웹 캐스트와 텍스트 마이닝을 이용한 축구 경기의 심층 분석

In-depth Analysis of Soccer Game via Webcast and Text Mining

저자 (Authors)

정호석, 이종욱, 유재학, 이한성, 박대희

Hoseok Jung, Jonguk Lee, Jaehak Yu, Hansung Lee, Daihee Park

출처 (Source)

한국콘텐츠학회논문지 11(10), 2011.10, 59-68 (10 pages)

JOURNAL OF THE KOREA CONTENTS ASSOCIATION 11(10), 2011.10, 59-68 (10 pages)

발행처 (Publisher)

한국콘텐츠학회

The Korea Contents Society

URL

http://www.dbpia.co.kr/Article/NODE01713822

APA Style

정호석, 이종욱, 유재학, 이한성, 박대희 (2011). 웹 캐스트와 텍스트 마이닝을 이용한 축구 경기의 심층 분석. 한국콘텐츠학회논문지, 11(10), 59-68.

http://dx.doi.org/10.5392/JKCA.2011.11.10.059

웹 캐스트와 텍스트 마이닝을 이용한 축구 경기의 심층 분석

In-depth Analysis of Soccer Game via Webcast and Text Mining

정호석*, 이종욱**, 유재학***, 이한성****, 박대희*****
한국생명공학연구원 지식정보과*, 고려대학교 전산학과**, 한국전자통신연구원 USN기반기술연구팀***,
한국전자통신연구원 휴먼인식기술연구팀****, 고려대학교 컴퓨터정보학과*****

Hoseok Jung(hsjeong@kribb.re.kr)*, Jonguk Lee(eastwest9@korea.ac.kr)**,
Jaehak Yu(dbzzang@etri.re.kr)***, Hansung Lee(mohan@etri.re.kr)****,
Daihee Park(dhpark@korea.ac.kr)*****

요약

축구 경기를 분석하고 이를 팀 전략 수립에 활용하는 축구 분석관의 역할이 강조됨에 따라, 방송용 축구 경기에서 주요 이벤트의 탐지와 같은 절차적 기능 이상의 고수준의 해석 방법들이 요구되고 있다. 본 논문에서는 인터넷 기반의 텍스트 방송인 축구 웹 캐스트에서 실시간으로 제공하는 텍스트 정보를 기반으로 텍스트 마이닝을 이용한 축구 경기의 전략 수립이 가능한 고수준의 해석 기법을 제안한다. 제안하는 해석 기법은 축구 웹 캐스트의 텍스트 정보와 도메인 지식을 기반으로 축구 경기의 다양한 속성, 동작 그리고 이벤트 등 메타데이터를 추출하고, 인덱싱하고, 텍스트 마이닝의 다양한 해석 기법인 연관 규칙 마이닝, 성장도 분석, 그리고 패스파인더 네트워크 분석 기법 등을 사용함으로써 유용한 지식을 추출한다. 실제 2010년 월드컵의 스페인 팀 경기들을 중계한 웹 캐스트의 텍스트 정보를 대상으로 제안된 기법의 타당성을 실험적으로 검증한다.

■ 중심어 : | 텍스트 마이닝 | 웹 캐스트 | 축구 경기 분석 |

Abstract

As the role of soccer game analyst who analyzes soccer games and creates soccer wining strategies is emphasized, it is required to have high-level analysis beyond the procedural ones such as main event detection in the context of IT based broadcasting soccer game research community. In this paper, we propose a novel approach to generate the high-level in-depth analysis results via real-time text based soccer Webcast and text mining. Proposed method creates a metadata such as attribute, action and event, build index, and then generate available knowledges via text mining techniques such as association rule mining, event growth index, and pathfinder network analysis using Webcast and domain knowledges. We carried out a feasibility experiment on the proposed technique with the Webcast text about Spain team's 2010 World Cup games.

■ keyword : | Text Mining | Webcast | Soccer Game Analysis |

접수번호 : #110706-007
접수일자 : 2011년 07월 06일
심사완료일 : 2011년 09월 16일
교신저자 : 박대희, e-mail : dhpark@korea.ac.kr

I. 서 론

현재 방송과 인터넷에서 제공하는 스포츠 비디오의 양이 폭발적으로 증가함에 따라, 대용량의 데이터베이스 및 멀티미디어 관련 컴퓨터 기술을 이용한 스포츠 비디오의 효과적인 저장 및 검색, 그리고 활용방안 등이 최근 관련 학계와 산업체의 주요 관심 대상이다. 특히, 축구 경기를 분석하고 이를 팀 전략 수립에 활용하는 축구 분석관의 역할이 강조됨에 따라, 방송용 축구 경기에서 주요 이벤트의 탐지와 같은 절차적 기능에서부터 고수준의 축구 비디오 해석 방법에 이르는 다양한 기능들이 요구된다[1]. 그러나 현재 방송과 인터넷에서는 축구 비디오의 하이라이트 검색과 같은 단순한 기능 및 기초적인 통계 정보만이 제공되고 있다. 결국 축구 경기의 승패에 직·간접적으로 영향을 미치는 다양한 변인들의 추출과 이들 변인 간의 과학적 분석을 통한 축구 경기의 전략 수립은 사실상 불가능하다.

축구 경기 분석과 관련된 대부분의 컴퓨터 연구들은 비디오 영상을 기반으로 진행되고 있다[2-5]. 그러나 사람이 비디오 영상을 인지하는 방식과 저차원 특징 값을 기반으로 한 컴퓨터 표현의 차이로 인하여 사람과 컴퓨터 사이의 의미적 차이(semantic gap)가 존재한다[6]. 결국, 이러한 의미적 차이로 인하여 영상에서 고차원의 해석 결과를 유추한다는 것은 매우 어려운 문제로 평가되고 있다[6]. 반면, 비디오 영상과 함께 텍스트 정보를 이용하는 혹은 비디오 영상과는 별개로 텍스트 정보만을 이용함으로써 의미적 차이의 문제를 간접적으로 해결하고자 하는 시도들[7][8]이 발견된다. 이 중 축구 경기를 인터넷을 통해 텍스트로 실시간 중개해주는 인터넷 기반의 텍스트 방송인 축구 웹 캐스트를 활용한 방법[7][8]이 흥미롭다. 축구 웹 캐스트에서 제공하는 텍스트 정보를 살펴보면, 각 문장은 하나의 동작을 나타내고, 슛, 파울, 프리킥 등과 같은 이벤트와 이벤트가 발생한 시간, 선수 정보, 선수가 행한 액션 등과 같은 다양한 정보들이 담겨 있다. 웹 캐스트를 이용한 기존의 연구들[7][8]은 현재 웹 캐스트에서 제공하는 텍스트 정보를 이용하여 축구 경기의 이벤트들을 탐지하고, 탐지된 이벤트의 발생 시간 정보를 통하여 그에 해당하는 축구 이벤트 영상을 탐색하는 등과 같은 절차적 기능에 초점이 맞추어져 있다. 결국, 축구 경기에 내재되어 있는 풍부한 잠재적 지식을 탐사하는 고수준의 의미 분석 단계에 까지는 이르지 못하고 있다.

전술한 바와 같이, 축구 경기의 승패에 직·간접적으로 영향을 미치는 다양한 변인들의 추출과 이들 변인 간의 과학적 분석을 통한 전략 수립이 축구 비디오 심층 분석의 최종 목적이라고 한다면, 다루기 어려운 동영상을 직접 대상으로 하는 분석방법보다는 보다 정형화되어 있는 자료구조인 텍스트를 대상으로 이미 상당한 수준에까지 올라와 있는 텍스트 마이닝 기법에 대한 연구로 방향을 설정하는 것이 현 시점에서 취할 수 있는 타당한 방법 중 하나이다. 텍스트 마이닝이란 자연어로 구성된 비구조적인 텍스트 안에서 패턴 또는 관계를 추출하여 지식을 발견하는 과정으로, 주로 텍스트의 자동 분류작업이나 텍스트로부터 새로운 지식을 생성하는 작업에 활용된다[9].

본 논문에서는 인터넷 기반의 텍스트 방송인 축구 웹 캐스트에서 실시간으로 제공하는 텍스트 정보를 기반으로 키워드 매칭을 통하여 축구 경기의 다양한 속성들을 추출하고, 텍스트 마이닝의 다양한 해석 기법인 연관 규칙 마이닝, 성장도 분석, 그리고 패스파인더 네트워크 분석 기법 등을 사용함으로써 축구 경기의 전략 수립이 가능한 고수준의 해석 기법을 제안하고자 한다. 심층 분석을 위한 과정은 다음의 3단계로 진행된다: 1) 웹 캐스트 텍스트의 매 문장들을 키워드 매칭을 통하여 해당 속성들을 추출하고, 메타데이터를 구성한다; 2) correlation based feature selection으로 주요 속성 선택 및 축소 과정을 수행 한다; 3) 선택된 주요 속성들로부터 텍스트 마이닝의 대표적 분석 기법인 연관 규칙 마이닝, 성장도 분석, 그리고 패스파인더 네트워크 분석 기법 등을 적용하여 축구 경기에 내재되어 있는 주요 속성들 사이의 규칙을 찾아내는 심층 분석을 실시한다. 실제 2010년 월드컵의 스페인 경기를 중계한 웹 캐스트의 텍스트 정보를 대상으로 제안한 방법론의 타당성을 실험적으로 검증한다.

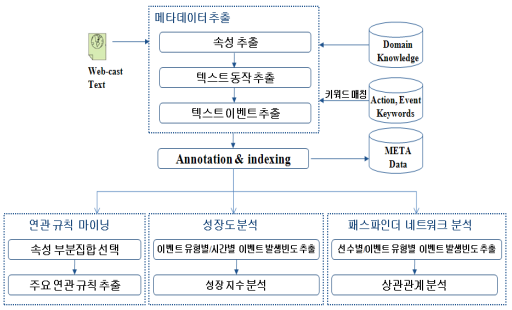


그림 1. 축구 경기 분석 시스템 구조도

본 논문에서는 축구 경기에 잠재된 유용한 지식을 탐사하기 위하여 저 처리비용의 인간친화적인 정보인 웹 캐스트를 적극적으로 활용하고자 한다. 즉, 웹 캐스트 텍스트와 축구 도메인 지식을 활용하여 속성, 동작과 이벤트 등 메타데이터를 추출하고 인덱싱하며, 메타데이터를 대상으로 연관 규칙 마이닝, 성장도 분석, 그리고 패스파인더 네트워크 분석 등으로 텍스트 마이닝을 통하여 축구 경기에 내재된 잠재된 유용한 지식을 추출한다. 추출된 지식은 코칭스태프가 원하는 고수준의 의미적 분석에 효과적으로 사용될 수 있다. 축구 경기 분석 시스템의 전체 구조도는 [그림 1]과 같다.

본 논문의 구성은 다음과 같다. 2장에서는 텍스트 마이닝의 전처리 작업 중 하나인 속성 부분집합의 선택 방법과 텍스트 마이닝의 다양한 해석 기법인 연관 규칙 마이닝, 성장도 분석, 그리고 패스파인더 네트워크 분석 기법 등을 소개한다. 3장에서는 축구 비디오의 심층 분석을 위한 실험 및 분석 결과를 기술하며, 마지막으로 4장에서는 결론 및 향후 연구과제에 대해 논한다.

II. 텍스트 마이닝 기법

2.1 속성 부분 집합의 선택

본 논문에서는 웹 캐스트에서 수집한 다양한 축구 관련 속성들 중에서 그 성능이 이미 검증된 Hall[10]의 correlation based feature selection 방법을 사용하여 속성 부분집합을 선택한다. 이는 최적 우선 탐색(best first search) 방법과 속성 또는 특징 값 Y에 대한 엔트로피(entropy) 그리고 목표 클래스(target class)와 속성들 간의 피어슨 상관 계수를 이용한 조건부 확률을 계산하여 전체 속성들의 확률 분포도를 가능한 가깝게 표현할 수 있는 최소 개수의 속성집합을 찾는 방법이다. 먼저 각 속성들에 대한 정보 이익(information gain)을 얻기 위해 임의의 Y속성에 대한 엔트로피를 식(1)로 계산한다.

$$H(Y) = -\sum_{y \in Y} p(y)\log_2(p(y)) \quad (1)$$

속성 X와 Y사이의 관계는 X가 주어졌을 때 Y가 발생하는 조건부 확률로써 식(2)와 같이 계산된다.

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x)\log_2(p(y|x)) \quad (2)$$

각 특징에 대한 정보 이익은 식(1)과 식(2)를 이용하여 식(3)으로 정의된다.

$$Gain = H(Y) + H(X) - H(X, Y). \quad (3)$$

식(3)에서 얻은 정보 이익을 기반으로 식(4)에서와 같이 symmetrical uncertainty를 이용하여 임의의 두 속성 X와 Y의 분포와 상관관계를 계산한다. 이때 속성 X를 기준으로 Y가 높은 분포와 상관관계를 보이면 전체 속성들을 효율적으로 표현할 수 있는 부분집합에 속성 X는 포함되지만 Y는 포함되지 않는다. 마찬가지로 목표 클래스와 속성들 간의 분포와 상관관계를 계산하여 부분집합을 구성한다.

$$Symmetrical\,uncertainty\,coefficient = 2.0 \times \left[\frac{Gain}{H(Y) + H(X)}\right] \quad (4)$$

각각의 부분집합 $F_S \subset F$ 가 전체 속성들을 얼마나 효율적으로 표현하는지를 평가하기 위하여 메리트 함수(merit function)(식(5))를 사용한다. 메리트 함수의 값이 가장 큰 부분집합이 전체 속성들을 최적으로 표현

할 수 있는 부분집합으로 결정된다[10].

$$Merit(F_S) = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (5)$$

2.2 연관 규칙 마이닝

연관 규칙 마이닝은 수많은 데이터 안에 존재하는 각 객체들 간의 의미 있는 연관 규칙을 찾아내는 데이터 마이닝의 여러 기법 중 하나이다. 먼저 항목들의 집합 $I = \{I_1, I_2, ..., I_m\}$와 각각의 트랜잭션 T는 $T \subseteq I$의 관계를 가진 항목들의 집합이 있을 때, 각각의 트랜잭션 T는 고유한 트랜잭션 구분자(transaction identifier)를 갖는다. A를 항목들의 집합이라고 하면, 트랜잭션 T가 필요충분조건으로 $A \subseteq T$를 만족하는 경우에만 트랜잭션 T가 항목 A를 포함한다고 한다. 여기서 $A \subseteq I$, $B \subseteq I$, $A \cap B = \varnothing$ 을 만족하는 경우, 연관 규칙은 $R: A \Rightarrow B$의 형식으로 표현되며, A를 규칙의 조건부(antecedent), B를 결과부(consequent)라 한다. 추출된 연관 규칙의 평가 기준으로는 지지도(support)와 신뢰도(confidence)를 사용한다. 규칙 $A \Rightarrow B$는 트랜잭션 집합 D에서 집합 A와 B를 동시에 포함하는 트랜잭션의 백분율이 s인 경우 지지도 s를 갖는다고 표현한다. 이는 확률 $P(A \cup B)$를 계산함으로써 얻을 수 있다. 집합 A를 포함하는 트랜잭션 중에서 집합 B도 포함하고 있는 트랜잭션의 백분율이 c인 경우, 규칙 $A \Rightarrow B$는 신뢰도 c를 갖는다고 표현한다. 신뢰도는 조건부확률 $P(B|A)$를 계산함으로써 얻을 수 있다[9].

$$\begin{aligned} &support: (A \Rightarrow B) = P(A \cup B) \\ &confidence: (A \Rightarrow B) = P(B|A) \end{aligned} \quad (6)$$

최소 지지도 임계값(minimum support threshold)과 최소 신뢰도 임계값(minimum confidence threshold)을 동시에 만족하는 규칙을 강한(strong) 규칙이라고 한다. 이때 최소 지지도 값 이상을 갖는 항목집합(itemset)을 빈발항목집합(frequent itemset)이라 하고 개의 항목들로 이루어진 빈발항목집합을 k-빈발항목집합이라고 한다. 이진 연관 규칙에 대한 빈발항목집합을 찾는데 유용한 Apriori 알고리즘[9]은 k번째 항목 집합이 $(k+1)$번째 항목집합을 발견하기 위해 사용되는 반복적 접근방법을 사용하는데, 이는 수준별(level-wise) 방법으로 알려져 있다. 먼저, 빈발 1-항목집합을 L_1으로 나타내며, L_1은 2-항목집합인 L_2를 찾는데 사용되고 이것은 다시 L_3를 찾는데 이용된다. 이러한 방법은 더 이상 빈발 k-항목집합이 없을 때까지 진행된다.

2.3 이벤트 성장도 분석

축구 이벤트의 기간별 성장도 분석[11]에서는 전반 경기를 1기로, 후반 경기를 2기로 나누어 식(7)의 이벤트 성장 지수(EGI: Event Growth Index)를 계산한다. 이벤트 성장 지수는 1기와 비교하여 2기에 각각의 이벤트가 성장한 정도를 상대적 비율로 비교하기 위한 지표이다.

$$EGI = \frac{\text{2기 이벤트 수} - \text{1기 이벤트 수}}{\text{2기 이벤트 수} + \text{1기 이벤트 수}} \quad (7)$$

이벤트 성장 지수는 상대적인 이벤트 성장도를 나타내는 지표로서, 절대적 이벤트의 증가 정도를 반영하기 어렵다. 따라서 이벤트 성장 지수에 각 이벤트의 성장 규모를 반영하는 지수로 아래 식(8)과 같은 가중 이벤트 성장 지수(WEGI: Weighted EGI)를 산출할 수 있다.

$$WEGI = |\text{2기 이벤트수} - \text{1기 이벤트수}| \times EGI \quad (8)$$

이벤트 성장 지수가 상대적 비율을 이용한 지표인데 반해, 가중 이벤트 성장 지수는 절대적인 이벤트 증가를 반영하는 지표가 된다. 즉, 가중 이벤트 성장 지수는 증가한 이벤트 수에 비례하여 값이 커진다. 따라서 가중 이벤트 성장 지수가 보다 분별력 있는 지표이며, 본 논문에서는 가중 이벤트 성장 지수를 사용하여 이벤트의 기간별 성장도 분석을 수행한다.

2.4 패스파인더 네트워크 분석

팀 내 선수간의 상관관계 및 연관성을 확인하기 위하여 패스파인더 네트워크 스케일링(pathfinder network scaling)을 이용한 네트워크 분석(network analysis)을 수행한다. 먼저, 다음의 코사인 유사도 함수(식(9))로 선수 사이의 공격 이벤트 유사도 행렬을 생성한다.

$$\cos(d_x,d_y)=\frac{\sum_{i=1}^{n}(w(a_i,d_x)\times w(a_i,d_y))}{\sqrt{\sum_{i=1}^{n}w(a_i,d_x)^2\times\sum_{i=1}^{n}w(a_i,d_y)^2}} \quad (9)$$

여기에서 a_i는 이벤트 유형을 d_x와 d_y는 각 선수를 의미하고, $w(a_i,d_x)$와 $w(a_i,d_y)$는 선수 d_x와 d_y가 이벤트 유형 a_i를 각각 몇 번씩 발생시켰는지를 나타낸다.

패스파인더 네트워크는 가중치가 있는 모든 링크가 생성된 상태에서 삼각 부등식(triangle inequality)을 위반하는 경로를 제거함으로써 생성되는 네트워크이다. 이를 생성하는 알고리즘을 패스파인더 네트워크 알고리즘이라 하며, 때로는 다변량 분석 기법의 일종으로 간주하여 패스파인더 네트워크 스케일링이라 부르기도 한다[12][13]. 삼각 부등식 위반 여부를 결정하기 위해서는 두 가지 파라미터인 q와 r이 필요하다. 파라미터 q는 노드사이의 경로거리를 산출하는데 고려하는 최대 링크의 수를 뜻한다. n개의 노드를 고려할 경우, 파라미터 q는 2에서 n-1까지로 설정된다. 파라미터 r은 민스코프스키 거리 공식의 제곱수로서, 두 노드 n_i와 n_j 사이의 특정 경로를 구성하는 여러 링크들이 가지고 있는 가중치를 거리 $W(n_i,n_j)$에 반영한 것이다.

$$W(n_i,n_j)\le\left(\sum_{l=1}^{k-1}W^r(n_l,n_{l+1})\right)^{1/r},\quad \forall\, k=2,3,\ldots,q \quad (10)$$

식 (10)에서 r이 1이면 각 링크 가중치의 합이 그대로 경로의 거리가 되고, r이 무한대이면 경로를 구성하는 링크의 가중치 중 최대값이 경로의 거리가 된다. r이 커질수록 경로의 길이가 짧아지므로 남는 링크의 수는 줄어든다.

Ⅲ. 실험 및 결과 분석

본 논문에서 제안한 축구 경기의 심층 분석 방법론을 평가하기 위하여, 2010년 남아프리카 공화국 월드컵의 스페인 팀이 예선부터 결승까지 참여한 7 경기의 웹 캐스트를 대상으로 텍스트 마이닝의 다양한 해석 기법인 연관 규칙 마이닝, 성장도 분석, 그리고 패스파인더 네트워크 분석 기법 등을 수행하였다. 실제, 2010년 남아프리카 월드컵에서 스페인과 독일과의 월드컵 경기를 중계하는 웹 캐스트의 예는 [그림 2]와 같다.

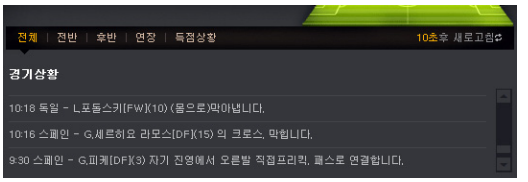


그림 2. 축구 웹 캐스트의 예

실험에서 사용한 컴퓨터 환경은 Pentium4 2.0GHz, Ram 2G 그리고 운영체제는 Window XP이며 java 기반의 기계학습 프로그램인 WEKA[14]를 사용하였다.

표 1. 웹 캐스트에서 추출한 축구 정보 속성들의 예

속성	속성 설명
선수	해당 이벤트에 포함된 선수 이름
위치	공격 위치(좌, 우, 중앙)
이벤트	필드공격, 프리킥, 코너킥, 패널티킥
동작	크로스, 패스, 슛, 공격차단 등
시간	분, 초
전·후반	전반, 후반 여부
슛 거리	단거리슛, 중거리슛, 장거리슛
슛을 한 발	왼발, 오른발
유효슛 여부	유효슛, 무효슛
스코어	현재 상태 스코어

이벤트에 관한 속성 값들은 축구 도메인 지식과 경기와 관련된 사전 정보들을 이용하여 데이터베이스를 구성한 후, 웹 캐스트의 정보들을 키워드 매칭을 통하여 속성 정보들을 수집하였다. 추출한 축구 정보 속성과 메타데이터의 예는 각각 [표 1][표 2]와 같다.

표 2. 추출된 메타데이터의 예

시간	이벤트	동작	위치	선수	설명
50:05	필드 공격	도움	1선 Right	나바스	'나바스' 오른쪽 측면에서 패스로 도움
50:06	필드 공격	슛	2선 Middle	비야	'비야' 오른쪽 중앙에서 오른발 슛
50:06	필드 공격	골인	2선 Middle	비야	'비야' 골인, 오른쪽 중앙에서 골대 정면

첫 번째 실험은 이벤트를 발생시키는 속성들의 패턴을 발견하는 실험이다. WEKA의 CFS(Correlation Feature Selection)를 이용하여 3개의 최적 속성 부분집합 {전·후반, 선수, 위치}를 찾았으며 이를 연관 규칙의 조건부로, 결과 요인은 이벤트로 설정하였고 최소 지지도는 3%, 최소 신뢰도는 70%로 설정하였다. 실험결과 추출된 속성들 간의 주요 연관 규칙들을 [표 3]에 정리하였다.

표 3. 이벤트 발생 요인 분석을 위한 연관 규칙들

번호	규칙 내용	신뢰도
1	'토레스' 이고 중앙 공격이면(9), 필드공격이다(8)	89%
2	'사비' 이고 중앙 공격이면(23), 프리킥이다(20)	87%
3	'사비' 이고 우측공격이면(33), 코너킥이다(25)	76%
4	후반이고 '비야' 이고 중앙공격이면(15), 필드공격이다(11)	73%
5	후반이고 '알론소' 이고 중앙공격이면(17), 프리킥이다(12)	73%

선수 및 공격 위치 속성과 관련한 [표 3]의 규칙 1~3으로부터, '토레스'가 중앙에서 필드공격을 많이 시도하고, '사비'가 중앙에서 프리킥과 우측에서 코너킥을 많이 시도하는 것이 스페인 공격의 전형적 패턴임을 알 수 있었다. 이러한 패턴은 특정 선수가 특정한 위치에서 선호하는 공격 스타일이 분명히 존재한다는 것을 보여준다. 규칙 4, 5의 경우에서는 전반에 비해 후반에 '비야'가 중앙에서 필드공격을 시도한 경우와 '알론소'가 중앙에서 프리킥 공격을 시도하는 공격 패턴이 빈번히 발생함을 확인할 수 있었다. 또한 후반전에 '비야'와 '알론소'가 중앙 공격을 많이 시도하였고, 필드공격은 '비야'가 프리킥은 '알론소'가 주도한다는 중요한 전략적 사실을 파악할 수 있었다.

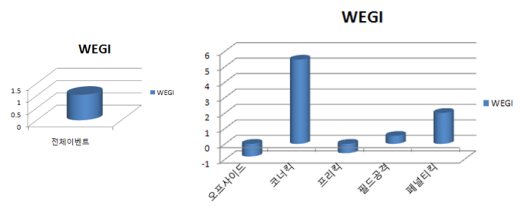
두 번째 실험은 슛 데이터에 내재되어 있는 유효슛 발생 요인 패턴을 분석하기 위한 실험이다. WEKA에서 CFS로 선택된 4개의 최적 속성 부분집합은 {전·후반, 이벤트 유형, 선수, 위치}이며 이를 연관 규칙의 조건부로, 유효슛을 결과부로 지정하였다. 최소 지지도는 5%, 최소 신뢰도는 65%로 설정하였다. 실험결과 추출된 속성들 간의 주요 연관 규칙들을 [표 4]에 정리하였다.

표 4. 유효슛 발생 요인 분석을 위한 연관 규칙들

번호	규칙 내용	신뢰도
1	필드공격이고 좌측 공격이면(12), 유효슛이다(10)	83%
2	좌측 공격이면(15), 유효슛이다(11)	73%
3	필드공격이고 '비야' 이면(23), 유효슛이다(16)	70%
4	필드공격이고 '비야' 그리고 중앙공격이면(16), 유효슛이다(11)	69%
5	'나바스' 이면(6), 유효슛이다(4)	77%

[표 4]의 규칙들을 살펴보면, 규칙 1, 2을 통하여 스페인 팀은 좌측 필드공격 또는 좌측 공격에서 유효슛을 많이 발생시키고 있음을 알 수 있다. 이러한 패턴 분석은 특정 위치와 특정 이벤트에서 해당 팀의 유효슛 성공률이 높다는 사실을 보여주고 있는 것으로, 이는 추후 이기기 위한 경기의 전략 수립에 소중하게 사용될 수 있을 것이다. '비야' 선수와 관련된 규칙 3, 4의 경우, '비야'가 다른 선수들에 비해 상당히 많은 필드공격을 통해 슛함을 알 수 있으며, 그 중 중앙 필드공격이 대부분임을 알 수 있다. 이러한 패턴은 특정 선수가 선호하는 공격 방향과 슛을 주도하는 선수가 누구인지를 미리 가늠할 수 있는 중요한 단서가 된다. 규칙 3~5를 통해 '비야'가 스페인의 슛을 전반적으로 주도하고 있다는 사실도 확인할 수 있다. 규칙 5의 '나바스'는 공격수가 아닌 미드필더 임에도 불구하고 공격의 유효성이 높음을 알 수 있다.

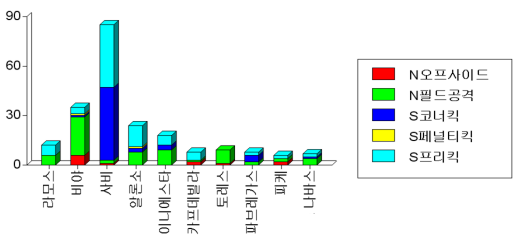
세 번째 실험은 경기 전반과 후반을 기준으로 전반적인 이벤트의 변화를 확인하기 위하여 전후반 별로 이벤트 유형별(오프사이드, 필드공격, 코너킥, 페널티킥, 프리킥) 공격 빈도를 산출하고 성장 지수 분석 실험을 실시하였다. [그림 3]은 경기 전반을 1기로, 경기 후반을 2기로 나누어 가중 이벤트 성장 지수를 나타낸 그림이다. 위 [그림 3](a)는 전체 이벤트 발생 수에 대한 기간별 분석을 나타낸 것이며, 1기에 비해 2기에 이벤트 수가 크게 증가된 것으로 나타났다. 전체 이벤트를 각 이벤트별 성장지수로 본 [그림 3](b)의 내용을 확인해보면 오프사이드는 1기에 비해 2기에 감소한 반면, 코너킥의 경우에는 1기에 비하여 2기에 두드러지게 증가한 것을 확인할 수 있으며 페널티킥의 수도 증가하였다. 결국 1기인 전반에 비해 2기인 후반에 축구 이벤트가 상대적으로 많이 발생하고 있음을 확인할 수 있다.



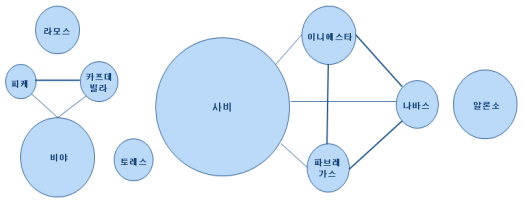
(a) (b)

그림 3. (a) 전체 이벤트에서의 이벤트 성장지수
(b) 각 이벤트별 성장지수

네 번째 실험은 팀 내 선수들 사이의 이벤트와 관련된 상관관계를 알아보기 위하여 각 선수들의 이벤트 유형별 공격 빈도를 산출하고 패스파인더 네트워크 분석으로 NWB(Network WorkBench) Tool[15]을 사용하여 실험하였다.



(a)



(b)

그림 4. (a) 선수별 이벤트 현황
(b) 선수별 유사도 행렬 기반 패스파인더 네트워크 분석

[그림 4](a)는 선수별 이벤트 유형별 공격 빈도를 도출한 결과이며, [그림 4](b)는 선수별 유사도 행렬을 기반으로 패스파인더 네트워크 분석을 수행한 결과이다. [그림 4](b)에서 노드의 크기는 선수가 공격한 이벤트의 수를 의미하고, 각 노드를 연결하는 링크의 굵기는 각 노드 간의 관련 정도를 나타낸다. 본 실험에서는 가중치 0.7이하인 링크는 제거하였으며, 0.9 이상은 굵은 선으로, 0.9 미만은 가는 선으로 표시하였다.

위 실험분석을 통하여, 각 선수 간의 흥미로운 상관관계 몇 가지를 다음과 같이 확인할 수 있었다:

첫째, 공격수들의 공격 패턴에 대한 상관관계 분석으로, '비야'와 '토레스'는 모두 최전방 공격수들이지만 서로 다른 군집(특성)을 보인다[그림 4](b). 이들 선수들의 공격 이벤트를 [그림 4](a)에서 보다 자세히 검토한 결과, 좌측 공격수인 '비야'는 필드공격을 주로 하며 동시에 프리킥과 코너킥 등 세트피스도 시도하는 패턴을 보이는 반면, 우측 공격수인 '토레스'는 필드공격만을 시도하고 있음을 확인하였다. 따라서 두 선수가 서로 다른 공격 패턴을 보임을 확인하였다.

둘째, 미드필더들의 공격 패턴에 관한 상관관계 분석으로, 미드필더인 '사비', '이니에스타', '파브레가스' 그리고 '나바스'는 하나의 군집을 이루며 서로 유사한 패턴을 보이고 있음을 보여준다. 그러나 '알론소'는 같은 미드필더임에도 불구하고 별도의 군집으로, 전체 공격 중 유난히 프리킥 비중이 높은 특이 공격 패턴을 보임을 확인하였다.

셋째, 수비수들의 공격 패턴에 관한 상관관계 분석으로, 좌측 수비수인 '피케'와 '카프데빌라'는 공통적으로 공격 참여율은 적지만 프리킥 참여율은 상대적으로 높

음을 확인하였다. 또한, 이들은 다양한 공격 패턴을 보임으로 공격수인 비야와도 약한 상관관계를 갖고 있음을 확인하였다. 그러나 '라모스'는 같은 수비수임에도 불구하고 공격에 가담하는 비율이 매우 높은 특이 패턴을 보이며, 다른 수비수들과 달리 세트피스 보다는 필드공격으로 오버래핑(overlapping)을 자주 한다는 사실을 알 수 있었다.

결과적으로 팀 내 선수 간의 위와 같은 상관관계 분석 결과는 코칭스태프가 속한 팀 또는 상대 팀 전력 분석 과정에서 의미 있는 자료로 활용될 수 있을 뿐 아니라, 향후 경기 전략 수립 시 상대 팀의 상관관계가 높은 선수들 별로 공격 패턴을 분석하고 대책을 수립하고 대비함으로써 보다 객관적이고 효율적인 축구 경기 분석을 위한 과학적 근거로 활용될 것으로 기대된다.

Ⅳ. 결론

기존의 축구 영상을 기반으로 한 컴퓨터 분석 방법들은 영상 자체를 이해하기 위한 전처리 과정과 저차원의 특징정보를 고차원의 의미와 연결하는 과정에서, 소위 의미적 차이 문제가 있었다. 이러한 문제점을 해결하면서, 동시에 축구 경기의 과학적 분석을 통한 방송용 축구 경기의 심층 분석을 위하여, 본 논문에서는 축구 경기 웹 캐스트에서 실시간으로 제공하는 텍스트 정보를 텍스트 마이닝의 다양한 해석 기법인 연관 규칙 마이닝, 성장도 분석, 그리고 패스파인더 네트워크 분석 기법 등을 이용하여 축구경기와 관련된 속성들 사이의 내재된 유용한 지식을 발견하는 시도를 하였다. 실제 2010년 월드컵의 스페인 경기를 중계한 웹 캐스트의 텍스트 정보를 대상으로 제안된 방법론의 타당성을 실험적으로 검증하였다.

본 논문에서 제안한 웹 캐스트와 다양한 텍스트 마이닝 기법들을 이용한 축구 경기의 심층 분석 기법은 아직까지는 그 활용 가능성만을 확인하는 차원의 프로토타입이다. 따라서 계속적인 추가 연구가 요구되며, 특히 비디오 정보와 텍스트 정보를 동시에 활용하는 멀티모달(multi-modality) 연구 방향이 바람직해 보인다.

참 고 문 헌

[1] C. Xu, J. Wang, H. Lu, and Y. Zhang, "A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video," IEEE Transactions on Multimedia, Vol.10, No.3, pp.421-436, 2008.

[2] J. Assfalg, M. Bertini, C. Colombo, A. Bimbo, and W. Nunziati, "Semantic Annotation of Soccer Videos: Automatic Highlights Identification," Computer Vision and Image Understanding, Vol.92, No.2-3, pp.285-305, 2003.

[3] X. Qian, G. Liu, H. Wang, Z. Li, and Z. Wang, "Soccer Video Event Detection by Fusing Middle Level Visual Semantics of an Event Clip," Advances in Multimedia Information Processing, Vol.6298, pp.439-451, 2010.

[4] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang, and H. Wang, "Automatic Player Detection, Labeling and Tracking in Broadcast Soccer Video," Pattern Recognition Letters, Vol.30, No.2, pp.103-113, 2009.

[5] M. Chen, S. Chen, and M. Shyu, "Hierarchical Temporal Association Mining for Video Event Detection in Video Databases," IEEE International Conference on Data Engineering Workshop, pp.137-145, 2007.

[6] E. Pazouki and M. Rahmati, "A Novel Multimedia Data Mining Framework for Information Extraction of a Soccer Video Stream," Intelligent Data Analysis, Vol.13, No.5, pp.833-857, 2009.

[7] C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, "Live Sports Event Detection Based on Broadcast Video and Webcasting Text," Proceeding of ACM International Conference Multimedia, pp.221-230, 2006.

[8] C. Xu, Y. Zhang, G. Zhu, Y. Rui, H. Lu, and Q.

Huang, "Using Webcast Text for Semantic Event Detection in Broadcast Sports Video," IEEE Transaction on Multimedia, Vol.10, No.7, pp.1342-1355, 2008.

[9] J. Han and M. Kamber, *Data Mining: Concepts and Techniques, 2nd edition,* Morgan Kaufmann Publishers, 2007.

[10] M. Hall, *Correlation-based Feature Selection for Machine Learning*, Ph.D. Diss. Department of Computer Science, Waikato University, Hamilton, NZ, 1998.

[11] 이재윤, 문주영, 김희정, "텍스트 마이닝을 이용한 국내 기록관리학 분야 지적구조 분석", 한국문헌정보 학회지, 제41권, 제1호, pp.345-372, 2007.

[12] R. W. Schvaneveldt, *Pathfinder Associative Networks: Studies in Knowledge Organization*, Norwood, NJ:Ablex, 1990.

[13] C. Chen, "Generalized Similarity Analysis and Pathfinder Network Scaling," Interacting with Computers, Vol.10, No.2, pp.107-128, 1998.

[14] http://www.cs.waikato.ac.nz/ml

[15] http://nwb.slis.indiana.edu

저 자 소 개

정 호 석(Hoseok Jung) 정회원

- 1991년 : 전남대학교 수학과 학사
- 1992년 ~ 1994년 : (주)한국전자계산 근무
- 1994년 ~ 현재 : 한국생명공학연구원 책임기술원
- 2000년 : 공주대학교 전산학과 석사
- 2011년 : 고려대학교 전산학과 박사

<관심분야> : 기계학습, 멀티미디어 마이닝, 얼굴 탐지 및 인식

이 종 욱(Jonguk Lee) 정회원

- 2003년 : 고려대학교 전산학과 학사
- 2005년 : 고려대학교 전산학과 석사
- 2005년 ~ 현재 : 고려대학교 전산학과 박사과정

<관심분야> : 멀티미디어 마이닝, 얼굴 탐지 및 인식, 기계학습, 지능 데이터베이스

유 재 학(Jaehak Yu) 정회원

- 2001년 : 건국대학교 전산과학과 학사
- 2003년 : 고려대학교 전산학과 석사
- 2010년 2월 : 고려대학교 전산학과 박사
- 2006년 3월 ~ 2008년 2월 : 고려대학교 컴퓨터정보학과 초빙전임강사
- 2010 ~ 현재 : 한국전자통신연구원

<관심분야> : 내용기반 영상검색, 멀티미디어

이 한 성(Hansung Lee) 정회원

- 1996년 : 고려대학교 전산학과 학사
- 1996년 ~ 1999년 : (주) 대우엔지니어링 근무
- 2002년 : 고려대학교 전산학과 석사
- 2008년 : 고려대학교 전산학과 박사
- 2006년 ~ 2007년 : 고려대학교 컴퓨터정보학과 초빙전임강사
- 2008년 ~ 2009년 : 고려대학교 BK21 연구교수
- 2009년 ~ 현재 : 한국전자통신연구원

<관심분야> : 기계 학습, 얼굴 인식, 멀티미디어 마이닝

박 대 희(Daihee Park) 정회원

- 1982년 : 고려대학교 수학과 학사
- 1984년 : 고려대학교 수학과 석사
- 1989년 : 플로리다 주립대학 전산학과 석사
- 1992년 : 플로리다 주립대학 전산학과 박사
- 1993년 ~ 현재 : 고려대학교 컴퓨터정보학과 교수

<관심분야> : 지능 데이터베이스, 데이터마이닝, 인공지능, 퍼지이론