|  |
| --- |
| **그래프 중심성을 이용한 프로야구 정규시즌 분석**    **요 약**  프로야구는 4대 스포츠라고 불리는 축구, 야구, 농구, 배구 중에 가장 인기 있는 종목이다. 야구는 대중의 관심 못지 않게 학자들에게 관심 받는 연구 대상이다. 야구는 통계자료의 양도 많고 종류도 다양하기 때문에 학자들이 다양한 주제로 분석하고자 한다.  본 프로젝트에서는 프로야구 정규시즌 분석을 통해 어떤 경기 지표가 승리에 결정적인 요인인지 분석하고 분석 결과가 실제 시즌 팀 순위와 얼마나 일치하는지 확인하려고 한다. |

**1. 서론**

**1.1. 연구배경**

기록경기가 아닌 대부분의 스포츠 경기에서는 수비와 공격이 동시에 진행되는 반면, 야구는 공격과 수비가 번갈아 이루어지는데 특히 공격에서는 선수 각각이 상대 팀 투수와 수비 선수들을 상대하게 된다. 따라서 야구에서의 통계자료의 양과 질은 다른 종목에 비하여 더 복잡한 양상을 보인다. 또한 야구는 수비와 공격에서 동료선수들과의 협력 뿐 아니라 선수들의 개별적 활약을 기록하여 통계자료를 만든다. 그만큼 통계자료의 양도 많고 종류도 다양하여 통계분석에 의한 게임 승패의 예측 가능성이 상대적으로 어렵다.

본 프로젝트에서는 이 같은 점에서 흥미를 갖고 프로야구 정규시즌을 분석하려고 한다. 팀 타율, 출루율, 장타율, 실책, era, whip 등의 지표 중에서 어떤 요인이 가장 승패에 영향을 주는지를 분석하려 한다. 또한 분석한 결과와 실제 시즌 팀 순위가 얼마나 일치하는지를 확인하려 한다.

**1.2. 연구목표**

우선, 프로야구 10개 팀의 정규시즌 전적 경기를 통하여 팀 간의 전력 차이(타자력 – 출루율, 장타율, 클러치타율 등, 투수력 – era, whip 등, 수비력, 실책 등)를 비교하려고 한다. 이후 그래프 중심성을 이용하여 중요도를 분석하여 결정적인 경기 요인을 알아내려 한다. 마지막으로, 분석한 각 팀의 중요도 서열과 시즌 순위를 비교하려 한다.

**2. 관련연구**

**2.1. 사회 연결망 분석(SNA)**

사회 연결망 분석은 개인이나 사회, 집단의 관계를 네트워크의 구조로 분석하는 방법이다. 노드와 엣지를 기반으로 하는 그래프를 사용하여 객체들간의 관계, 각 노드의 중요도를 파악하기 위해 사용한다. 사회 연결망의 분석 기준으로 밀도와 중심성이 있으며, 본 프로젝트에서는 그래프 중심성을 활용한다.

**2.1.1. 그래프 중심성**

네트워크에서 중심성은 하나의 노드가 전체 네트워크에서 어느 정도의 영향력을 가지고 있는지, 즉, 네트워크의 중심에 위치하는 정도를 뜻한다. 사회 연결망을 분석하는데 중심성을 사용하는 경우에는 네트워크에서 연결 정도, 노드의 영향력, 노드 간에 관계가 닿아있는 정도 등, 무엇을 중요하게 볼 것인지에 대한 기준을 정해야 한다. 각 노드의 중심성 수치를 구하기 위한 기준으로는 연결 정도(Degree), 근접(Closeness), 매개(Betweenness), 그리고 고유 벡터(Eigenvector)가 있다.  
 연결정도 중심성은 직접 연결된 노드의 수를 기반으로 측정한다. 이 경우, 하나의 엣지를 통해 직접적으로 연결되어 있는 수가 많을수록 중요도가 올라간다. 근접 중심성은 노드 사이의 최단거리를 기반으로 측정된다. 근접 중심성을 구하면 연결정도 중심성에서 측정할 수 없었던 직접 연결되지 않은 노드와의 관계 또한 수치에 포함시킬 수 있다. 매개 중심성은 다른 노드 사이의 최단거리에 포함된 수를 이용하여 측정한다. 노드들 사이에 위치하는 정도를 나타내고 있으며, 지나는 경로가 많을수록 중심성이 높다. 고유벡터 중심성은 인접행렬의 고윳값에 기반하여 노드에 네트워크에서의 영향력을 계산하고, 연결된 노드와 자신의 영향력을 반영하여 측정한다. 이외에도 고유벡터 중심성에서 파생된 Katz 중심성, 웹 페이지 및 하이퍼링크 등에 활용되는PageRank 중심성, 연결 정도 중심성에서 파생된 가중 연결 정도 중심성 등이 있다.

**2.2. 기존 통계자료를 이용한 프로야구 분석**

기존의 연구에서는 통계자료를 이용한 승리예측, 득, 실점의 원인 파악 등을 목적으로 분석을 시행한다. 타력, 득점력, 투수력 등의 통계 기록을 이용하여 머신러닝, 딥러닝 분석을 수행한 후, 각 방법의 정확도를 비교한다.

**2.3. 기존 연구의 문제점 및 해결 방안**

**2.3.1. 연구의 문제점**

모든 변수는 야구 경기에 작용하는 요소를 수치화한 것이므로, 경기의 승패와의 상관관계가 존재한다. 하지만 한국시리즈, 플레이오프 등 에서는 중요한 경기에서의 긴장이나 라이벌 매치 등 외부적인 요소로 인해 달라진 경기력이 존재할 수 있고, 역전이나 승부를 결정짓는 데에 기여한 결정적 변수는 다를 수 있다.

**2.3.2. 해결 방안**

본 프로젝트에서는 팀이나 경기, 선수와 같은 노드 사이에 존재하는 관계에 각 중요변수에 해당하는 가중치를 부여하고, 팀과 팀 사이의 경기에 승, 패로 이어지는 방향성 네트워크를 구성하여 중요도를 분석하고자 한다. 그리고 각 팀의 중요도의 서열이 실제 경기 결과와 얼마나 일치하는지 확인하여 중요변수가 승리에 기여하는 정도를 파악한다.

**3. 프로젝트 내용**

**3.1. 시나리오**

**3.1.1. 데이터 확보 및 정제**

데이터 확보는 웹 크롤링을 기반으로 한다. KBO 공식 홈페이지와 Statiz에서 제공하는 선수와 팀의 정보를 기반으로 확보한다. 확보한 데이터를 이용하여 선수, 팀 간의 네트워크를 구현할 수 있도록 데이터 정제를 수행한다. 정제를 수행하면서, 노드로 사용할 데이터와 관계를 나타낼 데이터를 구분짓도록 한다.

**3.1.2. Graph Centrality 알고리즘 구현**

앞서 정제한 데이터를 바탕으로, 네트워크 기반의 중요변수 도출을 위한 그래프 중심성 측정 알고리즘을 정한다. 이후 각 중요변수를 이용하여 엣지 혹은 노드에 방향성 및 가중치를 부여하여 승리, 패배, 소속의 관계를 가지는 그래프를 구현할 수 있도록 한다.

현재 구상중인 프로젝트의 네트워크는 국내 야구 프로팀, 각 날짜의 경기를 노드로 하고, 프로팀과 경기 사이의 관계를 승리, 패배를 이용한 방향성 엣지로 나타내어 구성할 예정이다. 이에 따라 방향성을 가지고, 각 수치에 따른 가중치를 적용하여 구할 수 있는 그래프 중심성 알고리즘을 계획 중에 있다. 현재로서는 경기와 팀 사이의 관계가 가장 중요한 요소로 작용할 것이라는 판단 하에 degree centrality를 기반으로 한 알고리즘을 계획하고 있다. 하지만 이것은 실제 네트워크 설계 전 이론적 그래프의 형태로 제안하는 것으로, 그래프의 형태나 네트워크의 구조를 단기간의 데이터로 먼저 구성해보고 적절한 중심성 측정 알고리즘이 무엇일지는 이후에 정하는 것이 올바른 결과를 도출할 수 있을 것이다.

**3.1.3. 결과 분석**

Gephi, networkX, igraph 등 그래프를 시각화할 수 있는 도구 중 정제 및 알고리즘이 적용된 데이터를 나타내기에 적절한 툴을 선택하여 네트워크를 시각화한다. 이때 수치로 노드의 중요도를 계산하며, 승패와의 상관성을 분석해서 각 중요변수가 승부와 연관된 정도를 파악한다.

도출된 결과를 바탕으로 한국시리즈 경기에 오른 두 팀의 경기에 대한 노드를 생성했을 때, 노드가 두 팀과 연결될 확률을 계산하여 분석의 결과를 평가한다.

**3.2. 요구사항**

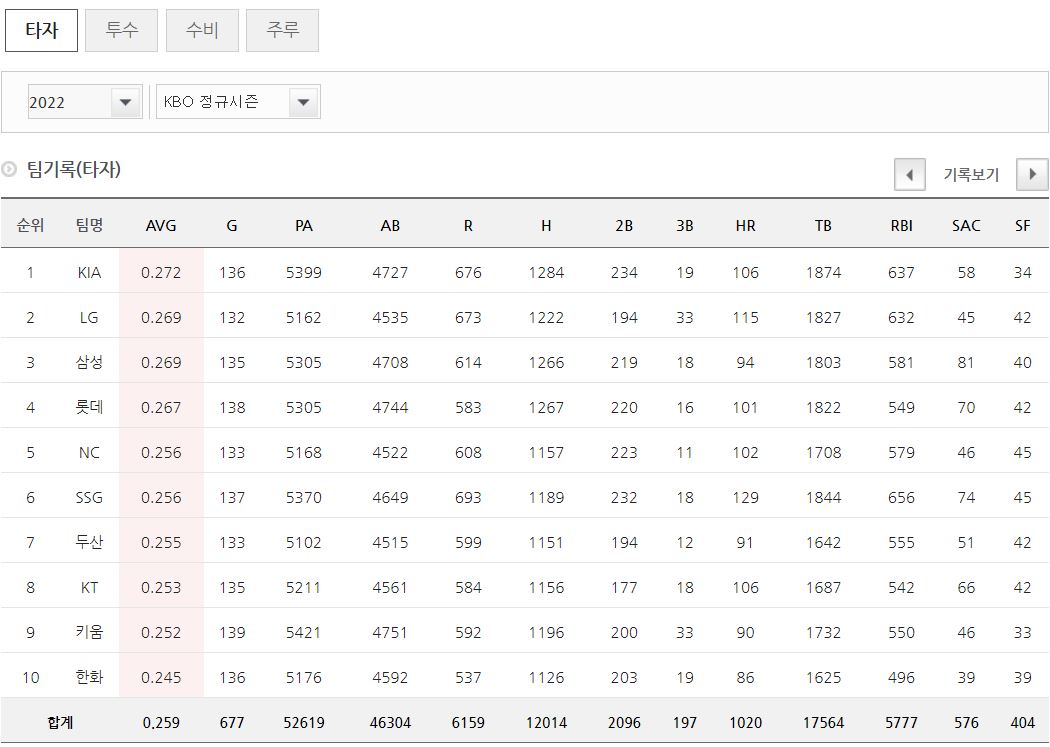
**3.2.1. 데이터 확보 및 정제에 대한 요구사항**

- 웹사이트의 기록실 및 팀 간에 치뤄진 경기의 승패 정보를 크롤링한다.

- 확보한 자료는 네트워크 구성을 위해 관계를 나타낼 수 있는 소속이나 상대한 팀, 경기 결과(승, 패)에 대한 정보를 담고있어야 한다.

 - 정제 과정에서 노드와 엣지를 명확하게 구분지을 수 있도록 한다.

**그림 1 [팀 간의 경기 정보]**

**그림 2 [팀 기록]**

**3.2.2. 프로토타입 제작**

- 수집한 데이터에서 일주일부터 한달 정도 크기의 데이터를 분할한다.

- 분할한 데이터를 활용하여 소규모 네트워크를 구성한다.

- 구성된 네트워크를 Graph Centrality 알고리즘을 결정할 인사이트로서 활용한다.

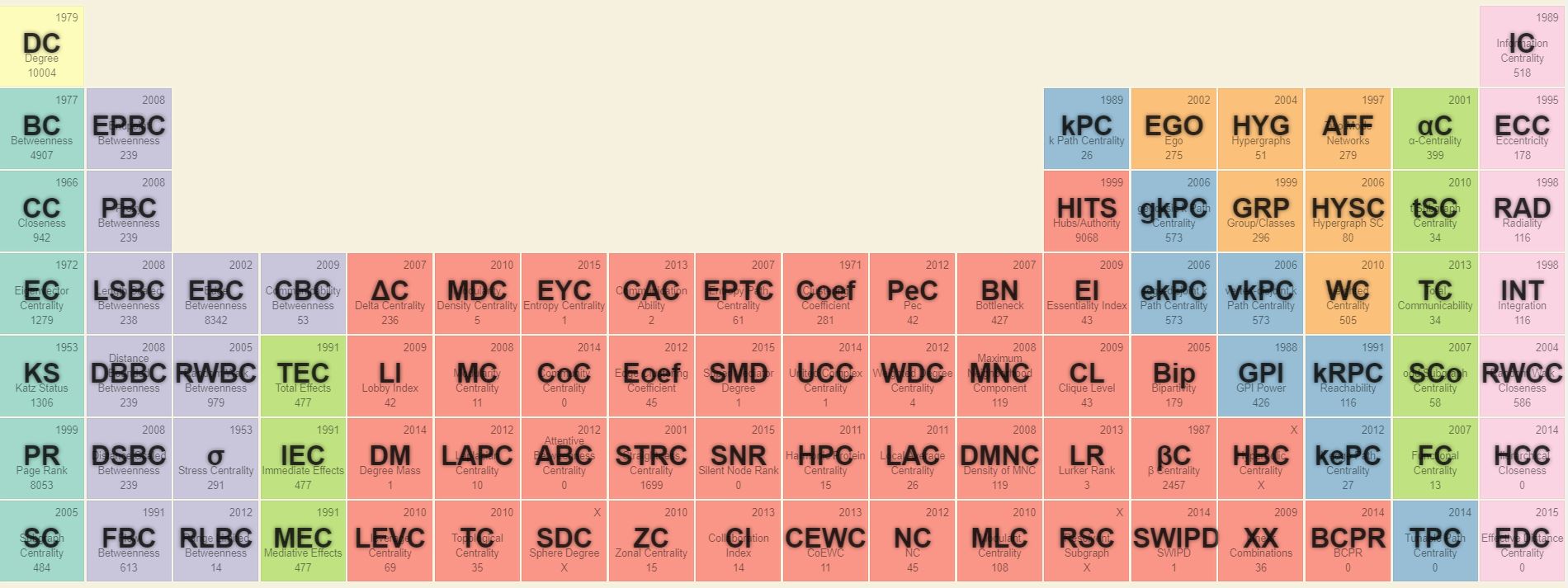
**3.2.3. Graph Centrality 알고리즘에 대한 요구사항**

- 유의미한 결과가 도출될 수 있도록 적절한 중심성 알고리즘을 선정한다.

- 시각화 툴이나 라이브러리에 맞는 형태의 결과를 이끌어낼 수 있는 알고리즘을 구현한다.

- 단순한 데이터의 일반화로 인해 생기는 기초적 형태의 결과 도출을 지양한다.

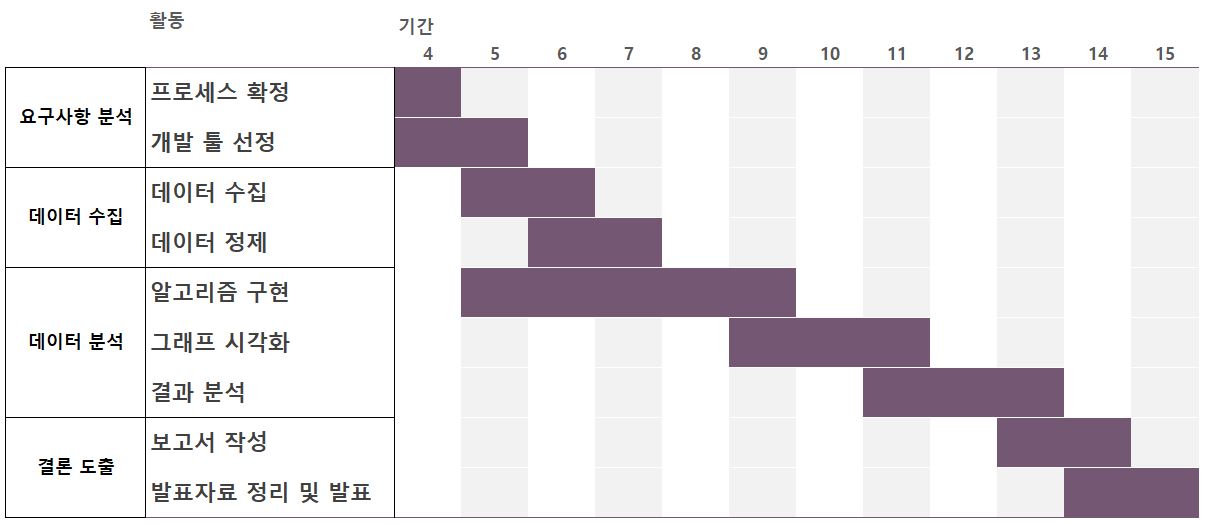
- 그래프 중심성을 구하는 알고리즘은 논리적 근거를 가져야 한다. (방향성 네트워크에서의 그래프 중심성, 여러 중심성 알고리즘 참고)



**그림 1 [중심성을 측정하기 위한 지표]**

**4. 향후 일정 및 역할 분담**

**4.1. 향후 일정**

****

**4.2. 역할 분담**

|  |  |
| --- | --- |
| **이름** | **활동** |
| 김창민 | - 크롤링 코드 작성  - 노드 / 엣지로 수집데이터 분리  - 그래프 중심성에 따른 노드별 중요도 추출  - 그래프 연결 가능성을 계산하여 경기 노드를 추가하였을 때의 승리예측 |
| 우하은 | - 수집 데이터 선정  - 중요변수별 네트워크에서 가질 수 있는 가중치 분석  - 네트워크의 노드 및 엣지에 가중치 할당  - 상관계수, 회귀계수 등을 활용하여 실제 경기결과와 노드 중요도 비교분석 |
| 공통업무 | - 네트워크 모형 설계 및 프로토타입 구현  - 프로야구 및 네트워크에 관한 연구 결과를 활용하여 네트워크에서 활용할 수 있는 가중치 탐색  - 그래프 알고리즘 구현 및 네트워크 시각화  - 보고서, 발표자료 등 문서 작성 |

**5. 결론 및 기대효과**

본 프로젝트는 프로야구의 승리에 기여하는 가장 중요한 변수를 네트워크 구성 및 중심성 측정이라는 새로운 접근 방법으로 분석해보고자 한다. 해당 분석은 전반적인 통계 결과를 통해 예측을 수행하는 일반적인 분석 모델과는 다르게 하나하나의 경기를 노드로 구성함으로써 서로 간의 관계를 경기 예측에 포함한다. 이를 통해 스포츠 분야의 승부 결과 예측에 경기, 팀 사이의 관계가 어느 정도 작용하고 있는지를 결과로 볼 수 있고, 해당 분석에서 계산한 중심성이 운동경기의 결과에 영향을 미치게 된다면, 스포츠 분석 분야에 새로운 인사이트를 제공하게 될 것이다.

**6. 참고문헌**

[1] 노택선 · 민충기 · 조용준\*, 한국외국어대학교 · 수원시정연구원\*, 『한국프로야구의 정규시즌과 포스트시즌에서 승리기여도의 차이에 관한 실증연구: 승리 결정방식을 중심으로 』 , 2021

[2] 김창권 · 진서훈, 고려대학교, 『프로야구 기록을 통한 승리 요인에 관한 연구』 , 2014

[3] 최영근 · 김형문, 건국대학교, 『한국프로야구 경기결과에 관한 통계적 연구』, 2014