韓國 컴퓨터情報學會 論文誌
第11卷 第5號, 2006. 11.

2006-11-5-4-1

# 확장된 시퀀스 요소 기반의 유사도를 이용한 계층적 클러스터링 알고리즘

오 승 준*

# A Hierarchical Clustering Algorithm Using Extended Sequence Element-based Similarity Measure

Seung-Joon Oh*

## 요 약

최근 들어 상업적이거나 과학적인 데이터들의 폭발적인 증가를 볼 수 있다. 이런 데이터들은 항목들 간의 순서적인 면을 가지고 있는 시퀀스 데이터들이다. 그러나 항목들 간의 순서적인 면을 고려한 클러스터링 연구는 많지 않다. 본 논문에서는 이들 시퀀스 데이터들 간의 유사도를 계산하는 방법과 클러스터링 방법을 연구한다. 특히 다양한 조건을 고려한 확장된 유사도 계산 방법을 제안한다. splice 데이터 셋을 이용하여 본 논문에서 제안하는 클러스터링 방법이 기존 방법 보다 우수하다는 것을 보여준다.

## Abstract

Recently, there has been enormous growth in the amount of commercial and scientific data. Such datasets consist of sequence data that have an inherent sequential nature. However, only a few of the existing clustering algorithms consider sequentiality. This study presents a similarity measure and a method for clustering such sequence datasets. Especially, we present an extended concept of the measure of similarity, which considers various conditions. Using a splice dataset, we show that the quality of clusters generated by our proposed clustering algorithm is better than that of clusters produced by traditional clustering algorithms.

▸ Keyword : 클러스터링(Clustering), 시퀀스(Sequence), 유사도(Similarity)

---

• 제1저자 : 오승준
• 접수일 : 2006.10.12, 심사일 : 2006.10.25, 심사완료일 : 2006.11.18
* 경기공업대학 산업경영시스템과 교수

# I. 서 론

클러스터링이란 물리적 혹은 추상적 객체들을 서로 비슷한 객체들의 집합으로 그룹화 하는 과정으로, 하나의 클러스터에 속하는 객체들 간에는 서로 다른 클러스터 내의 객체들과는 구분되는 유사성을 갖게 된다[1]. 클러스터링 기법들은 통계학(statistics), 패턴 인식(pattern recognition) 등의 분야에서 연구되어 왔으며, 현재는 데이터 마이닝 분야에서 이 기법을 응용하려는 연구가 활발히 진행되고 있다.

최근에는 상업적이거나 과학적인 데이터의 폭발적인 증가를 볼 수 있다. 이들 중 웹 로그, 단백질 시퀀스, 소매점 거래 데이터 등과 같은 분야의 데이터들은 순서적인 면을 가지고 있는 시퀀스 데이터(또는 시퀀스)들이다. 즉, 데이터의 항목들 간에 순서가 존재하는 것이다. 예를 들어, 두 개의 시퀀스들이 동일한 항목들로 이루어졌더라도 항목들 간의 순서가 다르면 서로 다른 시퀀스들이다. 그러나 기존의 클러스터링 방법들은 데이터들 내에 순서가 존재하는 면을 고려하지 않았거나, 효율적으로 시퀀스들 간의 유사도를 계산하는 방법을 사용하지 않았다.

항목들 간에 순서가 존재하는 시퀀스들을 클러스터링 하는 것은 많은 면에서 유용하다. 예를 들면, 웹 사용자들의 사이트 방문 기록을 보관한 웹 로그 파일들을 이용하여 웹 사용자들을 클러스터링 하는 것은 서로 다른 웹 사용자 그룹들을 발견하는데 도움을 준다[2]. 또한, 비슷한 구조를 공유하는 단백질 시퀀스들끼리 그룹화 하는 것은 비슷한 기능을 갖는 단백질 시퀀스들을 찾는데 도움을 준다.

본 연구에서는 웹 로그나 단백질 시퀀스, 소매점 거래 데이터 등과 같이 항목들 사이에 순서가 존재하는 시퀀스들을 클러스터링 하는 문제를 다룬다. 이를 위해서는 시퀀스들 간의 유사도를 구하는 것이 무엇보다 중요하다. 이를 위해, 본 연구에서는 기존의 유사도 계산 방법을 확장시킨 새로운 유사도 계산 방법을 제안한다. 또한, 이 방법을 이용한 계층적 클러스터링 알고리즘도 제안한다.

# II. 기존 연구

다양한 클러스터링 기법들에 대한 연구들은 [3]에 있으며, 클러스터링 기법들에 사용되는 데이터의 종류들에 대한 분류는 [1]에 있다. 기존의 클러스터링 기법들은 주로 수치형 값들의 데이터[4,5]와 범주형 값들의 데이터[6]들만을 문제영역으로 다루어 왔다.

최근에는 웹 마이닝 분야에서도 클러스터링 기법을 이용한 연구가 활발히 이루어지고 있는데[7], 여기에는 비슷한 내용의 웹 페이지끼리 클러스터링을 하는 웹 contents 마이닝 분야의 [8]와 웹 사용자의 웹 사용 패턴을 클러스터링 하는 웹 usage 마이닝 분야의 [9]과 [10]이 있다. 그러나, [8,9,10] 모두 항목들 간의 순서는 고려하지 않고 있다.

시퀀스에 대한 연구는 주로 빈발하는 순차 패턴을 찾는데 집중되어 왔다. 이 문제는 [11]에서 처음으로 제안되었는데, 이 분야의 순차패턴을 탐사하는 문제는 시퀀스의 지지도가 사용자가 정의한 최소지지도보다 큰 시퀀스를 발견하는 것이다. [12]에서는 순차 패턴을 일반화 시켜 표현하는 방법을 다루었다.

시퀀스들에 대한 클러스터링 연구로는 다음의 세 가지 연구가 있다. [7]은 빈발패턴이 주어져 있다고 가정을 하고, 이 빈발 패턴을 하나 이상 포함한 시퀀스들만을 대상으로 클러스터링을 수행한다. [13]은 시퀀스들 사이의 유사도로 edit distance 방법을 사용하여 클러스터링을 수행하고, [14]는 sequence alignment 방법을 이용하여 클러스터링을 수행한다. 그러나, 본 연구에서는 [7]와 달리 빈발패턴에 상관없이 모든 시퀀스들을 대상으로 클러스터링을 수행하고, [13,14]에서 사용한 유사도 계산 방법과 다른 새로운 유사도를 사용하여 시퀀스들을 클러스터링 한다.

# III. 시퀀스들 간의 유사도 계산 방법

## 1. 유사도 측정

데이터들을 그룹화하거나 클러스터링 하는데 있어서는 유사도의 개념이 중요하다. 그러나 데이터들 사이의 유사도는 데이터의 종류에 따라 달라지며 데이터의 특성에 따라 여러 종류의 유사도 측정 방법들이 존재한다. 즉, 두 데이터들 사이의 유사도가 어떤 유사도 측정 방법에서는 매우 높게 나올 수 있지만, 다른 유사도 측정 방법을 이용하면 낮게 나올 수도 있는 것이다.

일반적으로 두 시퀀스들 간의 유사도는 공통 항목이 많을수록, 또한 항목들의 순서가 동일할수록 높다고 할 수 있다. 따라서 이 두 가지 요소를 동시에 고려하기 위해서는 두 시퀀스 사이에 동일 서브 셋들이 얼마나 많이 존재하느냐를 고려한다. 본 연구에서는 동일 서브 셋들을 찾기 위해 순서를 가지는 두 항목 쌍들을 이용한다. 즉, 두 시퀀스들 사이

에 동일 항목 쌍들이 많을수록 유사도가 높게 나오는 성질을 이용한다.

[예제3.1] 두 시퀀스 $S_1$ = <A B C D>, $S_2$ = <A C D E>가 있다. $S_1$의 두 항목 쌍들의 모임은 (AB, AC, AD, BC, BD, CD)이고 $S_2$ 의 두 항목 쌍들의 모임은 (AC, AD, AE, CD, CE, DE)이다. $S_1$, $S_2$에 동일한 두 항목 쌍들이 많을수록 유사도는 높다. 여기서,(AC, AD, CD)가 공통 두 항목 쌍들이다.

## 2. 시퀀스 요소를 이용한 유사도 계산방법

시퀀스 $S = < x_1\ x_2\ \dots\ x_i\ x_j\ \dots\ x_n >$ 에서 순서를 가지는 2개의 항목들로 구성된 $x_i\ x_j$ 를 시퀀스 요소 $e_k$ 라고 하며, $e_k$ 들의 모임을 $E = (e_1, e_2, \dots, e_k, \dots)$ 라 한다. E의 크기는 E에 있는 요소들의 개수이며, $|E|$로 나타낸다.

[예제3.2] 시퀀스 S = < A B C E >에서 시퀀스 요소들의 모임은 E = (AB, AC, AE, BC, BE, CE)이며, $|E| = 6$이다.

시퀀스내의 항목들뿐만 아니라 항목들 간의 순서도 고려를 해서 식(1)과 같이 유사도 계산 방법을 제안한다.

[정의3.1] 두 시퀀스 $S_1$ 과 $S_2$ 의 시퀀스 요소들의 모임을 각각 $E_1$, $E_2$ 라고 하면, $S_1$, $S_2$ 의 유사도 $sim(S_1, S_2)$는 다음과 같이 정의한다.

$$sim(S_1, S_2) = \frac{|E_1 \cap E_2|}{\frac{|E_1| + |E_2|}{2}} \quad \cdots (1)$$

여기서, $|E_1 \cap E_2|$ 는 $E_1$과 $E_2$의 공통 요소들의 개수이며, $E_1$과 $E_2$사이에 공통 항목들이 많을수록 유사도는 높고, 이 값을 $(|E_1| + |E_2|)/2$로 나누는 것은 유사도를 0과 1사이의 값을 갖도록 하기 위해서이다[15,16].

[예제3.3] 두 시퀀스 $S_1$ = < A B D A >, $S_2$ = < A C D A C >에서 시퀀스 요소들의 모임은 각각 $E_1$ = ( AB, AD, AA, BD, BA, DA ) 과 $E_2$ =( AC, AD, AA, AC, CD, CA, CC, DA, DC, AC ) 이며, $|E_1| = 6$, $|E_2| = 10$, $E_1 \cap E_2$ = ( AD, AA, DA ), $|E_1 \cap E_2| = 3$ 이다. 따라서, 두 시퀀스의 유사도 $sim(S_1, S_2)$는 3/8이다.

## 3. 확장된 유사도 계산 방법

이번에는 2절에서 제안한 기본적인 유사도 계산 방법에 현실 세계의 다양한 조건을 추가적으로 고려해 보자.

2절에서는 시퀀스 $S_1$ = < A B C D E >과 $S_2$ = < A B C D P >의 유사도를 계산하는데 있어, [정의 3.1]을 이용하여 단순히 서로 공통이 되는 시퀀스 요소의 개수에 비례하여 유사도를 계산하였다. 즉, $S_1$과 $S_2$의 서로 공통이 되는 시퀀스 요소인 AB, AC, AD를 똑같이 고려한 것이다. 그러나, 실제적으로는 AB를 AC나 AD 보다 더 많이 고려할 수도 있고, 공통된 서브 시퀀스 요소를 찾을 때 간격이 3 이상인 AD를 제외할 수도 있다. 따라서 이러한 조건들을 고려한 확장된 개념의 유사도 정의 방법에 대해서 설명한다.

시퀀스 $S = < x_1\ x_2 \cdots x_i\ x_j \cdots\ x_n >$에서 시퀀스 요소들의 모임 E의 크기는 $|E|$로 나타내며, 다음과 같이 계산한다. $|E| = w_{1,2} + \dots + w_{i,j} + \dots + w_{n-1,n}$ , 여기서, $w_{i,j} = \frac{1}{d_{i,j}}$ 이다. 데이터베이스 D에서, xg (최대 간격, Maximum Gap)는 시퀀스 내에서 최대로 허용될 수 있는 간격, ng (최소 간격, Minimum Gap)는 시퀀스 내에서 최소로 요구되는 간격이다.

[예제3.4] 시퀀스 S = < A B C D E >에서 xg = 2, ng = 1이라고 하면,시퀀스 요소들의 모임은 E = (1AB, 1/2AC, 1BC, 1/2BD, 1CD, 1/2CE, 1DE)이고, $|E| = 1 + 1/2 + 1 + 1/2 + 1 + 1/2 + 1 = 11/2$이다.

ng, xg 등은 순차 패턴을 일반화하여 표현하는데 사용되었던 파라미터들이다[12]. 본 연구에서는 이들을 이용하여 기존의 edit distance 방법과 sequence alignment 방법에서 고려하지 못했던 다양한 조건들을 고려하여 유사도를 계산할 수 있다.

시퀀스 $S = < x_1\ x_2\ \dots\ x_i\ x_j\ \dots\ x_n >$에서 2개의 항목들로 구성된 $\frac{1}{d_{i,j}} x_i x_j$를 시퀀스 요소 $e_k$ 라고 한다. 여기서 $i < j$이고, $ng \le d_{i,j} \le xg$ 이다. 그러면, 확장된 유사도 계산 방법을 다음과 같이 제안한다.

[정의3.2] 두 시퀀스 $S_1 = < a_1\ a_2 \cdots a_i\ a_j \cdots a_{n-1}\ a_n >$, $S_2 = < b_1\ b_2 \cdots b_k\ b_l \cdots b_{m-1}\ b_m >$ 의 시퀀스 요소들의 모임을 각각

$$E_1 = (\frac{1}{d_{1,2}} a_1 a_2, \cdots, \frac{1}{d_{i,j}} a_i a_j, \cdots),$$

$$E_2 = (\frac{1}{d_{1,2}} b_1 b_2, \cdots, \frac{1}{d_{k,l}} b_k b_l, \cdots)$$

라고 하면 ($ng \le d_{ij} \le xg$, $ng \le d_{kl} \le xg$), $S_1$, $S_2$ 의 유사도 $sim(S_1, S_2)$는 다음과 같이 정의한다.

$sim(S_1, S_2)$=

$$\frac{\sum_{a_i a_j \in E_1, b_k b_l \in E_2} \frac{w_{i,j} + w_{k,l}}{2} \delta(a_i a_j, b_k b_l)}{\frac{|E_1| + |E_2|}{2}} \quad (2)$$

여기서, $w_{ij} = 1/d_{ij}$, $w_{kl} = 1/d_{kl}$,

$$\delta(a_i a_j, b_k b_l) \begin{cases} = 1 & \text{if } a_i a_j = b_k b_l \\ = 0 & \text{otherwise} \end{cases}$$

[예제3.5] 시퀀스 $S_1$ = < A B C >, $S_2$ = < A B D >, $S_3$ = < A E B >가 있다. ng=1, xg=3 일 때 $sim(S_1,S_2)$, $sim(S_1,S_3)$를 계산해 보자. [정의 3.1]을 사용하면 $sim(S_1,S_2)$ = 1/3, $sim(S_1,S_3)$ = 1/3로 $sim(S_1,S_2)$ = $sim(S_1,S_3)$이다. 그러나, [정의 3.4]를 사용하면, $S_1$의 시퀀스 요소들 모임 $E_1$은 (1AB, 1/2AC, 1BC)이고, $S_2$의 시퀀스 요소들 모임 $E_2$는 (1AB, 1/2AD, 1BD)이고, $S_3$의 시퀀스 요소들 모임 $E_3$는 (1AE, 1/2AB, 1EB)이다. $|E_1| = |E_2| = |E_3| = 5/2$ 이고, $sim(S_1,S_2)$ = 2/5, $sim(S_1,S_3)$ = 3/10로 $sim(S_1,S_2) \neq sim(S_1,S_3)$이다.

## Ⅳ. 계층적 클러스터링 알고리즘

계층적 클러스터링 알고리즘은 통합(agglomerative) 방법과 분리(divisive) 방법으로 나눌 수 있다. 통합 방법은 처음에 각각의 객체들을 하나의 클러스터로 설정 한 후 이들 쌍간의 거리 (혹은 유사도)를 기반으로 가장 가까운 클러스터(객체)들끼리 합병을 수행한다. 최종적으로 한 클러스터 내에 모든 객체들이 포함될 때까지 위의 과정을 반복한다. 분리 방법은 통합 방법과 반대로 위의 과정을 진행한다[1].

본 연구에서는 통합 방법의 계층적 클러스터링 알고리즘을 사용한다. n개의 시퀀스들을 클러스터링 하는 문제를 생각해 보자. 처음에는 $n\cdot(n-1)/2$개의 클러스터간 합병을 고려할 수 있는데, 이중에서 합병을 했을 경우 가장 높은 평가함수 값을 주는 두 개의 클러스터를 합병한다. 1번째 합병 후에는 $(n-1)\cdot(n-1-1)/2$개의 클러스터간 합병을 고려하며, 이중에서 가장 높은 평가함수 값을 주는 두 개의 클러스터를 합병한다. 최종적으로는 주어진 개수의 클러스터가 남을 때까지 위의 과정을 반복한다.

본 연구에서는 평가함수로 식(3)를 사용한다.

$$\text{maximize } Cf = \sum_{r=1}^{k} \frac{1}{n_r} \sum_{i,j \in C_r} sim(i,j) \quad \cdots\cdots (3)$$

여기서, $n_r$은 Cr 내의 시퀀스들 개수,
k는 클러스터 개수

식(3)은 [17]에 있는 평가함수들 중 하나인 식(4)를 변형한 것이다. 식(4)는 모든 클러스터에 대하여, 클러스터내에 있는 데이터 쌍들 간의 유사도 평균에 데이터 개수를 곱하여 모두 합한 값이며, 데이터들 간의 유사도로 코사인 유사도를 이용했지만, 본 연구에서는 3장에서 제안한 유사도 계산 방법을 사용한다.

$$\text{maximize } Cf = \sum_{r=1}^{k} n_r \left( \frac{1}{n_r^2} \sum_{i,j \in C_r} \cos(i,j) \right) \quad \cdots\cdots (4)$$

여기서, $n_r$은 Cr 내의 데이터들 개수,
k는 클러스터 개수

본 연구에서는 최단 거리법(shortest linkage method), 평균 거리법 (average linkage method), 최장 거리법(complete linkage method) 등 기존의 방법들 대신에 식(3)을 평가함수로 사용한다.

본 연구에서 제안하는 클러스터링 알고리즘의 단계는 [그림 1]과 같다.

Step 0. Initialize
For all i
$C_i \leftarrow S_i \in D$

Step 1. Compute the value of
the criterion function
For each $C_i, C_j \in D$
compute the criterion function (Eq.(3))

Step 2. Merge
$C_{new} \leftarrow$ merge $(C_i, C_j)$ for the max
value of the criterion function (Eq.(3))

Step 3. Check the condition
If ( $|C_i| > k$ ) then go to Step 1.
Else go to Step 4.

Step 4. Exit

그림 1. 계층적 클러스터링 알고리즘
Fig 1. The hierarchical clustering algorithm

Step 0은 초기화 단계로서 데이터베이스 D를 액세스하여 각각의 시퀀스를 하나의 클러스터로 설정한다. Step 1은 두 클러스터가 합병이 될 경우의 평가함수 식(3)의 값을 구하는 단계로, 현재 n개의 클러스터가 있다고 하면, $n \cdot (n-1)/2$개의 평가함수 값을 계산한다. Step 2는 합병 단계로서, Step 1에서 계산한 평가함수 값들 중 가장 큰 값을 주는 두 개의 클러스터를 합병한다. Step 3은 조건 검사 단계로서 클러스터의 개수($|C_i|$)가 지정된 클러스터 개수(k)보다 크면 Step 1로 간다. 그렇지 않으면 Step 4로 간다. 마지막으로, Step 4는 종료 단계로서 알고리즘을 끝낸다.

## V. 실험결과

본 연구에서 제안하는 방법을 기존 방법들과 비교 평가하기 위해, splice 데이터 셋과 합성 데이터 셋으로 실험을 하였으며, C++ 언어로 코딩을 하여 수행하였다.

splice 데이터 셋은 UCI KDD 아카이브에 포함되어 있는 데이터 셋이다[18].

표 1. splice 데이터 셋
Table 1. Splice data set

| 데이터 셋 | 시퀀스들의 개수 | 시퀀스의 크기 |
|---|---|---|
| splice | 1,535 | 60 |
| EI | 767 | 60 |
| IE | 768 | 60 |

표 1에서 보면, splice 데이터 셋은 60개의 항목을 가진 뉴클레오타이드(nucleotide) 시퀀스들을 포함하고 있으며, 각각의 시퀀스들은 엑손/인트론 경계 (exon/intron, EI라 부름)나 인트론/엑손 경계 (intron/exon, IE라 부름)에 속하는 클래스 레이블을 가진다. EI에 속하는 시퀀스들이 767개이며, IE에 속하는 시퀀스들이 768개이다.

세 가지 실험 방법은 표 2에 있으며, 실험 결과는 표 3에 있다. 제안하는 방법에서 최소간격(ng)은 1, 최대간격(xg)은 시퀀스의 크기인 60을 사용하였다.

splice 데이터 셋은 클래스 레이블이 EI와 IE의 두 가지로 구분되므로, 클러스터의 개수를 2개로 클러스터링 하여 한 클러스터는 EI로, 다른 하나는 IE로 클러스터가 구성되는지 실험하였다.

표 2. 세 가지 실험 방법 비교
Table 2. Comparison of method 1, method 2, and the proposed clustering method

| | 유사도 | 클러스터링 방법 |
|---|---|---|
| 방법 1 | edit distance 방법 [13] | 제안하는 계층적 클러스터링 알고리즘 |
| 방법 2 | edit distance 방법 [13] | 최장거리법을 이용한 계층적인 클러스터링 |
| 제안하는 방법 | 제안하는 방법 (식(2)) | 제안하는 계층적 클러스터링 알고리즘 |

표 3. splice 데이터 셋에 대한 실험결과
Table 3. Experimental results for the splice data set

| 클러스터 번호 | 방법 1 | | 방법 2 | | 제안하는 방법 | |
|---|---|---|---|---|---|---|
| | EI | IE | EI | IE | EI | IE |
| 1 | 614 | 577 | 766 | 768 | 553 | 266 |
| 2 | 153 | 191 | 1 | 0 | 214 | 502 |

표 3에서 보면 방법 1로 클러스터링 한 클러스터 1에는 EI가 614개, IE가 577개, 클러스터 2에는 EI가 153개, IE가 191개로 EI와 IE가 대략 반반씩 섞여있다. 방법 2로 클러스터링한 결과는 1개의 시퀀스를 제외하고는 모든 시퀀스가 클러스터 1으로 클러스터링 되어 있다. 이에 반해, 제안하는 방법에서는 클러스터 1에 EI가 553개, IE가 266개, 클러스터 2에 EI가 214개, IE가 502개로 구성이 된다. 즉, 대부분이 EI으로 구성된 하나의 클러스터와 IE로 구성된 또 하나의 클러스터를 얻을 수 있었다. 본 연구에서 제안하는 유사도를 사용함으로써 클러스터링 결과가 좋아진 것을 알 수 있었다.

다음으로, 방법 1, 2와 제안하는 방법의 F-measure는 표 4와 같다.

표 4. splice 데이터 셋에 대한 F-measure 비교
Table 4. Comparison of F-measure for the splice data set

| | 방법 1 | 방법 2 | 제안하는 방법 |
|---|---|---|---|
| F-measure | 0.4852 | 0.5010 | 0.6869 |

표 4에서 보는 바와 같이, 제안하는 방법의 F-measure가 방법 1,2보다 우수함을 알 수 있다.

## VI. 결론

본 논문에서는 범주형 항목들이 순서를 가지고 있는 시퀀스들의 클러스터링 문제를 연구하였다. 최근 들어 웹 로그, 단백질 시퀀스, 소매점 거래 데이터 등과 같은 데이터들의 폭발적인 증가를 볼 수 있다. 이런 데이터들은 항목들 간의 순서를 고려해야 하는 시퀀스 데이터들이다. 시퀀스들은 동일한 항목들로 이루어졌더라도 항목들 간의 순서가 다르면 서로 다른 시퀀스 데이터들이다. 그러나, 항목들 간의 순서적인 면을 고려한 클러스터링 연구는 많지 않았다. 본 논문에서는 이러한 시퀀스 데이터들을 클러스터링 하기 위한 문제를 연구하였다.

시퀀스 데이터들을 클러스터링 하기 위해서는 두 시퀀스들 사이의 유사도를 효율적으로 측정하기 위한 방법이 필요하다. 시퀀스들 간의 유사도를 측정하기 위해, 본 논문에서는 시퀀스 요소를 이용하는 확장된 유사도 측정 방법을 제안하였다.

본 논문에서는 두 시퀀스들 간의 유사도 계산 방법을 이용하여 시퀀스들을 계층적 클러스터링 알고리즘을 이용하여 클러스터링 하였다.

splice데이터 셋을 이용한 실험을 통하여, 제안하는 클러스터링 방법이 기존 방법보다 성능이 우수함을 보였다.

향후에는 다양한 데이터 셋들에 대해 본 연구에서 제안하는 알고리즘을 적용하는 것이 필요하며, 범주형 뿐만 아니라 수치형 값들을 포함하는 시퀀스들도 연구해야 할 과제이다.

## 참고문헌

[1] Han, J.and Kamber, M. (2001). Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 335-393.

[2] Perkowitz, M. and Etzioni, O. (1999). "Towards Adaptive Web Sites: Conceptual Framework and Case Study", Proc. 8th Int. WWW Conf., Canada.

[3] Ye, N. (2003). The handbook of data mining, Lawerence Erlbaum Associates, New Jersey.

[4] Guha, S., Rastogi, R. andShim, K. (2001). "CURE: An Efficient Clustering Algorithm for Large Databases", Information Syst. 26(1):35-58.

[5] Han, J., Kamber M. and Tung, A. K. H. (2001). "Spatial Clustering Methods in Data Mining: A Survey", Geographic Data Mining and Knowledge Discovery, eds. H. J. Miller and J. Han (Taylor and Francis, New York).

[6] Guha, S., Rastogi, R. andShim, K. (2000).

"ROCK: A Robust Clustering Algorithm for Categorical Attributes", Information Syst. 25(5):345-366.

[7] Morzy, T., Wojciechowski, M. andZakrzewicz, M. (2001). "Scalable Hierarchical Clustering Method for Sequences of Categorical Values", Proc. 5th Pacific-Asia Conf.Knowledge Discovery and Data Mining, Kowloon, Hong Kong.

[8] Roussinov, D. and Zhao, J. L. (2003). "Automatic discovery of similarity relationships through web mining", Decision Support Syst. 35(1).

[9] Fu, Y., Sandhu, K. and Shih, M. Y. (1999). "Clustering of Web Users based on Access Patterns", Proc. 1999 KDD workshop on Web Mining, San Diego, CA.

[10] Mobasher, B., Dai, H., Luo, T., Nakagawa, M., Sun, Y. andWiltshire, J. (2002). "Discovery of Aggregate Usage Profiles for Web Personalization", Data Mining and Knowledge Discovery, 6, 61-82.

[11] Agrawal, R., and Srikant, R. (1995). "Mining Sequential Patterns", Proc. Int. Conf. Data Engineering, Taiwan.

[12] Joshi, M., Karypis, G. and Kumar, V. (1999). "Universal Formulation of Sequential Patterns", Technical Report TR 99-021, Univ. of Minnesota, Dept. of Com. Sci.

[13] Hay, B., Wets, G.and Vanhoof, K. (2003). "Segmentation of Visiting Patterns on Web Sites using a Sequence Alignment Method", Journal of Retailing and Consumer Services, 10, 146-153.

[14] Wang, W. and Zaiane, O. R. (2002). "Clustering Web Sessions by Sequence Alignment", 13th Int. Workshop on Database and Expert Syst. Applications, France.

[15] 오승준, "범주형 시퀀스 데이터의 K-Nearest Neighbour 알고리즘", 컴퓨터정보학회 논문지, 제10권, 제2호, 2005.

[16] 오승준, 원민관, "텍스트 마이닝 기법을 이용한 컴퓨터 네트워크의 침입 탐지", 컴퓨터정보학회 논문지, 제10권, 제5호, 2005.

[17] Zho, Y.and Karypis, G. (2002). "Comparison of Agglomerative and Partitional Document Clustering Algorithms", 2nd SIAM Int. Conf. Data Mining, Arlington, VA.

[18] Blake, C. L. and Merz, C. J. (1998). UCI Repository of Machine Learning Databases.

## 저 자 소 개

**오 승 준**

2004년 8월 한양대학교 산업공학과, 공학박사

2005~ 현재 : 경기공업대학 산업경영시스템과 교수

〈관심분야〉 데이터마이닝, 인공지능, 전문가시스템