韓國 컴퓨터情報學會 論文誌
第13卷 第2號, 2008. 3.

2008-13-2-1-13

러프 셋 이론을 이용한 시퀀스 데이터의 클러스터링 알고리즘

오 승 준*, 박 찬 웅**

A Clustering Algorithm for Sequence Data Using Rough Set Theory

Seung-Joon Oh *, Chan-Woong Park **

요 약

월드 와이드 웹에는 거대한 양의 하이퍼링크들과 웹 사용 정보들을 포함하고 있는 동적인 페이지들이 모여 있다. 이러한 구조화되어 있지 않은 웹 데이터들과 온라인 정보들의 폭발적인 증가로 인해 효율적인 웹 데이터 마이닝 툴이 필요로 하게 되었다. 최근에는 웹 사용자들의 특성을 자동적으로 발견하기 위한 Web usage mining 분야에서 많은 연구가 진행되고 있다. 본 연구에서는 웹 사용자들의 방문 기록, 단백질 시퀀스, 소매점 거래 데이터 등과 같은 시퀀스 데이터를 분석하는 방법에 대하여 연구한다. 러프 셋 이론을 이용하여 시퀀스 데이터들을 클러스터링 하는 방법을 제안하고, 간단한 예제를 통하여 제안하는 절차를 소개하고 splice 데이터셋과 합성 데이터셋을 통한 실험 결과를 제시한다.

Abstract

The World Wide Web is a dynamic collection of pages that includes a huge number of hyperlinks and huge volumes of usage informations. The resulting growth in online information combined with the almost unstructured web data necessitates the development of powerful web data mining tools. Recently, a number of approaches have been developed for dealing with specific aspects of web usage mining for the purpose of automatically discovering user profiles. We analyze sequence data, such as web-logs, protein sequences, and retail transactions. In our approach, we propose the clustering algorithm for sequence data using rough set theory. We present a simple example and experimental results using a splice dataset and synthetic datasets.

▸ Keyword : 시퀀스 데이터(Sequence Data), 러프 셋 이론(Rough Set Theory), 웹 마이닝(Web Mining)

• 제1저자 : 오승준
• 접수일 : 2008. 1. 9, 심사일 : 2008. 2. 12, 심사완료일 : 2008. 3. 16.
* 경기공업대학 산업경영과 교수 **경원대학교 산업정보시스템공학과 교수

Ⅰ. 서 론

월드 와이드 웹 (WWW)에는 거대한 양의 하이퍼링크들과 웹 사용 정보들을 포함하고 있는 동적인 페이지들이 모여 있다. 이러한 구조화되어 있지 않은 웹 데이터들과 온라인 정보들의 폭발적인 증가로 인해 효율적인 웹 데이터 마이닝 툴이 필요로 하게 되었다.

웹 마이닝은 웹 문서나 서비스로부터 데이터 마이닝 기법을 사용하여 정보를 발견하거나 추출하는 기법[1]으로, Web content mining, Web structure mining, Web usage mining등 세 분야로 나눌 수 있다[2,3].

Web content mining은 웹 사이트의 컨텐츠, 자료, 정보 등의 관계를 분석하여 사용자의 요구에 가장 잘 부합하는 내용을 보여 줄 수 있도록 자동적으로 찾아주는 기법이다.

Web structure mining은 웹 사이트와 웹 페이지의 하이퍼링크를 데이터 마이닝 과정을 통해 구조화, 표준화 시키는 프로세스이다.

Web usage mining은 웹 서버로부터 사용자의 엑세스 패턴을 발견하는 자동화된 마이닝 기법을 말한다. Srivastava, et al. [4]와 같이 기존의 데이터 마이닝 기법들을 이용할 수도 있고 합성 연관 규칙[5]이나 순서 발견 기법[6]처럼 기존 알고리즘을 변형한 방법들도 있다.

특히 웹 사용자의 특성을 자동적으로 발견하기 위한 Web usage mining 분야 중, 본 연구에서는 웹 로그, 단백질 시퀀스, 소매점 거래 데이터 등과 같은 시퀀스 데이터를 분석 하는 방법에 대하여 연구한다. 특히 사용자들이 방문한 웹 사이트 하이퍼링크들의 시퀀스들을 웹 트랜잭션이라고 하는데, 본 연구에서는 이들 웹 트랜잭션들을 클러스터링 하는 방법을 제안한다.

본 논문에서 제안하는 방법을 이용하여 시퀀스들을 클러스터링 하는 것은 많은 면에서 유용하다. 예를 들면, 웹 사용자들의 사이트 방문기록을 보관한 웹 로그 파일들을 이용하여 웹 사용자들을 클러스터링 하는 것은 웹 사용자의 행위들 중 흥미로운 패턴들을 발견하거나 서로 다른 웹 사용자 그룹들을 발견하는데 도움을 준다. 또한 비슷한 구조를 공유하는 단백질 시퀀스들을 클러스터링 하는 것은 비슷한 기능을 갖는 단백질 시퀀스들을 찾는데 도움을 준다.

본 논문에서는 시퀀스들을 클러스터링 하기 위해 러프 셋 이론의 상한 근사 방법을 이용하는데, 이 방법은 데이터에 대한 사전정보나 부가적인 정보가 필요 없이 분석을 수행한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 연구를 다루며, 3장에서는 러프 셋을 이용하여 클러스터링을 수행하는 방법에 대해 설명한다. 4장에서는 3장에서 제안한 방법을 이용한 간단한 예제를 보여주며, 5장에서는 splice 데이터셋과 합성 데이터셋을 통한 실험결과를 제시한다. 마지막으로 6장에서 결론을 기술한다.

Ⅱ. 기존 연구

1. 러프 셋 이론

1980년대 초에 Pawlak에 의해 소개된 러프 셋 이론은 어떤 집합에서 확실하게 분류되는 하한 근사 공간(Lower Approximation)과 불확실하게 분류되는 상한 근사 공간(Upper Approximation)을 집합 이론을 통해서 나타낸다[7,8]. 러프 셋 이론의 가장 중요한 장점 중 하나는 부정확하거나 불완전 하고, 애매모호한 성질을 가진 데이터의 분류 분석 문제에 적합한 알고리즘이라는 것이며, 또한 데이터에 대해서 어떠한 사전정보나 부가적인 정보가 필요 없다는 것이라 할 수 있다.

동치 관계에 의해 정보 객체 집단은 동치류(equivalence class)로 구분될 수 있으며, 이들 동치류 원소의 집합을 기본 집합이라 하고, 이 기본 집합에 의해 정의되는 집합 공간을 근사(approximation) 공간이라고 한다. 근사 공간상에 하나의 결정에 대한 정보 객체를 분류하는 경우, 동일한 기본 집합 내에 있으면서도 서로 다른 결정을 내는 경우가 발생할 수 있다. 이런 결정상의 불일치(inconsistency)를 나타내고 처리하기 위해서 리프 셋 이론에서는 두 가지 근사를 정의한다.

하나는 결정에 의해 나타내어지는 개념 X에 항상 포함되는 기본 집합으로 정의되는 하한 근사이고, 다른 하나는 개념 X와 일치하는 부분이 하나라도 존재하는 모든 기본집합으로 정의되는 상한 근사이다.

U를 전체 집합이라 하고 R를 U에 대한 동치 관계라 하자. A=(U, R)은 근사 공간이 되며, 하한 근사와 상한 근사는 다음과 같이 표현된다.

$$\underline{R}X = \{x \in U : [x] \subseteq X\},$$
$$\overline{R}X = \{x \in U : [x] \cap X \neq \varnothing\},$$

여기서 [x]는 원소 x를 포함하는 R의 동치류를 나타낸다.

X의 러프 셋은 다음과 같이 정의된다.

$$A_R(X) = (\underline{R}X, \overline{R}X)$$

또한, 상한 근사에서 하한 근사를 제외시키면 불확실한 개체들만 또 다른 부분 집합으로 표현될 수 있으며, 이를 경계 영역이라 부르며, 다음과 같이 표현한다.

$$BN_R(X) = \underline{R}X - \overline{R}X$$

러프 셋 이론을 이용한 연구는 현실세계에 존재하는 불확실한 데이터를 다루는 데 있어 매우 유용하다는 평가를 받고 있기에 적용 분야 또한 대단히 넓다. 예를 들어 인공지능, 인지과학, 의료 데이터 분석, 패턴인식 등과 같은 응용 분야가 있으며, 데이터 마이닝 분야에도 널리 활용 되고 있다. S.K. De[9]에서는 러프 셋 이론을 이용한 클러스터링 방법을 제안하고 있는데, 본 연구와 달리 순서를 고려하지 않은 웹 트랜잭션들을 대상으로 연구를 수행하였다. 그러나 최근에는 소매점 거래 데이터, 단백질 시퀀스, 웹 로그 등과 같은 순서적인 면을 고려한 시퀀스 데이터들의 폭발적인 증가를 볼 수 있다. 따라서 본 연구에서는 이들 시퀀스 데이터들을 분석하는데 있어 러프 셋 이론을 이용하는 새로운 알고리즘을 제안한다.

2. 클러스터링 방법과 Web usage mining

클러스터링이란 패턴들을 그룹들로 나누는 비교사 학습 분류이며, 클러스터링 문제는 모집단을 클러스터들로 구분하는 문제이다[10]. 모집단이란 m 개의 속성들로 이루어진 n 요소들의 집합이다. 클러스터링의 목표는 적당한 유사도 측정 방법에 의하여 패턴들을 유사한 클러스터들로 그룹화 하는 것이다. 즉, 유사한 패턴들은 동일 클러스터에 할당이 되는 반면에 확실하게 구별이 되는 패턴들은 서로 다른 클러스터들에 할당이 되도록 하는 것이다.

웹 사이트에는 많은 양의 컨텐츠가 있으며, 하이퍼링크 구조가 복잡하게 얽혀 있다. 최근에는 web usage mining 분야에서 클러스터링 문제에 대한 많은 연구가 진행되고 있으며, 웹 로그 데이터에 대하여 적응적인 데이터 마이닝 기법을 적용하려는 연구가 진행되고 있다[11]. Schechter et al. [12]는 HTTP 요청들을 예측하기 위해 사용자 프로파일들을 이용하는 기법을 개발하였고, Cooley [13]는 웹 로그로부터 사용자 패턴을 추출하기 위해 데이터 마이닝 기법을 응용하였다.

본 연구에서는 러프 셋 이론을 이용하여 시퀀스 데이터를 클러스터링 하는 방법을 제안한다.

III. 러프 셋을 이용한 제안하는 방법

1. 러프 셋을 이용한 접근 방법

이절에서는 시퀀스 데이터를 클러스터링 하기 위한 러프 셋 접근 방법을 제안한다.

예를 들어, 사용자 트랜잭션은 사용자가 방문한 웹 페이지들로서 시퀀스로 표현하며, 웹 페이지는 항목으로 표현한다. 즉, 시퀀스 t는 n개의 항목들의 모임이며 〈 x1 x2 … xi … xj … xn 〉으로 표시하고, 여기서 xi 는 웹 페이지로서 범주형 값을 가지는 항목이다.

t의 크기는 t에 있는 항목들의 개수이며, | t | 로 나타낸다. 시퀀스 t에서 순서를 가지는 2개의 항목들로 구성된 xi xj (i<j)를 시퀀스 요소 ek 라고 하며, ek 들의 모임을 E = (e1, e2 , … , ek, …)라 한다. E의 크기는 E에 있는 요소들의 개수이며, | E | 로 나타낸다.

[예 3.1] 시퀀스 t = 〈 A B C E 〉에서 | t | = 4이고, 시퀀스 요소들의 모임은 E = (AB, AC, AE, BC, BE, CE)이며, | E | = 6이다.

시퀀스내의 항목들뿐만 아니라 항목들 간의 순서도 고려를 해서 식(3.1)과 같이 유사도 계산 방법을 제안한다.

[정의 3.1] 두 시퀀스 t1 = 〈 a1 a2 … an 〉과 t2 = 〈 b1 b2 … bm 〉의 시퀀스 요소들의 모임을 각각 E1 = (ea1, ea2, … , eai , …), E2 = (eb1, eb2, … , ebj , …)라고 하면, t1, t2 의 유사도 sim(t1, t2)는 다음과 같이 정의한다.

$$sim(t1,t2) = \frac{|E1 \cap E2|}{\frac{|E1| + |E2|}{2}} \quad \cdots \quad (3.1)$$

여기서, | E1∩E2 | 는 E1과 E2의 공통 요소들의 개수이며, E1과 E2사이에 공통 항목들이 많을수록 유사도는 높고, 이 값을 (| E1 | + | E2 |)/2로 나누는 것은 유사도를 0과 1사이의 값을 갖도록 하기 위해서이다.

[정의 3.1]의 유사도 계산 방법은 오승준[14]과 오승준

외[15]에서 사용한 유사도 계산 방법과 동일한 방법이다.

[예 3.2] 두 시퀀스 t1 = 〈 A B D A 〉, t2 = 〈 A C D A C 〉에서 시퀀스 요소들의 모임은 각각 E1 = (AB, AD, AA, BD, BA, DA) 과 E2 = (AC, AD, AA, AC, CD, CA, CC, DA, DC, AC) 이며, | t1 | = 6, | t2 | = 10, t1∩t2 = (AD, AA, DA), | t1∩t2 | =3이다. 따라서, 두 시퀀스의 유사도 sim(t1, t2)는 3/8이다.

m명의 사용자들이 존재한다면 사용자 트랜잭션은 $T = \{t_1, t_2, t_3, \ldots, t_m\}$ 이고, 사용자들이 클릭한 고유한 항목들의 집합은 U라 한다.

임의의 값 Θ와 두 명의 사용자 트랜잭션 t와 s에 대하여 T에 대한 관계 R은 tRs로 표현되며, 다음과 같이 정의된다.

$$tRs \text{ iff } sim(t,s) \geq \Theta$$

[정의 3.2] t에 대한 유사 클래스 R(t)는 t와 유사한 트랜잭션들의 집합이며, 다음과 같이 정의된다.

$$R(t) = \{ s \in T, sRt \}$$

Θ값의 변화에 따라 서로 다른 유사 클래스를 얻을 수 있다.

[정의 3.3] $P \subset T$라 하고, 고정된 Θ 값에 대하여, T에 대한 관계 R이 정의되면, 하한 근사 P와 상한 근사 P는 각각 다음과 같이 정의된다.

$$\underline{R}(P) = \{ t \in P, R(t) \subseteq P \}$$
$$\overline{R}(P) = \bigcup_{t \in P} R(t)$$

2. 제안하는 클러스터링 방법

상한 근사 $R(t_i)$는 t_i 와 유사한 트랜잭션들의 집합으로서, ti 내의 항목(페이지)들을 방문한 사용자들은 $R(t_i)$의 트랜잭션들내의 항목들을 방문할 가능성이 높다. 이와 유사하게, $RR(t_i)$는 $R(t_i)$와 유사한 트랜잭션들의 집합이며, 이 과정은 두 개의 연속적인 상한 근사 값이 똑같을 때까지 반복된다. 이것을 유사도 상한 근사라 부르며, $\underline{S}i$로 표현한다.

본 연구에서 제안하는 클러스터링 방법은 다음과 같다.

입력 : n 개의 시퀀스 데이터들과 임의의 Θ 값 ($0 \leq \Theta \leq 1$)

단계 1 : 하나의 시퀀스 데이터를 하나의 클러스터에 할당
$C_i = \{t_i\}$, $C = \{C_1, C_2, \ldots, C_n\}$

단계 2 : 각각의 클러스터 $C_i \in C$와 Θ에 대한 유사도 상한 근사 S_i를 구한다.

단계 3 : $S_i = S_j$ $(i \neq j)$인 모든 $C_i \in C$에 대하여 C_i 클러스터들을 합병하고 C를 업데이트 한다.

단계 4 : C를 출력한다.

그림 1. 제안하는 알고리즘
Fig. 1. The proposed algorithm

n 개의 시퀀스들과 임의의 Θ 값에 대하여, 단계 1에서는 하나의 시퀀스를 각각 하나의 클러스터에 할당한다. 단계2에서는 유사도 상한 근사를 구하며, 단계 3에서는 동일한 유사도 상한 근사를 갖는 클러스터들을 합병하고, 단계 4에서 최종적으로 클러스터들을 출력하며 알고리즘을 끝마친다.

IV. Case Study

이번 장에서는 본 논문에서 제안하는 알고리즘의 절차를 보여주기 위한 예제를 소개한다.

사용자 트랜잭션들의 집합은 $T = \{t_1, t_2, t_3, t_4\}$이며, 사용자 트랜잭션들이 방문한 고유한 하이터링크들의 집합은 U = {A, B, C, D, E}이다. t_1 = { A, B, C }, t_2 = { A, C, D }, t_3 = { A, B, D, E }, t_4 = { A, C, D, E }라 하자.

그림 1의 단계 1에서 C_1 = { t_1 }, C_2 = { t_2 }, C_3 = { t_3 }, C_4 = { t_4 }가 된다.

단계 2에서 각 트랜잭션들간의 유사도를 구하면 다음과 같다.

sim(t1, t2) = 1/3, sim(t1, t3) = 2/9,
sim(t1, t4) = 2/9, sim(t2, t3) = 2/9,
sim(t2, t4) = 2/3, sim(t3, t4) = 1/2

여기서 θ = 0.5라 하면,

$R(t1) = \{t1\}$, $R(t2) = \{t2, t4\}$, $R(t3) = \{t3, t4\}$,
$R(t4) = \{t2, t3, t4\}$,
$\overline{R}(t1) = \{t1\}$, $\overline{R}(t2) = \{t2, t4\}$, $\overline{R}(t3) = \{t3, t4\}$,
$\overline{R}(t4) = \{t2, t3, t4\}$,
$\overline{R\overline{R}}(t1) = \{t1\}$, $\overline{R\overline{R}}(t2) = \{t2, t3, t4\}$,
$\overline{R\overline{R}}(t3) = \{t2, t3, t4\}$, $\overline{R\overline{R}}(t4) = \{t2, t3, t4\}$,
$\overline{R\overline{R}\overline{R}}(t2) = \{t2, t3, t4\}$, $\overline{R\overline{R}\overline{R}}(t3) = \{t2, t3, t4\}$

다음으로 유사도 상한 근사를 구하면 다음과 같다.

$S1 = \{t1\}$, $S2 = \{t2, t3, t4\}$, $S3 = \{t2, t3, t4\}$,
$S4 = \{t2, t3, t4\}$

단계 3에서 두 개의 클러스터 {t1}, {t2, t3, t4}를 얻게 된다. 즉, t2에 속한 하이퍼링크들을 클릭한 사용자는 또한 t3나 t4내의 하이퍼링크들을 방문할 가능성이 높다.

V. 실험결과

본 논문에서 제안하는 방법을 평가하기 위해 splice 데이터셋과 합성 데이터셋으로 실험을 수행하였다. 본 실험은 인텔 3.0 GHz 사양의 펜티엄 IV 컴퓨터에서 C++ 언어로 코딩을 하여 수행하였다.

1. splice 데이터셋

splice 데이터셋은 UCI KDD 아카이브에 포함되어 있는 데이터셋이다[16]. 이 데이터셋은 60개의 항목을 가지는 뉴클레오타이드 시퀀스들을 포함하고 있으며, 각각의 시퀀스들은 EI나 IE에 속하는 클래스 레이블을 가진다. EI에 속하는 시퀀스들이 767개이며, IE에 속하는 시퀀스들이 768개이다.

splice 데이터셋을 3장에서 제안하는 알고리즘으로 θ 값을 변화시키며 실험하였다. 표 1에는 θ 값 이상의 값을 갖는 두 쌍의 시퀀스들 개수를 표시하였고, 그림 2에는 θ 값의 변화에 따른 최종 클러스터 개수를 표현하였다.

표 1. θ 값 변화에 따른 두 쌍의 시퀀스들 개수
Table 1. θ values and number of two pairs sequences

θ 값	0.90	0.92	0.94	0.96
splice dataset	810	510	346	275

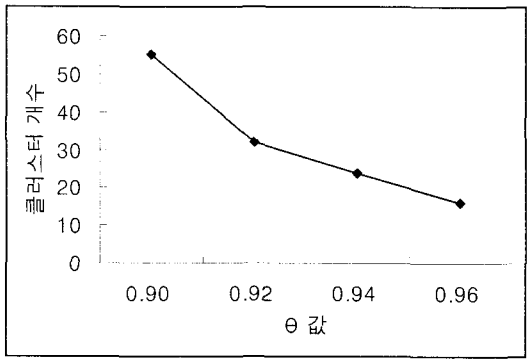

그림 2. θ 값 변화에 따른 클러스터 개수
Fig 2. θ values and number of clusters

표 1과 그림 2에서 보는 바와 같이 θ 값이 증가 할수록 두 쌍의 시퀀스들 개수와 클러스터 개수는 감소한다. 이는 θ 값이 증가 할수록 높은 유사도를 갖는 시퀀스들만을 같은 클러스터로 클러스터링 하기 때문이다.

2. 합성 데이터셋

본 논문에서 제안하는 알고리즘의 성능을 평가하기 위해 Quest 프로젝트의 합성 데이터셋 생성기[17]를 응용하여 합성 데이터셋 3개를 생성하였다. 이를 각각 DS1, DS2, DS3라 하며, DS1은 총 시퀀스 개수가 1,000개이며, DS2는 2,000개, DS3는 3,000개이다.

합성 데이터셋을 3장에서 제안하는 알고리즘으로 θ 값을 변화시키며 실험하였다. 표 2는 θ 값 이상의 값을 갖는 두 쌍의 시퀀스들 개수를 표시하였고, 그림 3에는 θ 값의 변화에 따른 최종 클러스터 개수를 표현하였다.

표 2. θ 값 변화에 따른 두 쌍의 시퀀스들 개수
Table 2. θ values and number of two pairs sequences

θ 값	0.35	0.40	0.45	0.50
DS1	119	42	18	8
DS2	541	249	112	40
DS3	1773	821	379	173

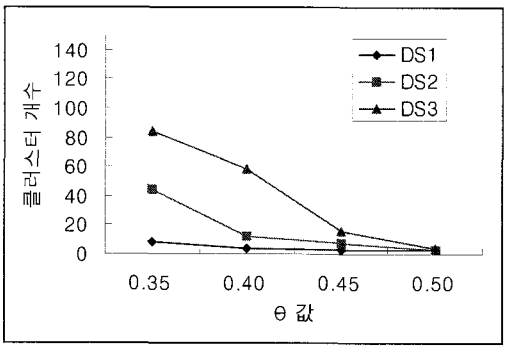

그림 3. θ 값 변화에 따른 클러스터 개수
Fig 3. θ values and number of clusters

표 2와 그림 3에서 보면, splice 데이터셋의 경우와 마찬가지로 θ 값이 증가 할수록 두 쌍의 시퀀스들 개수와 클러스터 개수는 감소한다.

VI. 결론

본 논문에서는 러프 셋 이론을 이용하여 시퀀스 데이터들을 클러스터링 하는 방법을 제안하였다. 본 논문에서 제안하는 방법을 이용하면, 웹 사용자의 방문 순서까지 고려하여 웹 트랜잭션을 클러스터링 할 수 있다. 이러한 방법은 웹 로그 데이터로부터 흥미로운 사용자 행동 패턴들을 발견하는데 유용하다. 또한 웹 사용자 행동 패턴들은 웹 사이트 디자이너들에게는 개인적인 행동들에 기반을 둔 적응적인 웹 사이트를 만드는데 도움이 된다.

웹 사용자들의 특성을 분석하기 위해 러프 셋 이론을 이용하였는데, 이 방법을 통해 데이터에 대한 사전정보나 부가적인 정보가 필요 없이 분석을 수행할 수 있었다.

향후 연구과제로는 효율적으로 θ 값을 설정하는 방법과 다양한 데이터 셋에 대하여 제안하는 알고리즘의 성능을 평가하는 것이 필요하다.

참고문헌

[1] O. Etzioni. "The world wide web: Quagmire or gold mine". Comunications of the ACM, pp 65-68, 1996

[2] R. Kosals, H. Blockeel, "Web mining research: a survey", ACM SIGKDD, July 2000

[3] S.K. Madira, S.S. Bhowmick, W.K. Nag, E-P. Lim, "Research issues in web data mining", Proc. Data Warehousing and Knowledge Discovery, First Int. Conf. DaWaK99, pp 303-312, 1999.

[4] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan. "Web usage mining: discovery and applications of usage patterns from web data", SIGKDD Explorations, 1, 2000

[5] J. Borges and M. levene, "Mining association rules in hypertext databases", In Proc. of the Fourth Int'l Conference on Knowledge Discovery and Data Mining, Aug. pp 27-31, 1998

[6] A. Buchner, M. Baumgarten, S. Anand, M. Mulvenna, and J. Hughes, "Navigation pattern discovery from internet data", In Proc. of the WEBKDD '99, Aug. 1999

[7] Z. Pawlak, Rough sets: theoretical aspects of reasoning about data, Kluwer Academy Publisher, 1991.

[8] S. K. De, P. R. Krishna, "Clustering web transactions using rough approximation", Fuzzy sets and systems, 148, pp 131-138, 2004.

[9] J. Han, M. Kamber, Data Mining: concepts and techniques, Morgan Kaufmann publishers, 2000.

[10] M. perkowitz, O. Etzioni, "Towards adaptive web sites: conceptual framwork and case study", Artificial Intelligence, 118, pp 245-275, 2000.

[11] S. Schechter, M. Krishnan, M. D. Smith, "Using path profiles to predict HTTP requests", Comput. Networks ISDN Systems, 30, pp

457-467, 1998.

[12] R. Cooley, Web usage mining: discovery and application of interesting patterns from web data, Ph. D. Thesis, Department of computer science, Univ. of Minnesota, 2000.

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

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

[15] C. L. Blake and C. J. Merz, UCI repository of machine learning databases, 1998.

[16] R. Agrawal, M. Mehta, J. Shafer, R. Srikant, A. Arning and T. Bollinger, "The Quest Data Mining System", Proc. 2nd Int. Conf. KDD, Portland, 1996.

저 자 소 개

오 승 준

2004년 8월 한양대학교 산업공학과 공학박사

2005~ 현재 : 경기공업대학 산업경영과 교수

박 찬 웅

1997년 2월 한양대학교 산업공학과 공학박사

1997~ 현재 : 경원대학교 산업정보시스템공학과 교수