생활 스포츠 콘텐츠 기반의 프로파일 처리 알고리즘 연구

고은미* · 안나영* · 이재동* ·이원진**

*단국대학교 소프트웨어학과, **단국대학교 정보문화기술연구원

A Study on Profile Processing Algorithm based on Sport for All Contents

Eun-mi Ko* · Na-Young An* · Jae-Dong Lee* · Won-Jin Lee**

*Dept of Software, Dankook University, **RICT, Dankook University

E-mail : koeunmio@daum.net, any9412@naver.com, letsdoit@dankook.ac.kr, god7300@dankook.ac.kr

요 약

본 논문에서는 생활 스포츠 콘텐츠 기반의 프로파일 처리 알고리즘에 대하여 제안한다. 제안한 알고리즘은 맞춤형 생활 스포츠 콘텐츠를 추천을 위해 필요한 연구이며, 추천의 신뢰성을 높이기 위해 선행되어야 할 연구이다. 그래서 제안한 알고리즘은 추천 시 고려되는 동적 정보를 포함하는 동적 프로파일을 처리하고, 추천 분류에 따라 변화되는 가중치 값을 처리할 수 있는 동적 프로파일 알고리즘을 제안하였다. 제안한 프로파일 처리 알고리즘은 콘텐츠 추천의 만족도 향상을 기대한다.

ABSTRACT

In this paper, we propose the profile processing algorithm based on in-life sports contents. The proposed algorithm is required research for recommending to sport for all contents, and is preceding research to improve reliability of recommendation. So the proposed algorithm processing dynamic profile based on dynamic information for recommendation, and processing weight values that depending on dynamic recommendation classification. The proposed profile processing algorithm is expected to improve satisfaction of contents recommendation.

키워드

생활 스포츠 콘텐츠, 프로파일, 맞춤형, TF-IDF

Ⅰ. 서 론

최근 생활 스포츠 팀의 증가는 생활 스포츠 시장 발전이 높아지고 있다. 특히 ICT가 접목된 스포츠 융합 콘텐츠 및 서비스 시스템 연구의 중요성이 높아지고 있다

본 논문에서는 생활 스포츠 활성화를 위한 맞춤형 스포츠 콘텐츠 큐레이션 시스템 구축을 위한 프로파일 처리 알고리즘에 대해 제안하였다. 여기에서 맞춤형 스포츠 콘텐츠 큐레이션 시스템이란, 개인과 팀 단위의 프로파일 특성에 맞춤화된 생활 밀착형 스포츠 융합 콘텐츠를 추천하는 시스템이다. 이러한 시스템을 위해서는 개인과 팀의 프로파일 정보가 필요하며, 프로파일 정보를 이용하여 개인과 팀에게 맞춤형 콘텐츠를 추천할 수 있다. 맞춤형 콘텐츠를 추천하기 위해서는 추천시 사용되는 다양한 프로파일 속성 정보와 사용자에 의해 전달되는 다양한 피드백 정보를 고려해야 한다. 즉 추천에 의해 변화되는 프로파일 정보는 동적 프로파일 정보로 분류하고, 처리 및 분석 시 신뢰성을 높이기 위한 알고리즘 처리 기술이 필요하다.

그래서 본 논문에서는 맞춤형 생활 스포츠 콘텐츠를 제공에 필요한 동적 프로파일을 처리하는 알고리즘을 제안한다.

Ⅱ.관련연구

프로파일 처리 및 분석 알고리즘과 관련된 기준 연구 중 대표적인 것이 TF-IDF(Term

Frequency - Inverse Document Frequency) 기법이다. TF-IDF는 정보 검색과 텍스트마이닝에서 이용하는 가중치로, 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하거나, 문서들 사이의 비슷한 정도를 구하는 등의 용도를 사용할 수 있다.

TF(Term Frequency, 단어 빈도)란, 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값으로, 이 값에 따라 문서 내에서 중요한 정도를 측정한다. 문서군 내에서 자주 사용되는 단어를 DF(Document Frequency, 문서 빈도)라 일컫는다. 따라서 IDF(Inverse Document Frequency)는 문서군 자체에서는 잘 사용되지 않는 단어의 빈도수를 의미하며 TF-IDF는 TF와 IDF의 곱으로 나타낸다. 논문 [1]에서는 한 세션 동안 사용자가 검색결과에 탐색한 웹문서들을 수집하여 문서집합을 구성한 후, TF-IDF 가중치를 이용하여 사용자가 탐색한 문서집합으로부터 적절한 키워드를 추출하는 방법을 사용하였다. 또한 논문 [2]에서는 기존의 TF-IDF를 변형하여 키워드 추출 연구에 적용하였다. 하나의 문서가 아닌 문서집합 수준에서 키워드를 추출하기 위해 기존의 TF-IDF 모델의 범위를 개별적인 문서에서 문서집합으로 확장시킨 것이다. 이를 위해 [2]에서는 두 가지의 정규화된 TF(Normalized Term Frequency)를 제안하며, 이를 NTF1, NTF2라 한다. 이 두 가지의 정규화된 TF를 이용하여 문서 길이가 달라서 생기는 가중치의 과도한 편차를 최소화한다.

Ⅲ. 제안한 프로파일 처리 알고리즘

본 장에서는 본 논문에서 제안하는 맞춤형 스포츠 콘텐츠 큐레이션 시스템 구축에 필요한 프로파일 처리 알고리즘에 대해 기술한다. 그림 1은 전체 프로파일 시스템의 구조이며. 정보 처리 과정은 다음과 같다.

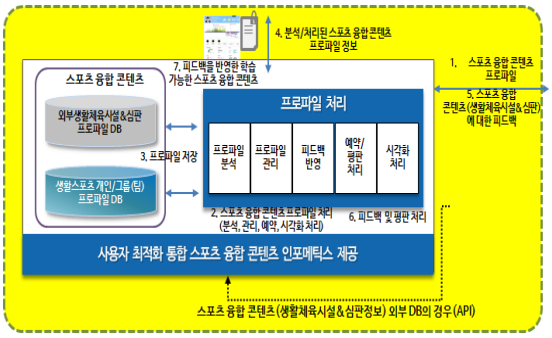

그림 1. 프로파일 시스템 구조

1) 프로파일 수집, 2)프로파일로 처리, 3) 프로파일 저장,. 4) 분석/처리된 프로파일 정보를 관리 5) 피드백 반영 및 처리

프로파일 처리를 통한 추천을 위해서는 개인의 피드백을 반영하며 다수의 프로파일을 취합해 하나의 팀으로서 프로파일을 처리 및 사용자간의 프로파일 처리가 필요하다. 이를 위해 프로파일의 다차원 속성의 값에 대한 거리 계산을 위해 유클리디안 거리 계산법 사용하여 평균 및 표준편차를 다음과 같이 계산한다.

$$TD_{ij} = \sqrt{\sum_{k=0}^{n} FW_n \times ((FA_{ik} - FA_{jk})^2 + (FSD_{ik} - FSD_{jk})^2)}$$

TD는 팀 i와 j간의 거리이며, k는 속성에 대한 인덱스이다. 이때, FA의 i와 j의 k번째 속성의 평균값이며, FSD는 팀 i와 j의 k번째 속성의 표준편차이다. 이를 통해 사용자 간의 분포도까지 고려한 신뢰도 높은 유사도를 추출하여 사용한다.

또한, 보다 신뢰성 높은 프로파일 처리 값을 추출하기 위해서 유사도와 평점을 이용한 팀 평균값과 표준편차 값을 이용한다. 평균값에 표준편차를 이용해 가중치를 부여하면, 분포도를 고려한 값을 추출할 수 있게 되어 비교적 신뢰도 높은 값을 추출할 수 있다.

이를 위해 본 논문에서는 그림 2와 같이 프로파일 처리 알고리즘 구조를 제안하고, 사용자가 선택할 수 있는 추천 분류에 따라 크게 TR(평점 기반 매칭 팀 추천)과 TS(유사도 기반 매칭 팀 추천)으로 나뉜다.

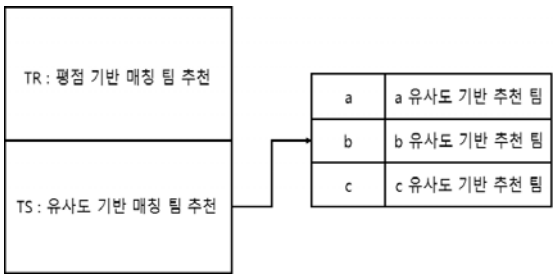

그림 2. 프로파일 처리 알고리즘 구조도

두 추천 분류 중 사용자가 어떠한 추천 분류를 선택하느냐에 따라 TR, TS 사이의 가중치 값의 변화를 통해 사용자에게 더욱 신뢰도 있는 추천을 제공한다. TR과 TS 사이의 상관 관계식은 다음과 같다.

$$TP_{ij} = \alpha_i \times TS_{ij} + (1 - \alpha_i) \times TR_{ij}$$

$$\alpha_t = \frac{\sum_{k=1}^{n} C_n}{MC_t}$$

a, b, c의 초기 값은 1을 가지며, MC(경기 수)

는 3의 값을 가진다. *FW*는 *a, b, c*와 같은 속성값들이 가지는 가중치이다. 따라서 FW_a, FW_b, FW_c의 값들은 $\frac{a,b,c}{MC}$로 $\frac{1}{3}$이다.

$$FW_n = \frac{C_n}{\sum_{k=1}^{n} C_n}$$

여기에서 a, b, c는 위치, 승률 등과 같은 하나의 속성을 기반으로 추천한 팀들의 리스트이며, a, b, c와 같은 속성들의 후보군은 다음과 같다.

1). 위치

대다수 스포츠 동호회의 특징은 위치(연고지)를 기반으로 한다는 것이다. 주로 학교 및 공공시설을 기준으로 하여 스포츠 활동이 이루어진다. 때문에 유사한 연고지의 팀들을 매칭 시켜주기 위해 위치를 속성 값으로 이용해 추천한다.

2). 경기 수 5판 이하 승률

2)와 3)을 나눈 이유는 5판 이하는 경기의 표본수가 매우 적어 승률의 신뢰도가 떨어진다. 예를 들어 1경기 1승의 팀은 승률이 100%인 경우와 같다. 이처럼 경기수가 적은 팀들은 따로 묶어서 경기를 치르게 하여 일종의 배치고사를 보게 하여 표본을 쌓아 승률의 신뢰도를 구축한다.

3). 경기 수 6판 이상 승률

경기수가 6판 이상인 팀은 경기의 표본이 쌓여 승률의 신뢰도가 확립됐다고 볼 수 있다. 그렇기 때문에 경기 수가 6판 이상인 팀은 따로 묶어서 경기를 치를 수 있도록 한다.

4). 평균 나이

현재 대부분의 스포츠 동호회는 나이 대를 기반으로 형성된 경우가 많다. 직장인 동호회, 대학생 동호회 등이 대표적인 예시이다. 이러한 기존의 방식을 유지하기 위해 평균 나이 대를 기반으로 하는 추천 리스트를 제시한다.

5). 경기 요일이 유사한 팀

대부분의 스포츠 동호회는 경기를 위한 요일을 정해놓고 주기적으로 매칭을 하는 경우가 많다. 그렇기 때문에 두 팀의 주요 경기 요일이 유사하여야 두 팀의 매치가 성사될 수 있다. 이를 근거로 경기 요일에 기반 한 추천리스트를 제시한다.

마지막으로, 팀의 평점을 계산할 때, 개인이 각 팀에 끼치는 기여도를 고려하여 주로 활동하는 팀원들에게 더 높은 평균 가중치를 부여하여 현재 팀에서 활동하는 팀원들의 평균을 도출하여 보다 높은 신뢰도의 평균값을 도출한다. 수식은 다음과 같다.

$$MW_u = \frac{(RC_u + U_\alpha)}{\sum_{k=1}^{u}(RC_u + U_\alpha)}$$

여기에서, RC_u이 사용자 u의 매칭에 대한 평점 입력 수일 때, U는 사용자의 직위(즉, 리더, 관리자 등)에 대한 가중치를 나타낸다. 이에 대한 정규화 값을 각 사용자의 가중치로 활용한다.

Ⅳ. 결 론

본 논문에서는 본 논문에서는 맞춤형 생활 스포츠 콘텐츠를 제공에 필요한 동적 프로파일을 처리하는 알고리즘을 제안하였다.

제안한 알고리즘은 사용자간의 프로파일 처리를 하고, 사용자가 선택한 추천 분류에 따라 TR, TS 사이의 가중치 값의 변화를 통해 사용자에게 더욱 신뢰도 있는 추천을 제공하는 프로파일 처리 알고리즘을 제안하였다. 제안한 프로파일 처리 알고리즘을 통해 맞춤형 생활 스포츠 콘텐츠를 추천의 신뢰성과 만족도를 높여줄 것으로 기대하며, 향후 제안한 프로파일 처리 알고리즘을 통해 실제 스포츠 매칭을 필요로 하는 생활 스포츠 팀의 만족도 조사와 서비스의 신뢰성 및 만족도 향상을 위한 연구를 진행할 예정이다.

Acknowledgement

위 논문은 문화체육관광부의 스포츠산업기술개발사업에 의건 국민체육진흥공단의 국민체육진흥기금을 지원받아 연구되었습니다.

참고문헌

[1] Lee Sung Jik, "Keyword Profile-based Personalized Web Search", University of Seoul Postgraduate School, Electronic Electricy Computer Engineering Dept, pp.10-13, 2010

[2] Lee Sung Jik, Kim Han jun, "Keyword Extraction from News Corpus usin Modified Tf-IDF", Korea Institute for Electronic Commerce, Vol. 14, No. 4, pp.61-65, 2009