상황인식정보 추출을 위한 클러스터링 알고리즘 기반 사용자 구분 알고리즘

김민섭* · 신인호 · 정병훈 · 손지원 · 조아현 · 도윤형 · 이강환*
*한국기술교육대학교

Context-awareness User parameter Analysis based on Clustering Algorithm

Min-seop Kim* · Shin-in Ho · Byoung-hoon Jung · Ji-won Son · Ah-hyeon Jo · yun-hyung do · Kang-whan Lee*
*Korea University of Technology and Education
E-mail : appleeji@koreatech.ac.kr

요 약

본 논문에서는 개개인의 사용자 상황인식 정보추출을 위해 구분이 필요한 시스템에서의 클러스터링 알고리즘을 이용한 대체 방법에 대한 알고리즘을 제안한다. 기존의 사용자 구분 시스템에서는 사용자가 직접 자신의 정보를 입력해야 하는 번거로움이 있었다. 본 논문에서는 이러한 사용자 관리 기반에 있어 개선된 알고리즘을 연구 적용한 상용자 인식정보 추출이 가능한 클러스터링 알고리즘을 적용한 시스템을 연구 개발하고자 한다. 일반적으로 같은 데이터를 가진 사용자들을 구분하는 알고리즘은 기록된 정보와 새로 입력된 정보가 일치하는지 확인 후 그에 따라 적절한 대처를 해준다. 하지만 그 새로 입력된 정보가 어떤 사용자의 정보인지를 직접 입력해줘야 하는 번거로움이 발생한다. 따라서 본 논문에서는 사용자 정보를 직접 입력하지 않아도 누적된 시스템 내의 워킹 메모리로부터 분석된 데이터를 바탕으로 시스템 스스로 클러스터링 알고리즘을 이용하여 사용자를 구분하는 방법을 제안한다. 연구 적용된 알고리즘을 적용한 시스템의 관리 기법이 기존의 시스템보다 인원 구성이 다양한 환경에서 적응성이 더 높음을 보여주었다(주관적 관찰자 실험방법으로 증빙).

ABSTRACT

In this paper, we propose an algorithm for an alternative method using the clustering algorithm in a system that needs classification to extract individual user context information. In the conventional user classification system, the user has to input his own information. In this paper, we will research and develop a system applying a clustering algorithm which can extract user 's perceived information applying the improved algorithm for user management base. Generally, the algorithm that distinguishes users with the same data makes sure that recorded information matches the newly entered information, and then responds accordingly. However, it is troublesome to manually input information of the new user. Therefore, in this paper, we propose a method to distinguish users by using the clustering algorithm based on the analyzed data from the working memory in the accumulated system without directly inputting the user information. The study shows that the management method applied to the applied algorithm is more adaptive in environments where the number of people is different from that of the existing system (as a subjective observer test method).

키워드

클러스터링 알고리즘, 다중 사용자, 적응성, 자동 분석

Ⅰ. 서 론

최근 컴퓨터기술이 발전하면서 사용자간의 가상 연결망 또한 발전하여 사용자들에 대한 보다 향상된 사용자 관리기법이 개발되고 있는 추세이다. 일례로, 사용자 계정을 만들어 여러 회사에서 제공하는 전자 서비스들을 받을 수 있는데 이러한 계정을 생성할 때 사용자는 직접 자신의 정보를 입력해야 하고, 여러 전자 서비스들을 받기 위해서는 여러 개의 계정을 만들어야 하는 번거로움을 겪는다. 뿐만 아니라 계정을 만들 때 필요한 개인정보는 완벽하게 보호받지 못하고 있는 실정이다[1]. 개인정보를 도용하여 계정을 생성할 경우 사용자가 누구인지 완벽하게 알 수 없기 때문에 쉽게 악용될 수 있다. 이처럼 계정 관리 및 계정 접속은 사람들에게 좋은 서비스이지만 거슬리는 계륵이라 느끼게 할 것이다. 이는 자연스러운 생활의 일부분이라기보다는 인위적인 시스템이 될 것이다. 또한, 기존에 자동으로 사용자를 구분할 수 있는 서비스가 있더라도 같은 데이터 값을 가진 사용자의 경우 같은 사용자로 인식하는 문제가 발생할 수 있다.

사용자들은 점차 자신에게 가장 알맞은 시스템 환경을 자동으로 제공 받기를 원하고, 자동으로 시스템 환경을 제공해주기 위해서는 사용자 구분의 필요성이 대두되고 관심이 높아질 것이다. 따라서 본 논문에서는 여러 명의 사용자가 한 시스템을 사용하는 환경에서의 사용자의 특징이 되는 특성 정보를 분석하여 사용자를 구분하는 방법을 제안한다. 이를 이용하여 사용자 유형에 따른 최적의 업무 환경을 구축할 수 있고, 샤워 시 최적의 물의 온도를 자동으로 제공하는 등 여러 환경 중에서 사용자에게 가장 알맞은 환경을 자동으로 제공할 수 있을 것이다.

Ⅱ. 관련 연구

샤오미의 스마트 체중계는 측정한 체중을 이용하여 사용자를 구분한다. 체중계와 스마트 폰을 블루투스로 연결하고, 어플을 연동하여 사용자가 방금 측정한 몸무게를 자동으로 불러온다. 측정한 몸무게는 자신의 계정의 일부로 등록하고 성별과 나이, 신장 등을 추가로 입력한다. 한번 자신의 체중이 등록이 되면 다음에 사용자가 몸무게를 측정할 때 별다른 설정을 하지 않아도 기존의 데이터베이스에 저장된 몸무게와 비교해 변화 값이 3kg이내이면 같은 사용자로 인식한다. 사용자의 몸무게 변화를 계속해서 저장하고 판단하여 건강관리 서비스를 제공하여 준다. 하지만 체중계를 사용하기 위해서는 샤오미 계정부터 만들어야 하는 번거로움이 있고, 만일 다른 사용자와 몸무게가 3kg이하의 차이가 나면 같은 사용자로 인식하는 한계점이 있다.

Ⅲ. 본 론

본 논문에서는 보다 향상된 사용자 관리기법을 제공하기 위해서 기존에 누적된 사용자의 데이터를 통한 K-means Clustering 알고리즘[2][3]을 사용하여 사용자 구분을 하려고 한다. 적용하고자 하는 k-means Clustering 알고리즘이란 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다. 이 알고리즘은 자율 학습의 일종으로, 레이블이 달려 있지 않은 입력 데이터에 레이블을 달아주는 역할을 수행한다.[4][5]

본 연구에서는 다양한 사용자를 구분하기 위해서 인바디 센서[6]를 통해 제공받은 데이터를 이용한다. 데이터로는 골격근량, 체지방량, 체수분, 단백질, 무기질, 몸무게 6개로 구성된다. 이 사용자를 나타내는 6개의 특징을 이용함으로써 위에서 보였던 스마트 체중계와는 다르게 좀 더 정교하게 사용자를 구분 할 수 있다.

1. K-means Clustering 알고리즘을 이용한 사용자 구분 알고리즘

K-means Clustering 알고리즘을 이용한 사용자 구분 알고리즘은 이 6가지의 특성들을 K-means clustering 알고리즘에 feature로 사용한다. 이런 정보를 가지는 n명의 데이터셋을 $X_1, \cdots, X_n$ 이라 하자. 총 N개의 데이터로 D개의 차원을 가지며 K개의 클러스터로 나눌 것이다. D개의 차원을 가지는 벡터를 μ처럼 표기하고 이 벡터가 k번 째 클러스터에 속하는 경우 μ_k와 같이 표기한다. 우리의 최종 목표는 주어진 데이터 집합으로부터 이러한 중심점 μ_k의 값을 결정하는 것이다. 위와 같은 문제는 데이터 집합을 특정 클러스터에 할당하는 과정으로 표기하는 것이 좋다. 그러기 위해 우선 변수 $r_{nk} \in 0,1$을 정의한다. 이때의 k 값은 당연히 $k = 1, \cdots, k$이다. 이는 어떤 n 번째 샘플 X_n이 k번째 클러스터에 속하는 경우 $r_{nk} = 1$이고 아닌 경우 0이 된다. 이제 목적 함수를 정의한다. 이를 왜곡 측정(distortion measure) 함수라고 한다.

N	전체 데이터셋의 개수
K	나누려는 클러스터의 개수
x_n	n번째 유저의 데이터
μ_k	k번째 클러스터에 속하는 값들의 평균
r_{nk}	어떤 n번째 샘플x_n이 k번째 클러스터에 속하는 경우 1이고 아닌 경우 0이 된다.

$$J=\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\|X_n-\mu_k\|^2$$

이 목적함수 J값이 최소가 될 때의 r_{nk}와 μ_k의 값을 구해야한다. 이를 위해서 아래와 같은 방법으로 r_{nk}의 값을 구한다.

$$r_{nk}=\begin{cases}1 & \text{if } k=\arg min_j\|X_n-\mu_j\|^2\\ 0 & otherwise\end{cases}$$

즉, 각 클러스터 중심과 샘플의 거리를 측정해서 가장 가까운 클러스터를 선택한다. 다음으로 μ_k의 값을 구한다. r_{nk}의 값을 고정하고 목적함수 J를 미분하여 최솟값이 되는 지점을 알아 낼 수 있다.

$$2\sum_{n=1}^{N}r_{nk}(X_n-\mu_k)=0$$

이를 전개하면,

$$\mu_k=\frac{\sum_n r_{nk}X_n}{\sum_n r_{nk}}$$

이와 같은 식을 얻을 수 있다. 결국 k클러스터에 속한 값들의 평균을 구하게 된다. 두 단계를 거치는 동안 데이터는 각각의 클러스터에 다시 할당이 되고, 이렇게 재 할당된 데이터를 이용하여 평균값을 다시 계산하는 과정을 반복하게 된다.

본 시스템의 흐름도는 그림 1과 같다.

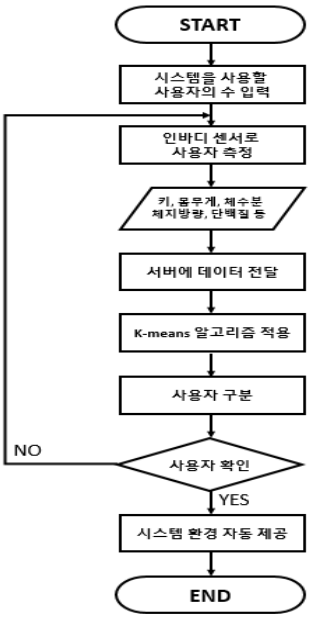

그림 1

우선적으로 군집의 수를 지정하기 위해 시스템에서 사용할 사용자의 수를 입력받는다. 이후, 인바디 센서로 측정된 사용자의 특성 정보를 받아 서버에 전달한다. 모여진 데이터들을 k-means Clustering 알고리즘에 적용하여 사용자를 구분한다. 사용자에게 feedback을 받아 사용자 본인이 맞는지 확인한다. 맞다면 사용자에게 최적화된 시스템 환경을 자동으로 제공한다.

2. 사용자 구분 알고리즘 실험

위의 알고리즘을 실행하기 전에는 그림 2와 같이 나타난다. 하지만 이 알고리즘을 이용해서 학습하면 처음의 랜덤한 값들이 클러스터링 된 곳으로 이동하고 사용자를 특정 지을 수 있게 되며 그림3과 같이 나타난다.

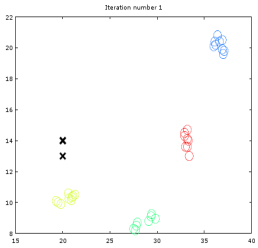

그림 2

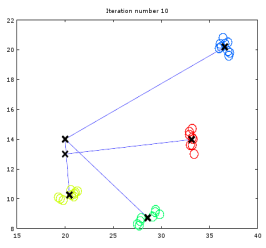

그림 3

만약 사용자가 다이어트를 해서 몸무게가 급감한다거나 수분이 부족해서 체수분지수가 급감한다고 하더라도 다른 특징들의 가중치를 이용해서 사용자를 특정 지을 수 있다. 그림4는 의도적으로

초록색의 골격근량을 빨간색과 같이 만든 것이다. 하나의 특성 요소가 같은 값이 되더라도 나머지 5가지의 특성 요소를 분석하여 알고리즘이 사용자를 식별하는데 문제가 생기지 않는다. 그림 5는 변경한 데이터의 수치를 나타내는 도표이다.

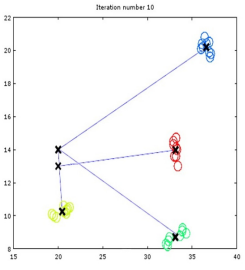

그림 4

파란색 골격근량 (변경전)	파란색 골격근량 (변경후)	빨간색 골격근량
36.0	33.2	33.2
36.1	33.0	33.0
36.2	33.3	33.3
36.9	33.4	33.4
36.4	33.2	33.2
36.9	32.9	32.9
37.0	32.9	32.9
37.1	33.2	33.2

그림 5

Ⅳ. 결 론

본 논문에서는 누적된 인바디 데이터로부터 k-means Clustering 알고리즘을 사용하여 사용자 구분을 하는 기법을 제안한다. 인바디 데이터로부

터 k-means Clustering 알고리즘을 활용해 다수의 사용자가 있는 시스템에서도 계정을 따로 만들 필요 없이 사용자의 구분을 명확히 할 수 있다. 또한, 6가지의 사용자 특성 정보를 이용하여 사용자를 구분하기 때문에 사용자의 정보가 급격하게 변화하더라도 같은 사용자로 인식이 가능하여 정보의 수정이 필요가 없어 유연하게 반응 할 수 있다. 기존 기술에서는 다중 사용자를 구분할 때 계정을 등록해야 하는 번거로움이 있다. 또한 같은 데이터를 가진 사용자를 구분해내지 못한다. 본 연구에서는 누적된 데이터를 인바디 데이터로 한정하여 사용하였지만 사용자를 구분할 수 있는 명확한 정보가 있다면, 다른 데이터를 사용하여도 사용자를 구분할 수 있을 것이라고 기대한다.

참고문헌

[1] 최해원, “I-pin의 위험성 분석과 개선에 관한 연구”, 학위논문(석사)--단국대학교 정보미디어대학원, pp. 12-20, 2014

[2] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm”, Journal of the Royal Statistical Society. Series C (Applied Statistics), pp. 100-108, 1979

[3] Aristidis Likas and Nikos Vlassis and Jakob J.Verbeek, “The global k-means clustering algorithm”, Department of Computer Science, University of Ioannina, 45110 Joannina, Greece, pp 452-456, 2003

[4] Lining Xue and Weixin Luan, Improved K-means Algorithm in User Behavior Analysis, Dalian Maritime University, pp.339-341, 2015

[5] TANVI SHEIKH and SHIKHA AHRAWAL, “Enhanced K-means based Facial expressions recognition system”, CSE Department CSIT Durg, Assistant Professor CSE Department CSIT Durg, pp. 39-41, May-2013

[6] Takuya Obichi and Yaraka Okaie, Tadashi Nakano, Takahiro Hara, Shojiro Nishio, “Inbody Mobile Bionanosensor Networks Through Non-diffusion-based Molecular Communication”, Department of Information and Communications Technology, Osaka University, Japan, pp. 1078-1083, 2015