한국컴퓨터정보학회 동계학술대회 논문집 제24권 제1호 (2016. 1)

R을 활용한 야구 통계 데이터 다차원 시각화 도구

김주희[O], 최용석[*]
[O]한양대학교 컴퓨터 소프트웨어
[*]한양대학교 공과대학 컴퓨터공학부
e-mail: kjhh1203@gmail.com[O], cys@hanyang.ac.kr[*]

Multi-dimensional Visualization Tool for Baseball Statistical Data Using R

Ju Hee Kim[O], Yong Suk Choi[*]
[O]Dept of Computer Software, Hanyang University
[*]Dept of Computer Science and Engineering, Hanyang University

● 요 약 ●

본 연구에서는 대용량의 야구 데이터를 R 패키지인 googleVis를 이용하여 시각화하는 웹페이지를 구축하고, 버블 차트로 시각화하여 표현하였다. 웹페이지에서는 시각화하는 객체를 버블로 나타내며, 객체는 타자, 투수, 팀 3가지이다. 각 객체의 속성들을 버블 색상, 버블 사이즈, X-Y좌표, 연도에 설정함으로써 5차원으로 시각화하여 표현할 수 있게 한다. 웹페이지 기능 중 타임 슬립 애니메이션을 사용하여 시간의 흐름에 따른 기록 변화를 한 눈에 관찰할 수 있으며, 선수 검색 기능을 통해 특정 선수들을 선택하여 비교 및 분석하는 것이 가능하다.

키워드: 시각화(visualization), googleVis, 빅데이터(Big Data)

I. 서 론

대용량 데이터에 대한 분석 기술이 집적, 심화됨에 따라 현대 스포츠에서 데이터의 중요성은 점점 더 증가하고 있다. 특히 야구는 '기록의 스포츠'라 불릴 정도로 한 경기에도 수많은 데이터가 생성되는 종목으로 잘 알려져 있어, 다른 종목에 비해 기록이 가지는 중요성은 더욱 크다고 볼 수 있다.[1] 이전 경기 또는 선수 데이터를 분석해 데이터를 기반으로 적재적소에 선수들을 배치해 승률을 높일 수 있으며 전략을 세울 때 중요한 요소가 된다[2].

이러한 야구 데이터에 대한 일반인들의 관심이 증가함에 따라 프로 야구 데이터를 제공하는 사이트에 대한 방문과 관심 역시 증가하고 있다.[3] 프로 야구 데이터를 제공하는 웹 사이트의 경우, 일반적으로 표로 기록에 해당하는 수치를 표시하거나 그래프를 통해 기록을 시각화하여 제공한다.

본 논문에서는 MLB 선수의 시즌별 기록 데이터와 한국 프로야구 선수, 팀 시즌별 기록 데이터를 기반으로 구글에서 제공하는 R 패키지인 googleVis를 이용하여 시각화하는 웹페이지를 구축하였다. 데이터는 버블차트로 나타내며 버블의 사이즈, 버블의 색상, 좌표에서 기록의 속성들을 변경해가며 비교 및 분석이 가능하다.

II. 관련 연구

II-1. 기존 시각화 방법

국내에서는 KBO(Korean Baseball Organization, 한국프로야구) 공식홈페이지[1]와 statiz[2]라는 웹사이트에서 야구 데이터를 표로 제공하고 있으며, Ballgraphi[3]라는 웹사이트에서는 시즌별로 Bar 차트, 버블 차트, 파이 차트 등으로 시각화해서 제공하고 있다. MLB(Major League Baseball, 미국프로야구)의 경우 MLB공식 홈페이지[4]와 Baseball-reference.com[5]이라는 웹사이트에서 야구 기록 데이터를 표로 제공하고 있다.

본 논문에서는 기존 방법에 더하여 타임 슬립 애니메이션을 이용하여 시간의 흐름에 따른 기록의 변화를 한눈에 볼 수 있도록 제공한다.

II-2. googleVis :구글차트를 위한 R 인터페이스

googleVis 패키지는 차트 API로써 R과 구글간의 인터페이스를 제공하고, 대화형 차트를 제공한다.

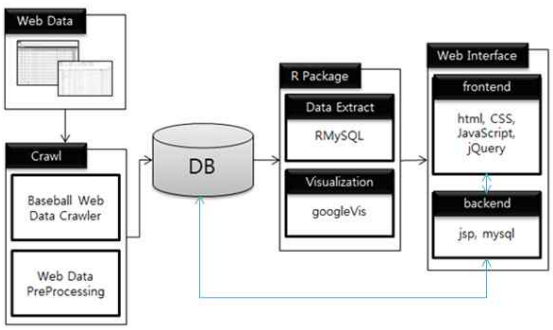
가장 널리 사용되는 차트로는 모션 차트가 있으며 Hans Rosling의

1) http://www.koreabaseball.com
2) http://www.statiz.co.kr
3) http://ballgraphi.com
4) http://mlb.mlb.com
5) http://www.baseball-reference.com

TED 강연에서 사용된 차트로도 잘 알려져있다. 본 논문에서는 모션 차트를 사용하며 시간에 따른 동적인 객체의 움직임을 관찰할 수 있다. googleVis 패키지의 기능은 사용자가 구글에 데이터를 직접 업로드 하지 않고 구글 차트와 R data frame에 저장된 데이터로 시각화할 수 있도록 도와준다. googleVis function의 산출물은 데이터와 구글에서 호스팅된 JavaScript function들의 html 코드이다. 차트는 Flash로 만들어진다[4].

III. 본론

본 논문 목적은 대용량의 야구 기록 데이터를 R을 이용하여 시각화하는 웹페이지를 생성하고, 생성된 웹페이지를 통하여 시즌별로 기록의 변화를 쉽게 알아볼 수 있도록 사용자에게 시각화 도구을 제공하는 것이다. 개발된 시각화 도구는 한양대학교 인공지능 연구실 서버[1]에서 프로토타입으로 제공하고 있다. [그림 1]은 본 논문에서 구축한 시스템 아키텍처이다. [그림 1]에서 보여지는 바와 같이 우선 웹에 공개되어 있는 야구 기록 데이터를 수집한 후 전처리하여 DB에 저장한다. 데이터 수집이 완료되면 DB에 저장한 데이터를 R에서 불러온 후 R 패키지를 이용하여 시각화 웹페이지를 생성한다. 이때 시각화 및 분석을 위해 불러오는 데이터는 선수 이름, 시즌, 팀, 데뷔년도, 안타, 득점, 완투, 홈런, 삼진 등이며 R 패키지는 RMySQL과 googleVis를 사용한다. 만들어진 시각화 웹페이지에서는 특정 선수들을 선택하여 비교 및 분석하기 위한 용도로 선수 이름으로 질의(query)할 수 있는 검색 기능을 제공한다.



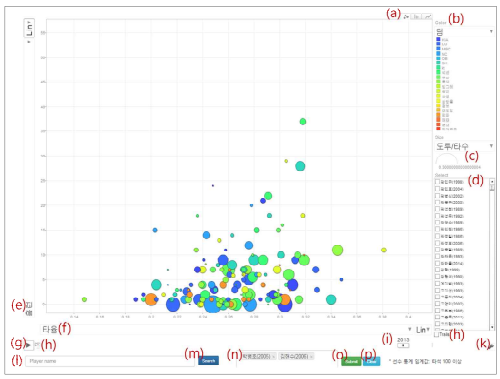
[그림 1] 시스템 아키텍처

III-1. 구현 및 화면 구성요소

본 논문에서는 R을 이용한 야구 통계 데이터 시각화를 위하여 R version 3.2.0, googleVis version 0.5.10, mysql version 5.1.73 환경에서 HTML, Jsp, CSS, JavaScript, jQuery로 구현하였다. [그림 2]는 처음 웹페이지에 방문하였을 때 화면이다. MLB 투수 기록 통계, MLB 타자 기록 통계, 한국 프로야구 팀 기록 통계, 한국 프로야구 투수 기록 통계, 한국 프로야구 타자 기록 통계 웹페이지 순으로 나열되어 있으며, 그림을 클릭하면 해당 웹페이지로 이동하도록 하이퍼링크로 연결하였다.

[그림 2] 야구 통계 시각화 웹 페이지 메인

각 웹페이지는 공통적으로 [그림 3]과 같은 형태로 구성되어있다. [표 1]은 [그림 3]의 각 요소별 기능을 설명한다. 각 웹페이지에서는 시각화하는 객체를 버블로 나타내며, 각 객체들의 속성을 색상, 버블 사이즈, 좌표 3가지 방법으로 표현한다. 본 논문에서 시각화하는 객체는 타자, 투수, 팀 3가지이며 각 객체의 속성은 [표 2]에서 확인할 수 있다.



[그림 3] 야구 통계 시각화 웹 페이지

요소	기능
(a)차트 종류	차트 타입 변경(googleVis에서는 버블, bar, 라인 3가지 차트를 제공하는데 본 논문에서는 버블 차트만 사용함)
(b)Color	버블의 색상 변경
(c)Size indicator	사이즈로 나타낼 속성을 선택
(d)Select variables	박스안의 객체를 선택하면 특정 버블이 선택됨.
(e)Y-axis	Y축 야구 기록 속성 변경
(f)X-axis	X축 야구 기록 속성 변경
(g)Play / Stop	타임 슬립 애니메이션을 컨트롤할 play/stop 버튼
(h)Speed of animation	애니메이션 속도 조절
(i)Time	클릭과 드래그로 년도를 변경
(j)Trails	타임 슬립 애니메이션이 재생될 때 선택된 버블의 년도 별 움직임을 트레일로 나타냄
(k)Settings	선택된 버블 외에 선택되지 않은 버블의 투명도 선택
(l)Player name input bar	비교 및 분석을 하고자하는 선수의 이름을 입력(본 논문에서는 팀

1) http://166.104.142.94:62000/baseballStatVisualization/main.html

	검색은 제공하지 않음)
(m)Search button	선수 검색 버튼
(n)Selected player name bar	선수 다중 선택
(o)submit button	선택된 선수들의 객체만을 그래프에 나타냄
(p)clear button	그래프 초기화 버튼

[표 1] 웹페이지 요소별 기능

객체	속성
타자	타석, 타수, 득점, 안타, 2타, 3타, 홈런, 루타수, 타점, 도루, 도루 실패, 볼넷, 사구, 고의4구, 삼진, 병살, 희생타, 희생플라이, 타율, 출루율, 장타율, OPS
투수	출장 경기, 완투, 완봉, 선발 출장 수, 승, 패, 세이브, 홀드, 이닝, 실점, 자책점, 상대 타자, 피안타, 피2타, 피3타, 피홈런, 볼넷, 고의 4구, 사구, 삼진, 보크, 폭투, ERA, FIP, WHIP
팀	경기당홈런수, 관중수, 도루율, 등수, 방어율, 승률, 에러율, 타율

[표 2] 야구 통계 시각화 웹 페이지에 사용된 속성

III-2. 시각화 분석

1. KBO 타고투저 현상 분석

[그림 4]는 타임 슬립 애니메이션 기능을 이용하여 2011년 투수 및 타자 통계와 2015년 투수 및 타자 통계를 비교한 것이다. 타자 통계 차트에서는 X-Y좌표를 타율-홈런, 버블 색상은 팀, 버블 사이즈는 피홈런으로 설정하였고, 투수 통계 차트에서는 X-Y좌표를 평균자책점-승, 버블 색상은 팀, 버블 사이즈는 득점으로 설정하였다. 2011년 타자 통계와 2015년 타자 통계를 비교해보면, 2015년도 그래프 우측 상단에 더 많은 객체들이 분포되어있음을 알 수 있다. 이는 2011년에 비해 2015년에 좋은 성적을 낸 타자들이 더 많았음을 의미한다. 반면 2011년과 2015년 투수 통계를 비교해보면, 객체 집단이 대체적으로 오른쪽으로 이동한 것을 확인할 수 있다. 이는 전체적인 투수들의 평균자책점이 증가했음을 나타내는 것으로 2011년에 비해 2015년의 투수 성적이 좋지 않았음을 의미한다. 이와 같이 KBO리그가 지난 수년간 타고투저 현상이 있었다는 것을 시각화 도구로 확인할 수 있다.

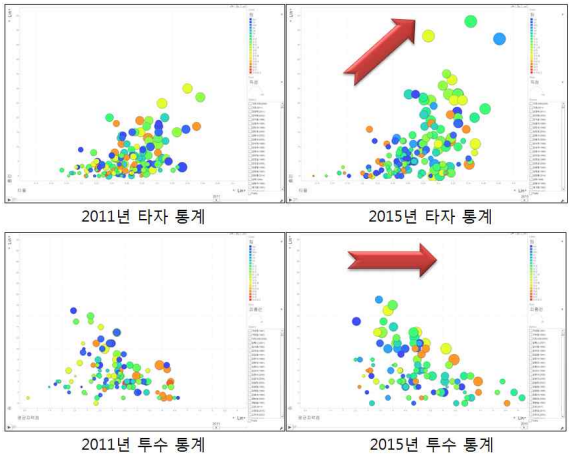
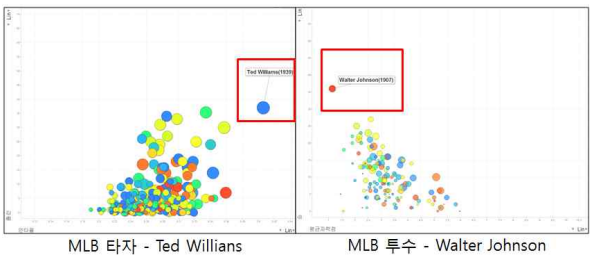


그림 4] 타고투저 현상 시각화

2. 객체 집단에서의 우수 선수 검출

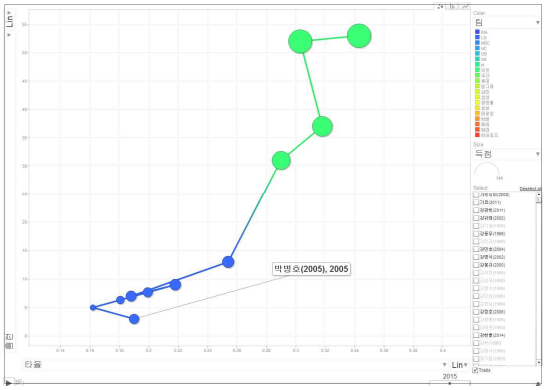
[그림 5]는 MLB 타자 및 투수 통계를 시각화 한 그래프이다. 타자 차트에서는 X-Y좌표를 타율-홈런, 버블 색상은 팀, 버블 사이즈는 득점으로 설정하였고, 투수 차트에서는 X-Y좌표를 평균자책점-승, 버블 색상은 팀, 버블 사이즈는 피홈런으로 설정하였다. 해당 그래프에서 다른 객체 집단과 멀리 떨어진 객체를 발견할 수 있는데, 이를 통해 당해년도 기록이 타 선수에 비해 월등한 선수를 쉽게 찾아낼 수 있다.



[그림 5] 객체 집단에서의 우수 선수 검출

3. 단일 선수 분석

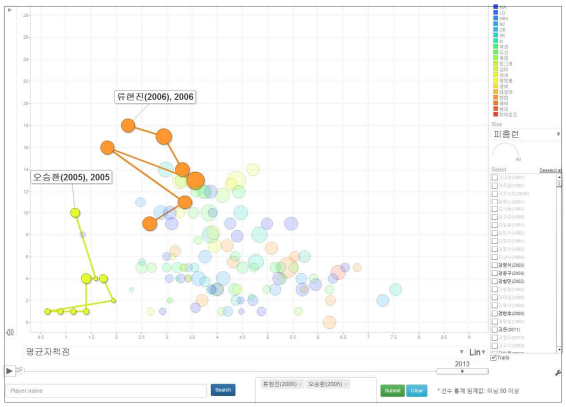
[그림 6]은 검색 기능을 통해 특정 선수의 버블만 보이도록 선택되지 않은 버블의 불투명도를 0%로 설정하고, 선수의 년도 별 기록 변화를 트레일로 나타낸 것이다. 차트에서 X-Y좌표를 타율-홈런, 버블 색상은 팀, 버블 사이즈는 득점으로 설정하였다. 예를 들어, 2005년 데뷔 KBO 타자 박병호 선수의 년도 별 기록으로 알 수 있는 것은 버블의 색상 변경으로 팀의 변경이 있었다는 것을 알 수 있고, 시간이 지날수록 성적이 좋아지는 것을 확인할 수 있다.



[그림 6] 단일 선수 분석

4. 선수간 비교 분석

[그림 7]은 선수 검색 기능을 통해 특정 선수 여러명을 선택하고, 선수들의 년도 별 기록 변화를 트레일로 나타낸 것이다. 이때 선택되지 않은 버블의 불투명도는 20%로 설정하였다. 차트에서 X-Y좌표를 평균자책점-승, 버블 색상은 팀, 버블 사이즈는 피홈런으로 설정하였다. 예를 들어, 2005년 데뷔 KBO 투수 오승환 선수와 2006년 데뷔한 류현진 선수의 기록을 2005년부터 2013년까지 비교 분석하자면 오승환 선수에 해당하는 버블은 좌측 아래에 포진되어 있고, 류현진 선수에 해당하는 버블은 오승환 선수의 버블보다 오른쪽 위쪽에 포진되어 있다는 것을 알 수 있다. 오승환 선수는 류현진 선수보다 승이 적고 평균자책점이 낮은 것을 확인 할 수 있는데, 이는 오승환 선수가 마무리 투수임을 의미한다. 류현진 선수의 경우 선발 투수로 오승환보다 승이 많고 평균자책점은 높은 것을 확인할 수 있다.



[그림 7] 선수간 비교 분석

Ⅳ. 결론 및 향후 연구 계획

본 논문에서 구축한 시각화 웹페이지는 대용량의 야구 기록 데이터를 R 패키지인 googleVis를 통해 시각화하여 제공하는 웹페이지이다. 해당 웹페이지를 통해 시간 순으로 기록의 변화를 관찰할 수 있고, 기록의 속성들을 바꿔가면서 다각도적으로 선수의 기록을 분석할 수 있음을 확인하였다.

향후 본 논문에서 구축한 시각화 웹페이지를 통해 기록의 유사성을 나타내는 선수들의 상관계수를 분석하여, 특정 선수의 향후 기록을 예측하는 연구를 진행할 예정이다.

참고 문헌

[1] SeokMi Hong, KyungSook Jung, TaeChoong Chung, "Win/Lose Prediction System : Predicting Baseball Game Results using a Hybrid Machine Learning Model", Journal of KIISE, Vol. 9, No. 6, pp. 693-698, December. 2003.

[2] wikipedia. https://ko.wikipedia.org/wiki/%EB%B9%85_%EB%8D%B0%EC%9D%B4%ED%84%B0

[3] Younhak Oh, Han Kim. Jaesub Yun, Jong-Seok Lee “Using Data Mining Techniques to Predict Win-Loss in Korean Professional Baseball Games”, Journal of the Korean Institute of Industrial Engineers, Vol. 40, No. 1, pp. 8-17, February 2014.

[4] Markus Gesmann, "Using the Google Chart Tools with R: googleVis-0.5.10 Package Vignette", August 26. 2015.