Project Proposal: Baseball KG

October 2022

Fandel Lin
fandel.lin@usc.edu

Zihao Han
zihaohan@usc.edu

1 PROJECT DOMAIN AND GOAL

The project objective is *building MLB knowledge graph for player and game prediction with statistical analysis.*

Baseball is one of the major professional sports in the United States, and can attract tens of thousands of on-site spectators per match. However, instead of simply skimming through the match results, there are complex interactive and synergistic relationships among players in baseball.

In this project, we plan to build a knowledge graph about baseball games. The knowledge graph will hold teams of a baseball league, match schedules of teams, game results of matches, players of teams, historical records of players, and statistics of players in matches. For different types (e.g., pitchers and hitters) of players in a team, the statistics here span from pitch types, spin direction, and pitch movement for pitchers; to pitch tracking, plate discipline, and batted-ball profiles for hitters. This brings a total of at least 11 semantic types.

1.1 Project Goal

With visualization, this knowledge graph could help people dig into the interactive and synergistic relationship among players in baseball games. Furthermore, we plan to exploit this knowledge graph to provide a data-centric approach to inferring the performance of players and the results of matches.

1.2 Motivating Example

As a motivating example, spectators for baseball games not only want to know whether the team they support can win a match, but also want to take part in some 'special events' in person! Such events could be intuitive, for instance, a team winning the championship, or a pitcher (a player whose routine is stable) winning a match. However, most of such events are uncertain, for instance, a team playing a 'perfect game' (no opposing player ever reaches base by any means in a match, e.g., there are only 21 perfect games in MLB since 1903), a pitcher getting his shot outs or complete games, or a hitter getting his career home run in round hundreds (e.g., Albert Pujols gets his 698^{th} home run on 9/16, but his 699^{th} and 700^{th} ones both on 9/24). These types of events are difficult to be expected if a spectator only relies on the impression of recent situations of players or teams.

2 DATASET

(1) MLB.com
- The official website of Major League Baseball.
- Contains all season data for 30 total teams.

(2) Baseballsavant.mlb.com
- The official statistics of Major League Baseball.
- Contains statistics of each player.

(3) Wikipedia.org
- Any other information of players beside match statistics.

There are 2430 matches and at least 1200 active players per year in MLB. We will crawl the game and schedule info from the first dataset, and the player info from all three datasets.

All three datasets have structured tables, this brings around 6030 pages and more than 2400 structured records. We will design custom ontologies that capture the concepts from the structured sources, for instance, (team, has match against, team), (team, has, player), (player, plays in, match), (player, plays against, player), (pitcher, has, pitch type), (hitter, has, pitch tracking), (player, belongs to, player type), (player, plays against, player type), and (player type, plays against, player type), etc.

3 TECHNICAL CHALLENGE

The project will solve technical challenges in analyzing knowledge graphs. The difficulties lie in categorizing entities (players) with similar characteristics, understanding the relationship among different types of entities, and exploiting the uncertainty in such relationships for an inference. To the best of our knowledge, few addresses the inference problem based on uncertain relationships in a knowledge graph. We conduct the experiment on a real-world dataset from the Major League Baseball (MLB), and evaluate the performance for the inference problem based on the accuracy of game results and player performance. (e.g., we take the first 70% matches in a season for constructing the predictive model, and use the last 30% matches for evaluation.)

We use SPARQL queries to extract and build our knowledge graph for players from MLB in various aspects. Precisely, we harvest different pitch types, spin direction, and pitch movement for pitchers; different pitch tracking, plate discipline, and batted-ball profiles for hitters from the baseball savant dataset. Based on these statistics, we can categorize these pitchers and hitters. We will use these trends to build a probabilistic predictive model among hitters and pitchers based on their categories. We then exploit the player list (roster) and game schedule of each team from the MLB dataset. However, the starting lineup for a match is only part of the player list of a team. Therefore, without knowing the exact players who are playing in matches, we have to infer the game results and player performance based on our predictive model under such an uncertainty in the KG relationship.

Compared to other ball games, MLB has more teams and players, so we have more relationships between players on each team. Plus with more game data, we could face a very large knowledge graph and then may slow down our predictions.

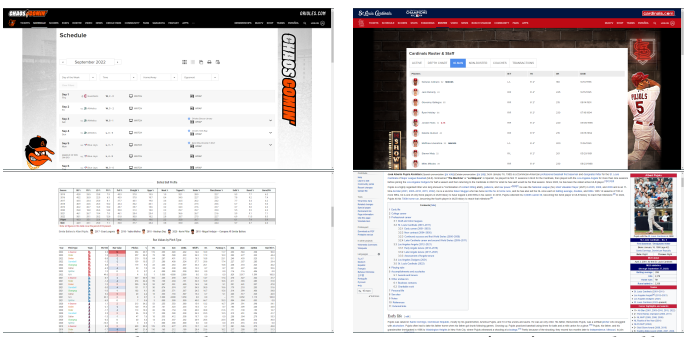

Figure 1: Three data sources: MLB.com (top), Baseballsavant.mlb.com, and Wikipedia.org (bottom, left to right).