그래프 중심성을 이용한 프로야구 정규시즌 분석

김창민, 우하은, 한치근
경희대학교 컴퓨터공학과
kcmin19@khu.ac.kr suaveh97@khu.ac.kr

Analysis of Korean professional baseball regular season using graph centrality

Changmin Kim, Haeun Woo, Chigeun Han
Department of Computer Science and Engineering, KyungHee University

요 약

본 연구에서는 한국 프로야구 정규시즌에 진행되는 720 경기 중 무승부 12경기를 제외한 708 경기를 그래프 중심성을 이용하여 분석하고자 한다. 타자지표, 투수지표, 수비지표 등의 야구 지표들을 Betweenness, pagerank centrality를 이용하여 중심성을 측정해내어 어떤 지표가 경기의 승리에 가장 중요한지 알아본다.

1. 서 론

프로야구는 국내 프로 스포츠 중 가장 사랑받는 종목이다. 대중에게 받는 관심만큼 많은 학자들이 분석하기를 사랑하는 종목이기도 하다. 야구는 어느 스포츠보다 분석할 지표가 넘치기 때문이다. 공격 시간과 수비 시간이 따로 구분되어 있지 않은 축구 등의 여타 스포츠와는 달리, 야구는 공격 이닝과 수비 이닝이 정확히 나누어져 있다. 게다가 공격 이닝에는 야수만 나서지만 수비 이닝에는 투수까지 참여하기 때문에 야수 지표와 투수 지표가 별도로 존재한다. 본 연구에서는 넘쳐나는 야구 지표들 가운데 어떤 지표가 경기에 가장 결정적인 요인인지 그래프 중심성을 이용하여 분석하고자 한다.

2. 관련 연구

2.1. Graph Centrality

네트워크에서 중심성은 하나의 정점이 전체 네트워크에서 어느 정도의 영향력을 가지고 있는지, 즉, 네트워크의 중심에 위치하는 정도를 뜻한다. 사회 연결망을 분석하는데 중심성을 사용하는 경우에는 네트워크에서 연결 정도, 정점의 영향력, 정점 간에 관계가 닿아있는 정도 등, 무엇을 중요하게 볼 것인지에 대한 기준을 정해야 한다.

그래프 중심성 중 Betweenness Centrality는 한 정점이 서로 다른 두 정점의 최단거리에 속하는 비중을 나타낸 것이다. 즉, 해당 정점이 다른 정점들의 bridge 역할을 하는 정도를 나타낸다. Eigenvector Centrality는 그래프에서 영향력 있는 정점들과 연결되어 있는 정도를 나타낸다[1]. PageRank Centrality는 고유벡터 중심성에서의 영향력에 더해 정점 사이의 간선의 방향을 추가로 고려하여 중심성을 구하는 방법이다. 들어오는 방향의 연결을 가진 정점의 영향력이 커져, 웹 검색엔진에 사용되는 중심성이다[2].

2.2. 다중회귀분석을 이용한 메이저리그 승률의 모형구축과 예측

해당 연구에서는 메이저리그의 경기로부터 79개의 변인을 조사하여, 승률을 반인 변수로, 나머지 78개의 변인을 설명변수로 가지는 다중회귀분석을 수행한다[3]. 전처리 시 반인 변수와 상관관계가 없는 것으로 나타나는 변수와 설명변수 간에 다중공선성이 높게 나타나는 변수들을 제거하고 20개의 설명변수를 사용하여 회귀모형을 구성하였다. 그 결과 득점, 홈런, 삼진, 실점, 세이브, 완봉경기 수의 6가지 요인이 승률에 영향을 미치고 있다는 결론을 도출했다. 본 연구에서는 국내 프로야구 정규시즌 경기에서 기록되는 변인들을 조사하여 각 변인이 경기, 경기의 승패 유무와 관계를 가지는 정도를 네트워크로 나타내고 그래프 중심성을 이용하여 각 변인이 경기의 승패에 작용하는 중요도를 파악한다.

3. 연구 내용 및 연구 방법

3.1. 연구 내용 요약

본 연구에서는 한국 프로야구 정규시즌의 경기, 경기에 작용할 수 있는 변인, 경기의 승리 여부를 정점으로 하고, 경기로부터 승리까지 도달하는 네트워크를 구성한다.

해당 네트워크에서는 승리 팀의 기록에서 패배 팀의 기록 간의 차를 구한 후 정규화하여 변인과 팀 사이의 가중치를 설정하고, 양수인 경우 승리에 기여한 변인으로서 승리 여부 정점에 간선을 연결한다. 각 요인이 경기가 승리로 연결되는 데에 bridge 역할을 하는 정도를 나타내는 Betweenness Centrality와 중요한 정점(경기의 승리 여부)과의 연결 정도를 방향을 포함하여 나타내는 PageRank Centrality를 사용하여 각 변수가 해당 네트워크에서 가지는 중요도를 측정하고자 한다.

연구는 데이터를 수집하고 수집한 데이터를 이용하여 가중치를 구할 수 있도록 정제한 후, 정점과 간선을 포함한 그래프를 제작, 해당 그래프에서의 중심성을 추출하는 순서로 진행한다.

3.2. 데이터 수집

데이터는 국내 야구 통계사이트인 statiz에서 2022년 정규시즌 720경기의 각 팀 박스스코어 기록으로부터 수집하였다. 수집 방법으로 Python selenium 모듈로 웹사이트를 크롤링하는 방법을 사용했다. 경기 자체의 변수를 이용한 네트워크를 구성하기 위하여 각 선수의 기록을 통합한 팀 단위의 기록을 수집하였으며, 타자기록에서 19개, 투수기록에서 18개, 수비기록에서 7개의 요인을 조사하였다. 일차적으로 수집한 기록은 다음과 같다.

batter		pitcher		defense
TPA(총타석)	AB(타석)	IP(이닝수)	TBF(타자수)	IP(이닝수)
R(득점)	H(안타)	H(피안타)	R(실점)	PO(풋아웃)
HR(홈런)	RBI(타격득점)	ER(자책점)	BB(볼넷)	A(어시스트)
BB(볼넷)	HBP(데드볼)	HBP(데드볼)	K(삼진)	E(실책)
SO(삼진아웃)	GO(땅볼아웃)	HR(피홈런)		_P(직접실책)
FO(뜬공아웃)		PIT-S(투구수-스트라이크)		_A(간접실책)
GDP(병살)		GSC(투수평점)		
AVG(타율)		WHIP(이닝당 볼넷안타 허용수)		
PIT(투구수)		WPA(승리확률기여도)		
LOB(잔루)		RE24(기대득점)		
LI(현재상황의 중요도)		LI(현재상황 중요도)		
RE24(기대득점)		ERA(평균자책점)		
OPS(출루율+ 장타율)		IR-IS(승계주자-승계주자득점)		
WPA(승리확률기여도)		GO-FO(땅볼아웃-뜬공아웃)		

3.3. 데이터 정제

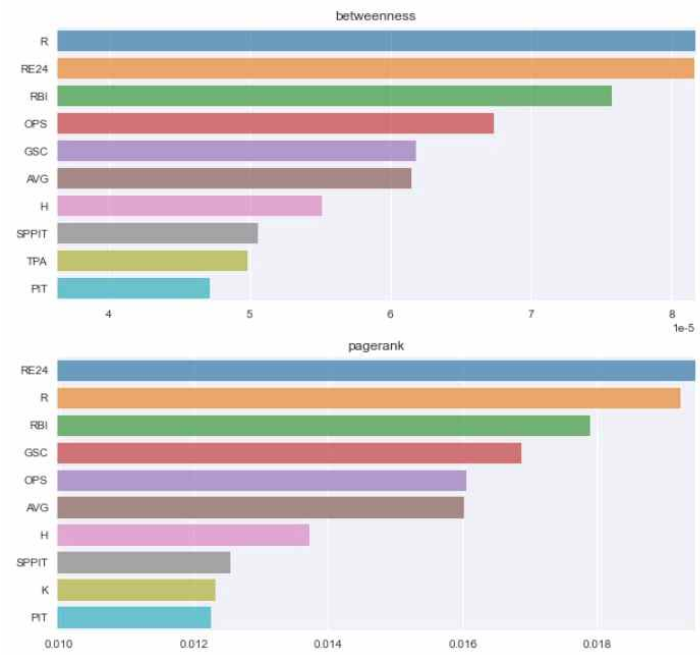
수집한 데이터를 구현하고자 하는 그래프로 나타내기 위하여 여러 전처리 과정을 필요로 한다. 전처리의 모든 작업은 Python pandas dataframe 모듈을 이용하여 수행하였으며, 먼저 변환과 분리를 진행했다. 확보한 데이터는 경기를 치른 두 팀 각각의 타자기록, 투수기록, 수비기록으로 나누어져 있으므로, 득점을 기준으로 승리팀, 패배팀을 구분하여 각각의 기록을 (승리팀의 기록) - (패배팀의 기록)으로 나타냈다. 이때, 무승부인 경기는 승리로 이어질 수 없으므로, 제외하였다. 해당 과정을 위해 투수기록의 GO-FO, PIT-S를 각각 비율(GO/FO, S/PIT)로 나타내는 GOPFO, SPPIT로 변환했으며, IR-IS는 IR, IS 각각 하나의 열으로 분리했다.

이후 불필요하거나 중복되는 열을 제거하는 과정을 진행했다. 경기 내용과 관계없이 동일하거나 유사한 값을 가지는 수비기록의 GDP, PO, 타자, 투수기록의 WPA, 투수, 수비기록의 IP를 제거하였으며, 남은 기록 중 타자의 기록, 투수의 기록에 중복되어 나타나는 7가지 항목(R, H, HR, BB, HBP, LI, RE24)는 행위의 주체자가 되는 측의 기록을 사용했다. 즉, 공을 치거나 경기를 결정지을 수 있는 특정 중요한 상황을 맞이하는 경우에 포함되는 H, HR, LI, RE24는 타자의 기록을, 공을 던지는 것에 관여하는 BB, HBP는 투수의 기록을 사용하도록 했다.

마지막으로 세 분류의 가중치를 하나의 테이블로 통합한 후, 각 변수를 정규화하는 과정을 진행했다. 각 변수는 비율로 나타난 값도 있고, 횟수로 나타나는 값들도 있어서 그대로 간선의 가중치에 할당한다면, 각각의 가중치의 규모에 크게 의존할 수밖에 없는 결과를 도출할 것이다. 모든 요인을 같은 범주로 나타내어 수치에 따른 중요도를 나타내어보기 위해 해당 연구에서는 최소-최대 정규화를 사용했다. 0부터 1 사이의 같은 범주 내에서 (승리팀 기록) - (패배팀 기록)의 수치가 클수록 높은 값을 가지도록 하여 서로 다른 요인이 승리에 작용할 수 있는 정도를 동일하게 구성하였다. 전처리 과정을 수행하며, 총 708개의 경기와 32개의 요인을 추려낼 수 있었다.

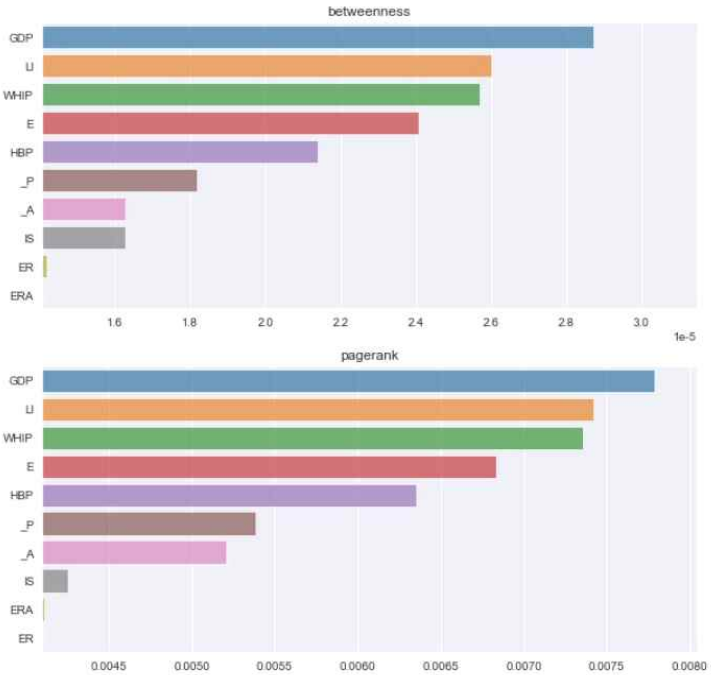
3.4. 그래프 생성 및 중심성 측정

해당 그래프는 708개의 경기, 32개의 요인, 1개의 승리 결과의 741개의 정점과 각 경기에서 (승리팀의 요인) - (패배팀의 요인) 값이 양수인 경우 해당 요인 정점과 정규화된 가중치를 가지고 연결된 간선 10,162개, 각 요인과 승리 결과 정점을 연결하는 32개의 간선을 합하여 10,194개의 간선으로 구성된다. 하나의 경기가 다른 경기의 중심성에 영향을 주지 않도록 모든 간선에는 방향을 설정해두었다. 간선의 방향은 경기에서 요인으로, 요인에서 승리결과로 이어진다. 해당 그래프에서 각 요인의 Betweenness Centrality와 PageRank Centrality를 구하고, 중심성을 높은 순, 그리고 낮은 순으로 10개를 시각화하였다. 중심성 수치는 다음과 같다.



[그림1 상위 10개 요인의 매개, PageRank 중심성]

두 중심성에서 R, RE24, RBI, OPS의 타자기록과 관련된 변수가 최상위 중심성을 보여주었고, 상위 10개 기록 중 Betweenness 중심성의 경우 8개, PageRank 중심성의 경우 7개의 값을 타자기록이 차지하고 있다. 수비기록은 상위 10개의 요인에 등장하지 않았지만, 투수기록 중 GSC는 타율(AVG)보다 높은 위치에 모습을 보이며, SPPIT(투구 중 스트라이크 비율) 또한 높은 중심성을 보인다.



[그림1 하위 10개 요인의 매개, PageRank 중심성]

하위 10개 중심성의 경우 두 중심성이 동일한 순위를 보인다. ER, ERA가 가장 작은 중심성을 보이고 있으며, 2개의 타자기록, 3개의 수비기록, 5개의 투수기록이 그래프 상에 나타난다.

3.5. 결과분석

중심성 수치는 결과적으로 그래프에서 각 요인과 승리와의 연관성을 나타낸다. 일반적으로 많은 경기와 연결될수록 높은 중심성을 추출할 수 있다. 야구 경기에서 중요한 요인일 것으로 추측할 수 있는 대표적인 요인인 홈런과 완봉경기가 해당 그래프의 중심성에서 상위권을 보여주지 못한 이유도 발생한 경우 승리로 이어질 확률이 매우 높지만, 모든 경기에서 발생하지 않기 때문에, 승리한 경기에서 일반적으로 높은 중심성을 보이는 요인들보다 낮은 값을 보인다고 추측한다.

또한 중심성이 낮은 요인들은 승리하는 경기에서 패배한 상대에게 일반적으로 밀리는 요인들이거나 드물게 등장하는 요인임을 나타낸다. 실책과 같이 실제로 패배와 연관된 경우도 있을 수 있지만, 극단적인 예로 720경기 중 홈런이 단 한 번 나왔다면, 해당 경기가 승리하였더라도, 그래프 중심성은 매우 낮을 것이다. 이를 구분하기 위해서는 경기의 패배로 이어지는 요인들에 대한 그래프를 구성해 보아야 할 것이다.

4. 결론 및 향후연구

본 논문에서는 프로야구 정규시즌의 여러 기록과 경기, 승리 사이의 기록들에서 중요도를 파악하고자 네트워크를 구성하고 그래프 중심성을 구하였다. 2022년의 각 경기와 수집 후 전처리를 거친 32개의 기록을 정점으로, 각각의 기록에서 승자팀과 패자팀의 차를 구하고 정규화한 값을 가중치로 활용하여 간선을 만든 그래프에서 Betweenness Centrality와 PageRank Centrality를 사용하여 중심성을 측정하였고, 승리한 경기와 연관이 높은 요인들을 확인할 수 있었다.

향후에는 경기 패배로 이어지는 그래프를 만들고, 해당 그래프에서의 중심성을 측정하여 패배에 기여한 요인, 드물게 등장해 평가할 수 없는 요인을 분류하고자 한다.

참고문헌

[1] 조태수, 한치근, 이상훈, “그래프 중심성들을 이용한 그래프 유사도 측정”, 한국컴퓨터정보학회논문지, 2018.

[2] Andrew Disney “PageRank centrality & EigenCentrality”, https://cambridge-intelligence.com/eigencentrality-pagerank/, 2022.

[3] 이석원, 천영진, “다중회귀분석을 이용한 메이저리그 승률의 모형구축과 예측”, 한국자료분석학회, 2017.