Building English Premier League (EPL) Knowledge Graph and Game Prediction

A Final Report submitted for CSCI563 Project

Demo video link: https://www.youtube.com/watch?v=KnKZoMn4OZc

Jaeyoung Kim, Dong-Ho Lee
Email: jkim2458@usc.edu,dongho.lee@usc.edu

1 Introduction

Extrapolation, a reasoning process of predicting new facts from history, has gained increasing interests. In this project, we propose an event forecasting framework based on a temporal knowledge graph (TKG). We aim to collect data that contains temporal information and build a TKG in which each edge between entities is associated with temporal information and a set of interactions builds a multi-relational graph at each time. To effectively show the event forecasting as a project demo, we choose the sports domain that has a limited set of entities (e.g. team) and their interactions as edges that have temporal information (e.g. game, time).

2 Data Processing

Domain. The project aims to build a Knowledge Graph of English Premier League (EPL) and predicts future game results. First, the knowledge graph is constructed with following entities: 'Team', 'Player', and 'Game'. Each entity will contain following attributes:

(1) 'Player' - 'name', 'url', 'birth date', 'place of birth', 'nationality', 'height', 'weight', 'position', 'foot usage', 'age', 'current team', 'team history', 'strengths', 'weaknesses', 'style of play'
(2) 'Game' - 'season', 'round', 'score', 'date', 'home team', 'away team', 'home team players', 'away team players', 'goals and assists', 'player rating'
(3) 'Team' - 'name', 'url', 'date founded', 'chairman', 'owner', 'league', 'website', 'stadium'

Here we crawl data using Scrapy [1] and Selenium [2] from 3 unstructured sources and 1 structured data. To find records that refer to the same entity (e.g., player, game) across different data sources, we adopt probabilistic soft logic to conduct entity resolution. we adopt a blocking algorithm by player's birthday and game's home-away team information for player and game entity resolutions. For game, we conduct entity resolution twice (worldfootball-whoscored & whoscored-datahub).

category	url	Player	Game	Team	Season
Unstructured	https://worldfootball.net	2,832	4,444	39	09/10 - 20/21
Unstructured	https://whoscored.com	2,155	4,614	39	09/10 - 21/22
Unstructured	Wikipedia	-	-	39	-
Structured	https://datahub.io/sports-data/english-premier-league	-	3,420	-	09/10 - 18/19

	Player	Game
MC	11	27, 23
RR	99.94%	99.86%, 99.87%

Table 1: Data statistics & Blocking scheme results. MC is maximum canopy size and RR is reduction ratio.

3 Framework

To provide an experience of 1) analyzing knowledge graph for better game strategy and 2) future game record prediction, we construct the knowledge graph by rdflib and import it into the graph database management system Neo4J [3] and temporal knowledge graph (TKG) prediction model. Then, we create a back-end API with Flask [4] to request queries and get responses. We visualize the results using Vue.js [5] and Vis.js.

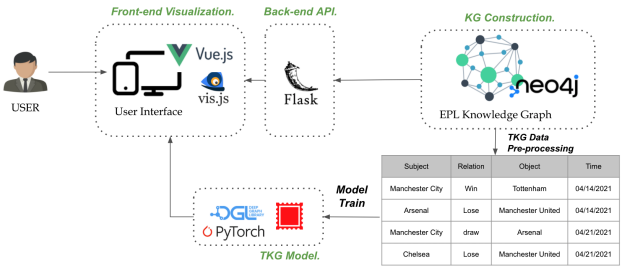

Figure 1: Framework overview.

[1] https://scrapy.org/
[2] https://www.selenium.dev/
[3] https://neo4j.com/
[4] https://flask.palletsprojects.com/en/2.0.x/
[5] https://vuejs.org/

3.1 Query Engineering & Front-end Visualization

To provide useful information to users, we first collect game analysis questions from news articles such as performance of certain players against other teams or best players in the match. Then, we construct the query templates. The reason why we construct the template is that it is difficult to parse the natural language form of question into query. By providing template style query and asking users to fill out the mask, we could give flexibility of query construction to users. We construct a mapping between template and Cypher query and exploit it to convert the user-filled template into Cypher query.

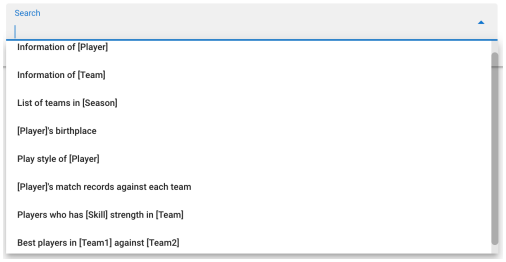

Figure 2: Query selection.

3.2 Temporal Knowledge Graph (TKG)

Many facts occur repeatedly along the history such as global economic crisis and diplomatic events. It also applies to domain of sports. Winning team is more likely to win in the sports. Inspired by this, CyGNet [6] models the historical facts with the same entity and relation to each query of entity prediction, and thus mainly focuses on predicting facts with repetitive patterns. Here, we exploit CyGNet to forecast future game results based on previous game records.

Problem Formulation. Given subject $s \in \mathcal{E}$, relation $p \in \mathcal{R}$ and object $o \in \mathcal{E}$ at time stamp $t \in \mathcal{T}$, where $\mathcal{E}, \mathcal{R}$ denote the corresponding vocabularies of entities and relations respectively and $\mathcal{T}$ is the set of timestamps, $g = (s, p, o, t)$ denotes a quadruple fact in $\mathcal{G}_t$. TKG is a set of quadruple facts ordered by timestamp: $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_T\}$. The learning objective is multi-class classification task to minimize cross-entropy loss on predicting a missing subject given $(s, p, ?, t)$ or object given $(?, p, o, t)$. Since the model can only deal with subject and object prediction, we need to reformulate the inference. Given $(s, ?, o, t)$, we reformulate the task into object prediction over enumerated quadruples by all the relation $p \in \mathcal{R}$ as follows: $\{(s, p_k, ?, t) \mid p_k \in \mathcal{R}\}$. Then, we choose p_k that shows the highest $\mathbf{p}(o \mid s, p_k, t)$. Here, the size of $\mathcal{E}$ and $\mathcal{R}$ is 39 and 3.

Evaluation. For subject and object prediction, we present results in Table 2. Results show that correct prediction mostly appears in top 15 among $\mathcal{E}$.

	# Train	# valid	# Test	# Total
	3,848	482	482	4,812
Start time	09/10 round 1	09/10 round 35	20/21 round 4	09/10 round 1
End time	18/19 round 34	20/21 round 4	20/21 round 13	20/21 round 13

	Hit@1	**Hit@3**	**Hit@5**	**Hit@10**	**Hit@15**	**MRR**
Subject Prediction	0.2714	0.5219	0.6785	0.8288	0.9102	0.4447
Object Prediction	0.2881	0.5741	0.7035	0.8622	0.9311	0.4711
Total	0.2797	0.5480	0.6910	0.8455	0.9207	0.4579

Table 2: Data statistics & TKG performance.

Inference. To make an inference on future game (starting from 20/21 round 14), we re-train the model. We recreate a dataset in Table 1 by aggregating train and valid set into train and consider test set as validation set. Here, we could find that our predictions show accuracy 60% on 20/21 round 14 which is the most recent game. It shows that sports game usually affects a lot from historical facts and shows repetitive patterns. For example, matches between strong and weak teams usually show a repeating pattern (e.g. Tottenham Hotspur vs. Brentford FC).

Home	Away	Game result	Prediction
Aston Villa	Manchester City	Lose	**Lose**
Manchester United	Arsenal FC	Draw	Win
Southhampton FC	Leicester City	Lose	Draw
Tottenham Hotspur	Brentford FC	Win	**Win**
Wolverhampton Wanderers	Burnley FC	Draw	**Draw**
Everton FC	Liverpool FC	Lose	**Lose**
Newcastle United	Norwich City	Win	Draw
West Ham United	Brighton Hove Albion	Draw	**Draw**
Watford FC	Chelsea FC	Lose	**Lose**
Leeds United	Crystal Palace	Lose	Win

Table 3: Home team's results against away team at 20/21 round 14.

4 Conclusion

In this paper, we present *Web framework* for analyzing knowledge graph of English Premier League and forecasting the game results using temporal knowledge graph. We believe that our work provides handful experiments of building a better game strategy and future game record forecasting to users.

[6] Learning from History: Modeling Temporal Knowledge Graphs with Sequential Copy-Generation Networks, Zhu et al., AAAI 2021