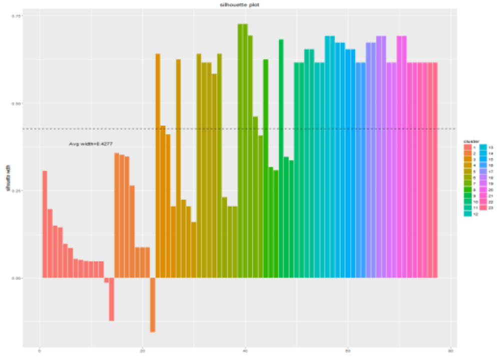


Figure.8: Walktrap silhouette (Average total silhouette width: 0.427683)

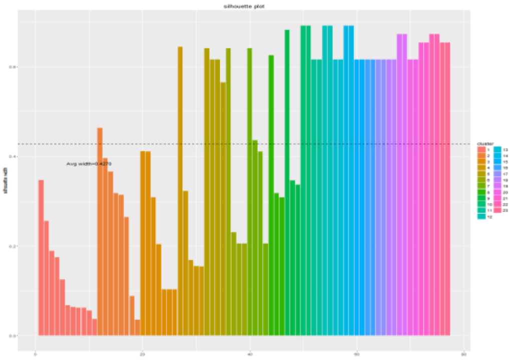


Figure.9: Infomap silhouette (Average total silhouette width: 0.426997)

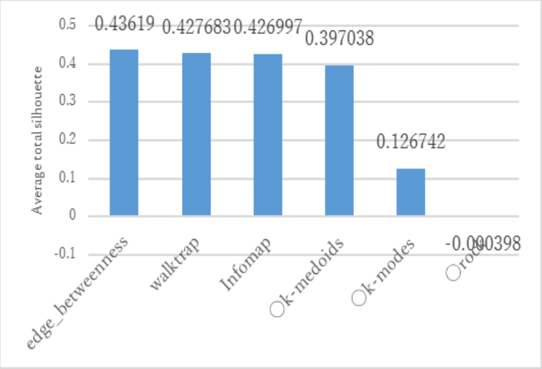


Figure.10: Average total silhouette widths

By comparing the Fig. 10 average silhouette widths, it was found that the Proposed Method, when applied to the edge betweenness, walktrap and infomap approaches, achieved scores that were higher than the k-medoids, k-modes and ROCK methods. The reason that the k-modes and ROCK scores were lower could be that these methods calculated the similarities based on the number of matching categorical variables and not on the similarities between the categorical variable sets. The mode for the k-modes is the mode value for all the members within a cluster, and can be as many as the number of clusters, or *k*. The withindiff values in the following Table 4 are the sums of the differences between these modes and the cluster members, which can also be as many as the number of clusters *k*. In other words, a smaller withindiff value indicates that the cluster members do not deviate much from the mode value used for the optimization. The rows in Tab. 4 represent the number of clusters, or *k* values, while the columns represent the within diff values for specific *k* values. The right-hand average column shows the average withindiff values for each row. As the Table shows, the withindiff values do not become small enough unless the cluster number k values are large enough (in other words, the difference within the cluster will not be small enough). This suggests that the cluster partition was less than optimal, which may be a result of the highly sporadic nature of the data caused by all members having significantly different values.

Table.4: Within diff values by number of clusters

		Within diff values by number of clusters																							average
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	
k	2	533	424																						479
	3	0	533	405																					313
	4	463	0	307	125																				224
	5	507	20	356	17	0																			180
	6	20	9	0	449	26	376																		147
	7	432	330	16	12	11	6	48																	122
	8	192	257	190	9	9	119	11	20																101
	9	0	402	33	125	11	213	0	14	7															89
	10	417	109	146	0	67	9	12	20	11	0														79
	11	402	0	0	216	8	8	63	0	0	31	30													69
	12	24	337	63	67	0	152	22	8	7	49	39	9												65
	13	284	0	0	12	9	29	0	0	71	270	34	0	48											58
	14	0	256	20	33	317	0	14	0	0	55	0	0	0	38										52
	15	20	0	25	327	0	118	16	0	84	6	30	8	7	11	65									48
	16	27	333	17	69	0	48	0	45	0	62	7	26	45	0	12	16								44
	17	20	127	7	63	234	14	0	41	6	36	0	13	0	9	0	0	122							41
	18	0	12	86	8	21	0	0	27	0	9	27	25	77	0	309	0	55	8						37
	19	11	20	9	0	8	99	27	10	16	7	12	101	66	0	11	219	13	0	0					33
	20	0	20	139	24	7	8	36	11	0	240	30	11	16	7	0	0	48	12	0	6				31
	21	9	23	20	67	124	47	8	0	0	182	11	17	0	0	8	23	8	12	9	27	9			29
	22	13	41	0	218	200	20	38	19	0	0	0	9	48	0	0	30	0	7	7	0	0	0		30
	23	149	13	9	7	32	11	10	0	30	0	8	9	0	0	12	256	0	13	26	8	0	0	0	26

It should also be noted that, even though the k-modes method uses the number of simple matches with a cluster's mode for measuring the similarities and only the nodes that satisfy a certain threshold value are used, all the categorical values are taken into account, including many values that have a single occurrence, and such small noises may have some impact. As for the ROCK method, which uses binary (Jaccard) distance matches to measure the similarities, it may not be suitable for variable length data, e.g. fashion coordination data, because the coordinations with many items become the denominators.

On the other hand, the reason that the k-medoids method had a score close to the Proposed Method may be that the k-medoids model does not rely on categorical variable matches, but uses a distance matrix directly instead, as does the Proposed Method. Nevertheless, the k-medoids scores were lower than the Proposed Method. This may be because of the nature of the fashion coordination data that was analyzed in this study—it is possible that the distances were difficult to separate when calculating them from the categorical variables' similarity scores. As Fig. 3 shows, the k-medoids method, which directly uses a distance matrix, generates better partitions when the cluster number *k* is large enough; whereas, when only one cluster based on its representative point is assigned under the k-medoids approach, trying to place the data in one cluster for optimization does not yield a satisfying result because the other clusters' data are placed in too close proximity, i.e. the data correlation in terms of the distance becomes too strong to have good separation. In contrast, the Proposed Method, unlike the k-medoids method, does not base the similarity on individual nodes (representative points), but instead uses the similarity between nodes of graph network communities, leading to better separation than the k-medoids approach. The conclusion of this section is that, when processing fashion coordination data having item sets as categorical variables, the Proposed Method generates better cluster partitioning result than those methods that calculate the similarities based on categorical variable matches, as well as the traditional methods for partitioning clusters based on representative points.

Ⅳ-2 Research on trends in items, brands and colors for individual items using a cluster characteristic analysis

The next Figure (11) shows the results of applying the Proposed Similarity Evaluation Method and the edge betweenness approach to cluster the undirected graph created in section 4.1.

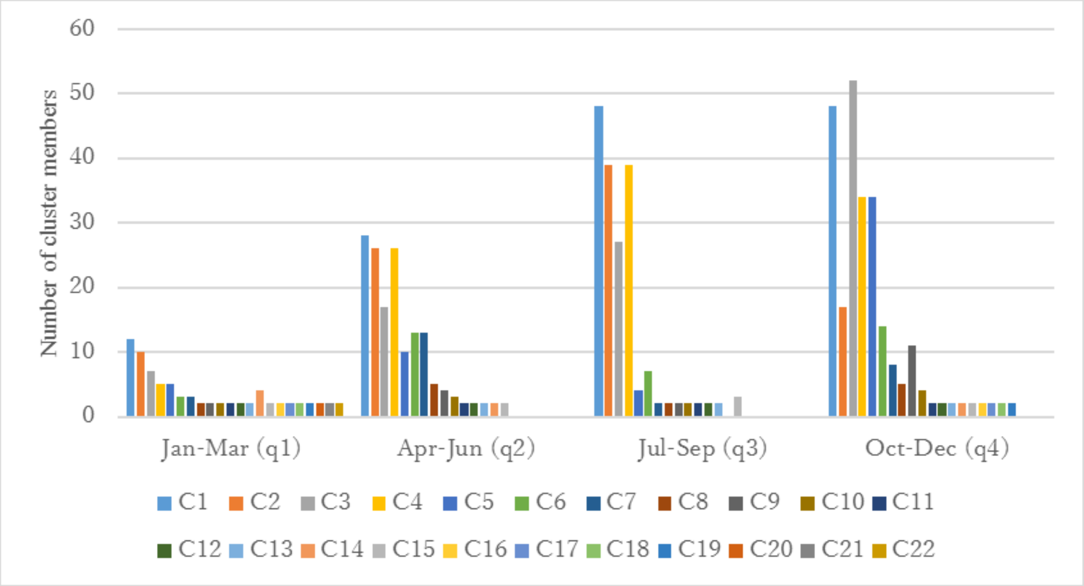


Figure.11: Number of cluster members for 2014 by quarter (Cn: Cluster number *n*)

In this section, five clusters with the most members for each cluster are chosen to analyze the trends for item types, brands and colors. For this study, the trend item types (TI) were defined as the item types with a more than 50% frequency across the entire item types within a cluster.

$$TI = IF \geq \frac{IA}{2} \quad (6)$$

IF is the frequency of this item type. IA is the total number of item types within the cluster. Figures 12 to 15 show the patterns of the clusters by quarter (and which trend item types were pronounced). "qn_cn" in Figs. 12 to 15 indicates the quarter, e.g. q1 representing the January–March period. cn represents an individual cluster. The vertical axis in each figure shows the relevant frequency of each TI within the cluster. All TIs were further categorized into "hats/caps", "coats/jackets", "tops", "bottoms", "underwear", "footwear", "bags", and "accessories". For example, in Fig. 12, the q1_c2 data can be interpreted to be a fashion coordination group centering on "shirts and blouses", "outers", "pants", and "sneakers". In the January-March quarter illustrated in Fig. 12 and the October-December quarter in Fig. 15, winter TIs such as the "standard fall collar coat" and "knitted sweater" were visible, but in the April-June quarter illustrated in Fig. 13 and the July-September quarter in Fig. 14, summer TIs like "sandals" and "t-shirt" were more popular. These are examples of a cluster analysis that help us to understand the seasonal changes in fashion coordination. Additionally, the “shirt and blouse”, “t-shirt and cut-&-sew shirt”, “pants” and “sneakers” were the TIs that remained noticeable in many clusters throughout the year (Figs 12 to 15).

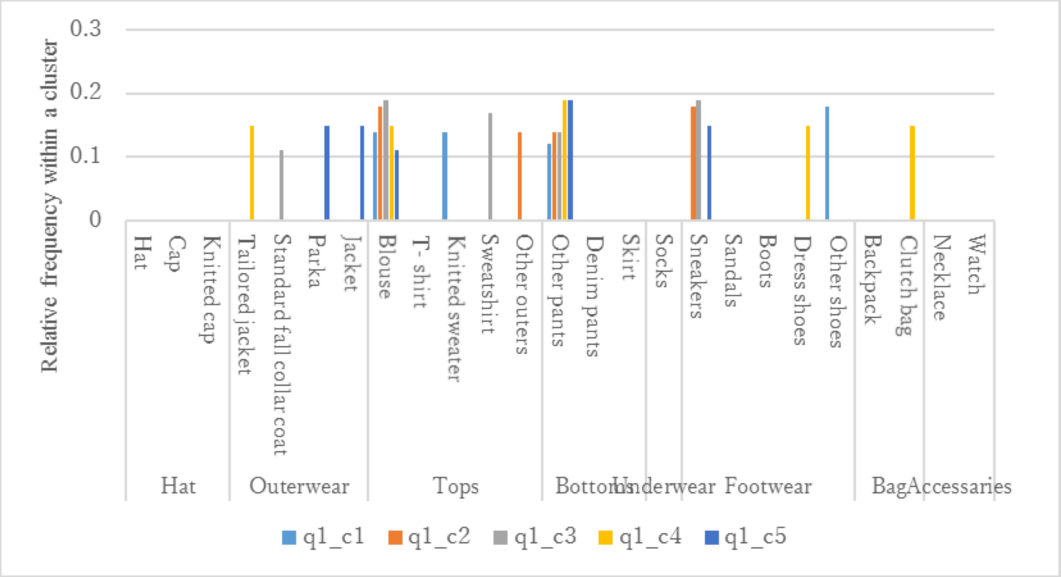


Figure.12: Cluster characteristics, January–March 2014

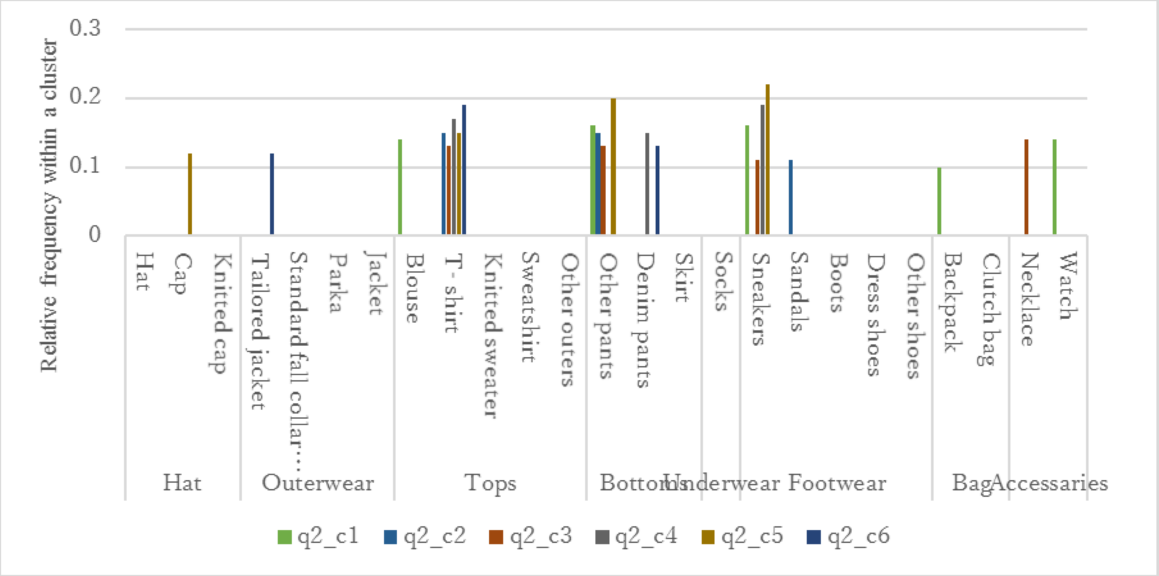


Figure.13: Cluster characteristics, April–June 2014

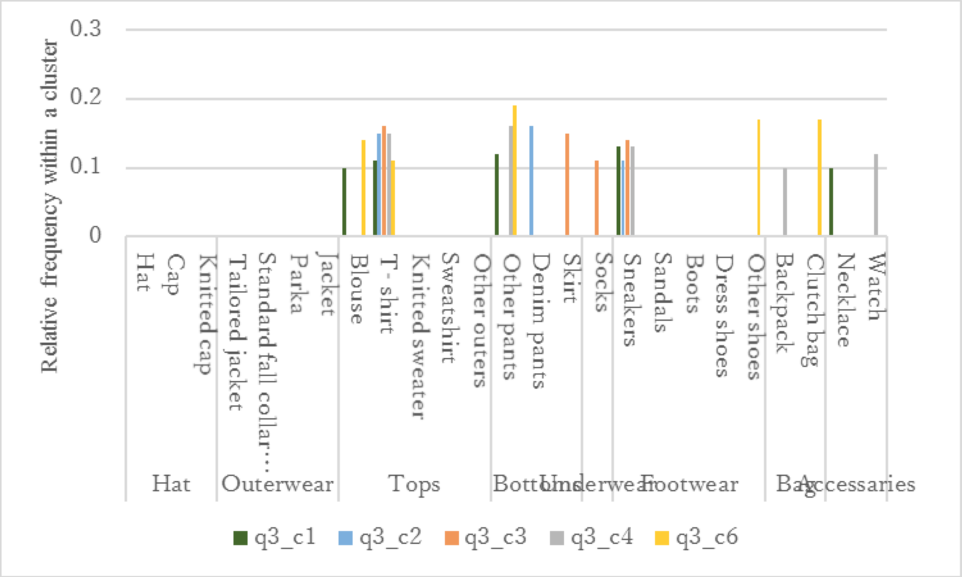


Figure.14: Cluster characteristics, July–September 2014

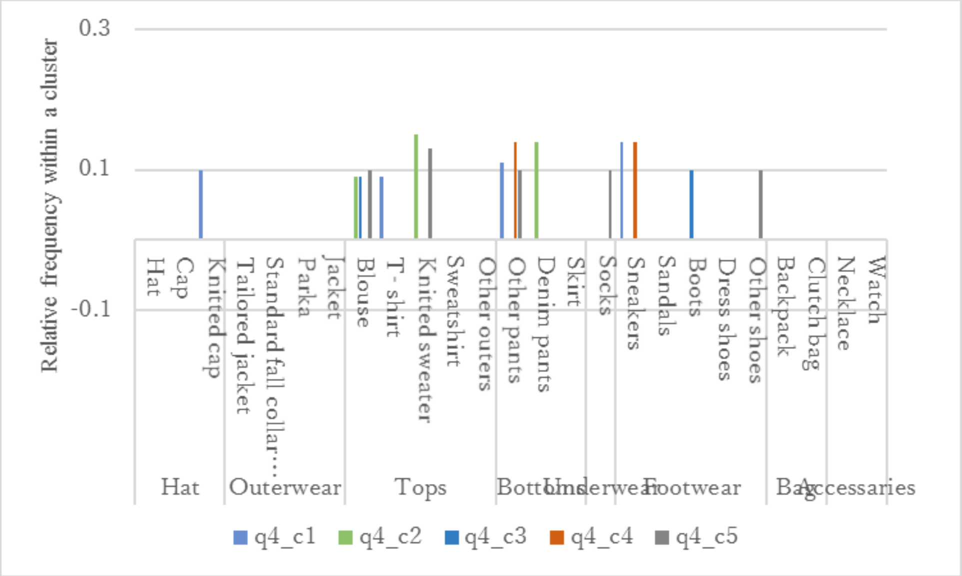


Figure.15: Cluster characteristics, October–December 2014

Figures 16 to 19 show the TI brands in each cluster, while Figures 20 to 23 show their colors. The vertical axis in each Figure represents the relative frequencies of the TI brands and colors in the cluster. When we look at Figs. 16 to 19, the most popular TI brand (i.e. the brand that was used in the most clusters) in the January-March quarter was "HARE" for shirts, but the pants and sneakers brands all had a score of 1.0. During the April-June quarter the most popular brands were "UNIQLO" and "RAGEBLUE" for t-shirts and cut & sew shirts, "UNIQLO" for pants, and "VANS" for sneakers. In the July-September quarter, the most popular brands were "UNIQLO" for t-shirts, cut & sew shirts and pants, and "NIKE" for sneakers. Finally, during the October-December quarter, the leading brand was "UNIQLO" for shirts, blouses and pants, but all the sneaker brands had the same score of 1.0.

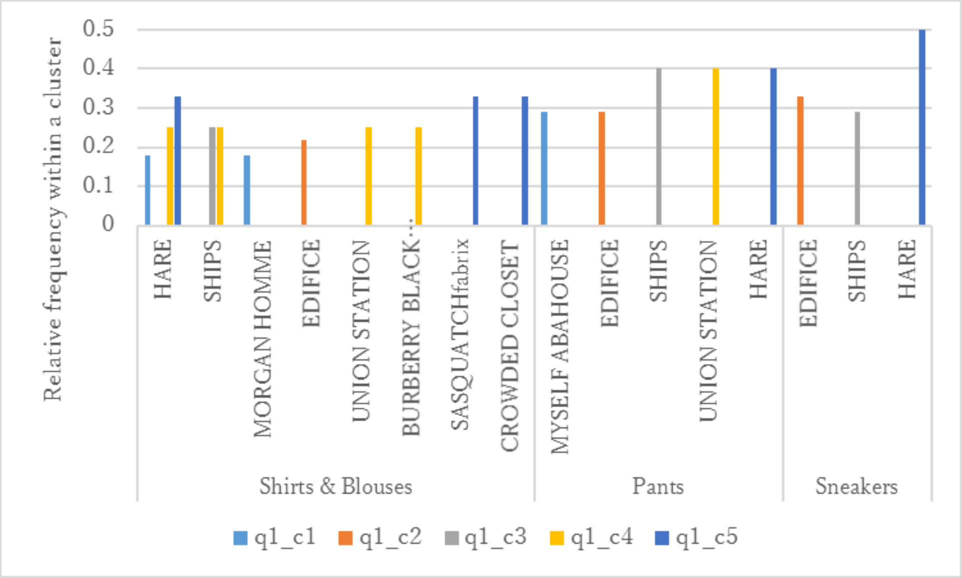


Figure.16: Trending brands, January–March 2014

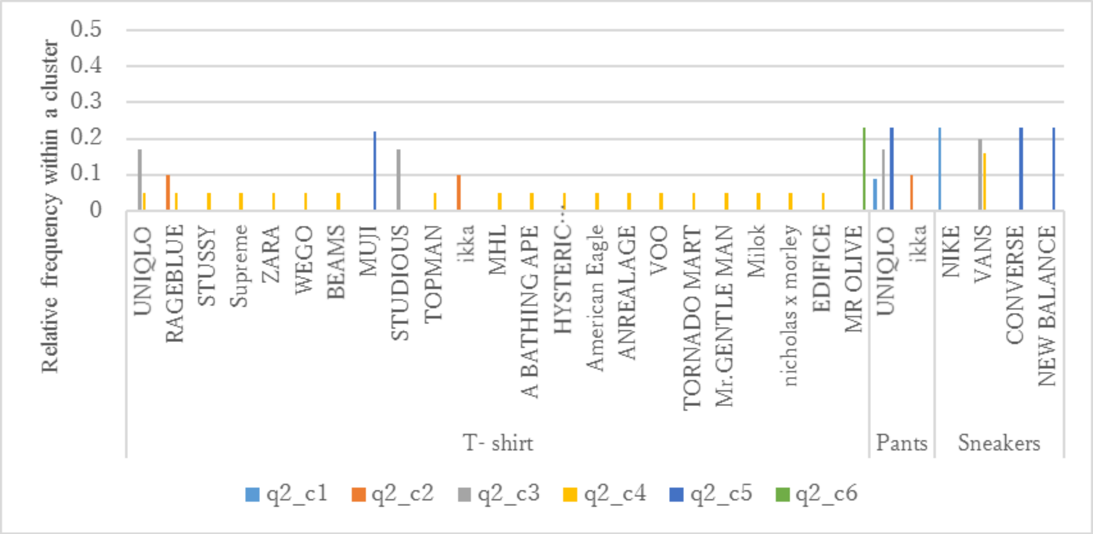


Figure.17: Trending brands, April–June 2014

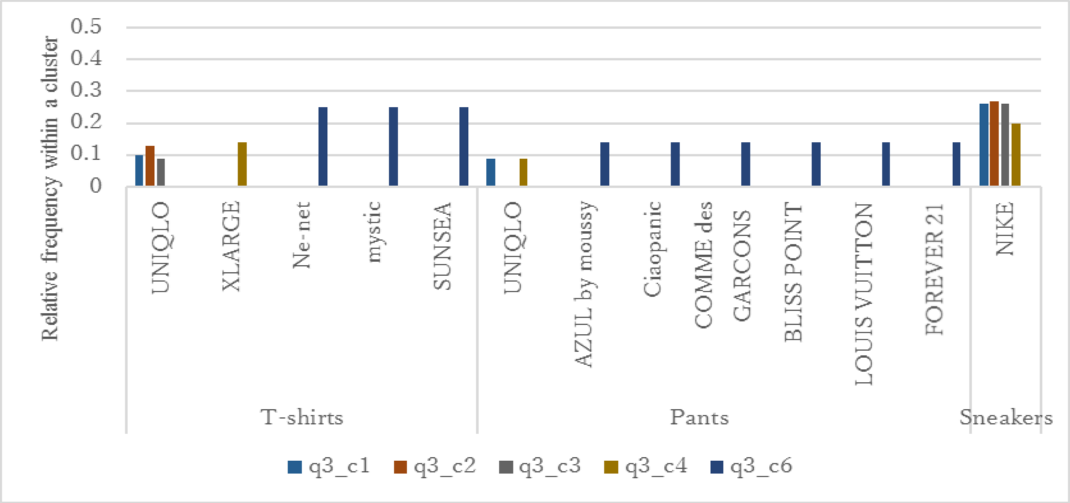


Figure.18: Trending brands, July-September 2014

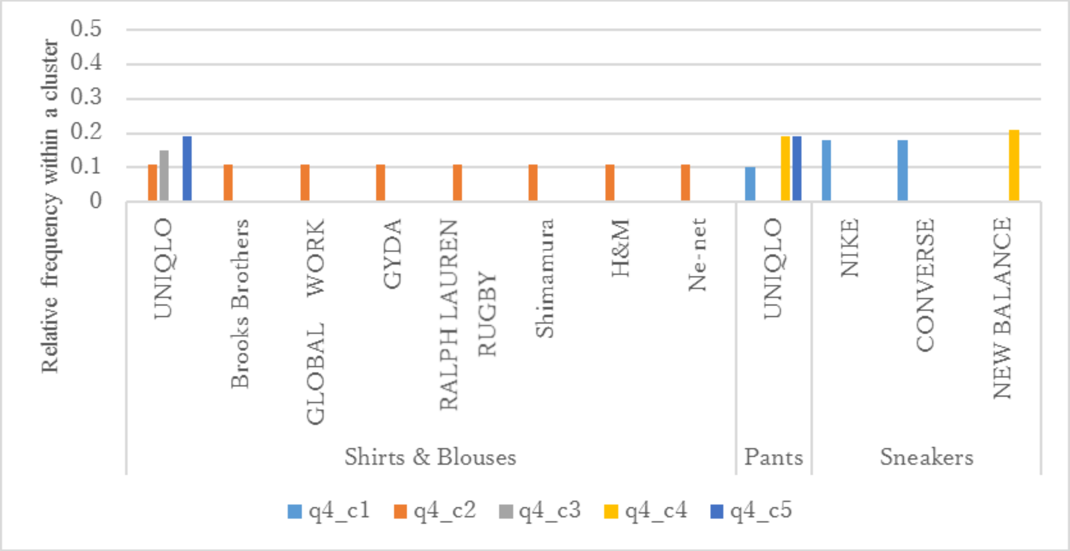


Figure.19: Trending brands, October–December 2014

The next Figures (20 to 23) show the TI colors that were used the most (i.e. the colors used in the most clusters). During the Jan-Mar quarter, "white" dominated for the shirts & blouses while "black" ruled in pants, but all of the colors scored 1.0 point for sneakers. In the Apr-June quarter, the prevailing colors were "white" for t-shirts and cut & sew shirts, and "black" for pants and sneakers. In the July-Sept quarter the popular colors were "white" for t-shirts and cut & sew shirts, "black" for pants, and "white" plus "black" for sneakers. Finally, in the Oct-Dec quarter, "white" was the trend for shirts & blouses, "black" was the most popular choice for pants and sneakers.

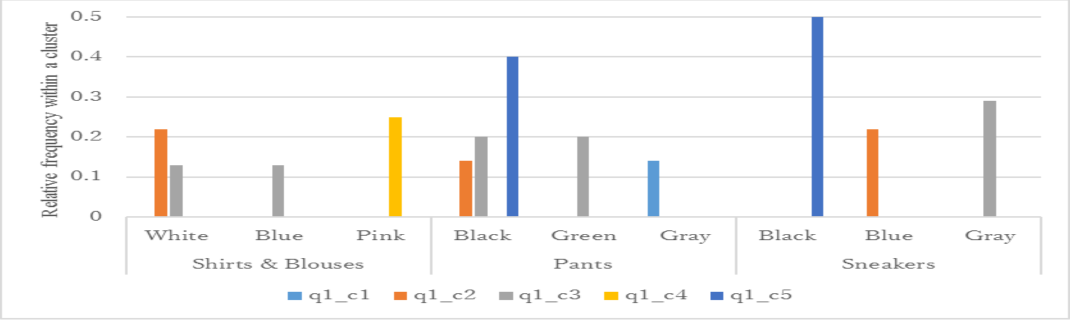


Figure.20: Color trends, January–March 2014

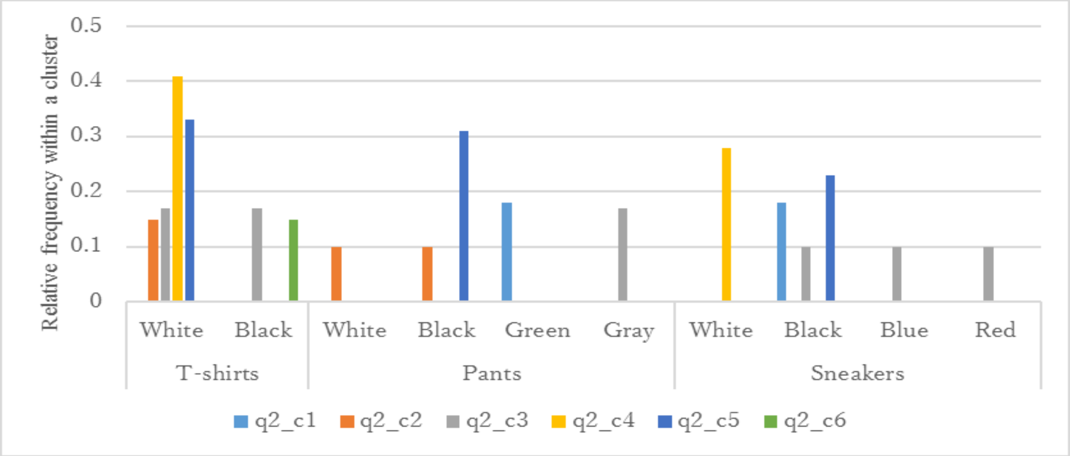


Figure.21: Color trends, April–June 2014

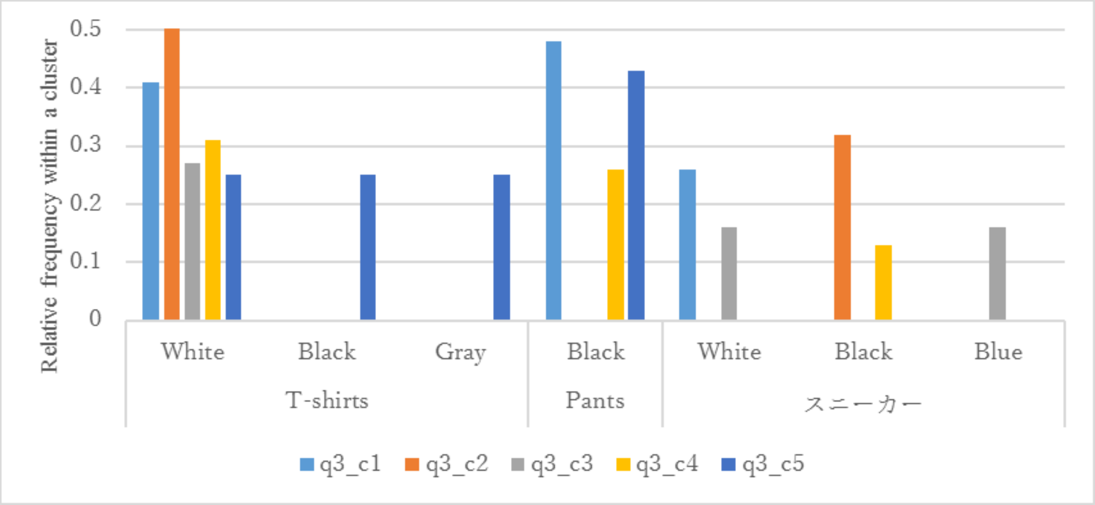


Figure.22: Color trends, July–September 2014

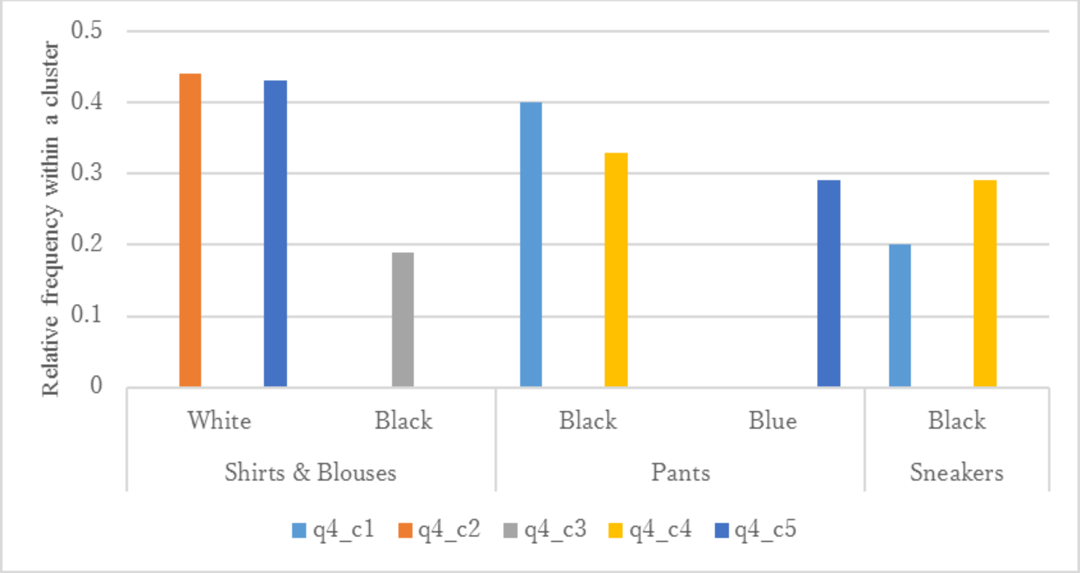


Figure.23: Color trends, October–December 2014

V. CONCLUSION

While various clustering methods have been proposed for categorical data, few studies exist on the clustering of categorical variables that are represented as item sets (groups), e.g. fashion coordination data. In order to fill this gap, this study focused on the patterns of similarities found between coordinated fashion items in order to define a new set of score value rows for calculating the similarity indexes from the pattern. These similarity indexes were then inverted into distances, based on which the clustering method proposed in this study was achieved. The clustering method was validated in two ways, each using 150,000 pieces of fashion coordination data available from the Wear website. First, a silhouette method evaluation confirmed that the Proposed Method offers better cluster partitioning in comparison with previous studies. Secondly, an evaluation approach consisting of a characteristics analysis of the clusters formed by the Proposed Method identified the seasonal patterns in the fashion coordination, as well as the trends in brands and colors for individual fashion items.

Future targets from this study may include adding more data to the evaluation process described in Chapter 4, since only a limited amount of samples were actually used.

Adding more evaluation data will also lead to further discussions on the optimum clustering method for the Proposed Similarity Evaluation technique, as only a small number of clustering approaches were tested in this study.

Acknowledgements
The author expresses his deepest gratitude to Start Today Co., Ltd. for generously approving the use of Wear website data for this study.

REFERENCES

[1] Wear: http://wear.jp/

[2] Teruji SEKOZAWA, Hiroyuki MITSUHASHI, Yukio OZAWA, “One to One Recommendation System in Apparel On-Line Shopping”, The transactions of the Institute of Electrical Engineers of Japan. C, A publication of Electronics, Information and System Society, Vol.128, No.8, pp.1333-1341, 2008.

[3] Kentaro KAWASAKI:” Information processing for prediction of fashion analysis “, The Japan Research Association for Textile End-Uses, Vol.20, No.5, pp.161-168, 1979.

[4] Choi, Tsan-Ming , ”Liu, Shuk-Ching; Tang, Christopher S., et al: A cross-cluster and cross-region analysis of fashion brand extensions”, The Journal of The Textile Institute, Vol.102, No.10, pp.890-904, 2011.

[5] Jun KANAMITSU: Structure of High Fashion Brands: Examination of A Three-partite Network Model of Brands Linking Consumers with Life Values/Personas, Kyoto Management Review, Vol.20, pp.93-109, 2012.

[6] Carter, R., Morris, R., & Blashfield, R, ”On the partitioning of squared euclidean distance and its application in cluster analysis”, Psychometrika, Vol.54, No.1, pp.9-23, 1989.

[7] Groenen, PJF; Jajuga, K , ”Fuzzy clustering with squared Minkowski distances”, Fuzzy Sets and Systems, Vol.120, No.2, pp.227-237, 2001.

[8] Anupam Gupta and R. Ravi,” Algorithmic Applications of Metric Embeddings”, http://www.cs.cmu.edu/~anupamg/metrics/.

[9] Gordon, A. D., ”2 Measures of similarity and dissimilarity”, Classification. 2nd ed. Chapman &H all, pp.15-34, 1999.

[10] Mahalanobis, Prasanta Chandra,” On the generalised distance in statistics”. Proceedings of the National Institute of Sciences of India 2 (1): 49–55. Retrieved 2013-12-17, 1936.

[11] Takayuki NOZAWA, MasaakiIDA, Fuyuki YOSHIKANE, Kazuteru MIYAZAKI, Hajime KITA, ” Construction of Curriculum Analyzing System Based on Document Clustering of Syllabus Data”, IPSJ Journal, Vol.46, No.1, pp.1-12, 2005.

[12] Yuki YASHUDA, ” Can marketing control the relationship – From possibility of the customer relationship and specific items by network analysis-, Marketing Journal, Vol.101, pp.4-17, 2006.

[13] Akira Otsuki, Masayoshi Kawamura , ”The Study of the Role Analysis Method of Key Papers in the Academic Networks”, The International Journal of Transactions on Machine Learning and Data Mining (MLDM), ibai-publishing, Volume 6, Number 1, pp.3-13, 2013.

[14] Huang, ZX. , ”Extensions to the k-means algorithm for clustering large data sets with categorical values”, Data Mining and Knowledge Discovery, Vol.2, No.3, pp.283-304, 1998.

[15] Huang, ZX , Ng, MK. , ”A fuzzy k-modes algorithm for clustering categorical data”, IEEE Transactions on Fuzzy Systems, Vol.7, No.4, pp.446-452, 1999.

[16] Ahmad, Amir , ”Dey, Lipika: A k-mean clustering algorithm for mixed numeric and categorical data”, Data & Knowledge Engineering, Vol.63, No.2, pp.503-527, 2007.

[17] Guha, S., Rastogi, R. and Shim, K., ”ROCK: A Robust Clustering Algorithm for Categorical Attributes”, ICDE, 1999.

[18] Kaufman L., Rousseeuw P.J., ”Finding Groups in Data: An Introduction to Cluster Analysis”, John Wiley & Sons, 1990.

[19] Girvan, M. & Newman, M. E. J., ”Community structure in social and biological networks”, Proc. Natl. Acad. Sci. USA 99, 7821-7826, 2002.

[20] Pascal Pons, Matthieu Latapy , ”Computing Communities in Large Networks Using Random Walks”, Journal of Graph Algorithms and Applications, Vol. 10, no. 2, pp. 191-218, 2006.

[21] Martin Rosvall,Carl T. Bergstrom, ”Maps of random walks on complex networks reveal community structure”, PNAS,vol.105, no.4, pp.1118–1123, 2008.

[22] Rousseeuw, P.J., ”Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. J. Comput. Appl. Math., 20, pp.53–65, 1987.

Author Profile’s:

Akira Otsuki

Received his Ph.D. in engineering from Keio University (Japan), in 2012. He is currently associate professor at Nihon University College of Economics (Japan) and Officer at Japan society of Information and knowledge (JSIK). His research interests include Big Data Landscape Analysis, Data Mining, Academic Landscape, and new knowledge creation support system. Received his Best paper award 2012 at JSIK. And received his award in Editage Inspired Researcher Grant, in 2012.