The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 20, No. 4, pp.177-185, Aug. 31, 2020. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2020.20.4.177
JIIBC 2020-4-25

빅데이터 기반 프로야구 데이터 분석

Analysis of Professional Baseball Data based on Big Data

신동진[*], 황승연[**], 이돈희[***], 문진용[****], 김정준[*****]

Dong-Jin Shin[*], Seung-Yeon Hwang[**], Don-Hee Lee[***],
Jin-Yong Moon[****], Jeong-Joon Kim[*****]

요 약 최근 프로야구의 스포츠 인기는 나날이 증가하고 있으며, 다양한 포털 사이트에서 프로야구와 관련된 데이터를 소유하고 있다. 프로야구의 인기를 증가시키고, 관련된 데이터를 활용한 분석을 통해 결과를 만들어 낸다면 프로야구를 접하는데 이점이 있다. 본 논문에서는 프로야구와 관련된 데이터를 활용하여 3가지 분석을 시행하였다. 따라서 본 논문에서는 특정 사이트에서 조회된 특정 프로야구단과 관련된 기사 개수와 트렌드를 알아보고, 프로야구 성적과 관중 수의 상관관계에 대해서 분석하였다. 마지막으로 2016, 2017년도의 프로야구 타자 타율 성적과 출루율 성적에 대한 현황 분석을 실시하였다.

Abstract Recently, the popularity of professional baseball is increasing day by day, and it has data related to professional baseball on various portal sites. If you want to increase the popularity of professional baseball and produce results through analysis using relevant data, you have the advantage of accessing professional baseball. In this paper, three analyzes were conducted using data related to professional baseball. Therefore, in this paper, the trend related to the number of articles retrieved from a specific site of a professional baseball team was examined, and the correlation between professional baseball scores and the number of spectators was analyzed. Finally, we analyzed the current status of professional baseball batting average and on base percentage in 2016 and 2017.

Key Words : Big Data, KBO Statistics Data, Team Trend Analysis, Team Record Analysis

Ⅰ. 서 론

최근 대한민국에서는 가장 핫한 스포츠로 프로야구가 뽑히고 있으며 아래의 표를 보면 확실히 야구의 인기는 다른 스포츠들을 압도하고 있다. 매년 프로야구팬들이 증가하고 있지만, 그들 모두가 프로야구에 대해서 잘 아는 것은 아니다. 어떤 스포츠도 마찬가지겠지만, 그 스포츠에 대해 잘 아는 팬층과 야구를 모르는 팬층 간의 괴리감은 있기 마련이다. 물론 이러한 괴리감이 크게 영향을 끼치지 않을 수도 있지만, 처음 프로야구를 접하거나 접하

[*]준회원, 안양대학교 컴퓨터공학과 박사과정
[**]준회원, 안양대학교 컴퓨터공학과 석사과정
[***]정회원, SK 주식회사 수석
[****]정회원, 강동대학교 방송영상미디어과 교수
[*****]정회원, 안양대학교 ICT융합학부 소프트웨어전공 교수
접수일자 2020년 4월 3일, 수정완료 2020년 6월 12일
게재확정일자 2020년 8월 7일

Received: 3 April, 2020 / Revised: 12 June, 2020 /
Accepted: 7 August, 2020
[*****]Corresponding Author: jjkim@anyang.ac.kr
Dept. ICT Convergence Engineering, Anyang University, Korea.

려고 하는 사람들한테는 큰 영향을 끼칠수도 있다. 따라서, 이러한 괴리감을 조금이나마 없애고, 자신이 응원하고 싶은 팀을 선택하는데 보탬이 돼서 초심자가 더욱 프로야구를 즐길 수 있는 환경이 필요하다.

프로야구 특정 구단의 트렌드를 분석하기 위해서 필요한 데이터의 수집은 크롤링과 셀레늄을 이용하여 기사의 본문과 기사 개수를 측정하며, 하이브를 이용하여 불필요한 단어를 필터링 하고, R 프로그래밍을 통해 그래프로 표현한다[1].

프로야구 성적과 관중 수의 관계를 분석하기 위해서 필요한 데이터의 수집은 KBO 사이트를 참고한다. 관중 및 성적 데이터란 연도별 각 프로야구 구단들의 관중 수 및 성적, 또는 월별 각 프로야구 구단들의 관중 수 및 성적을 나타낸다[2]. 프로야구 타율 성적과 출루율 성적을 분석하기 위해서 필요한 데이터는 Betman 공식 사이트를 이용하여 수집하며, R프로그래밍을 통해 그래프로 표현하였다[3].

본 논문의 전체 구성은 다음과 같다. 2장의 관련 기술을 소개하고 3장에서는 사용하는 기술을 바탕으로 데이터와 주요 기술 코드를 소개하고, 4장에서 분석된 결과를 보여주며, 5장 결론을 기술한다.

II. 관련 기술

1. 빅데이터(Big Data)

4차 산업혁명의 기술이 발전하면서 다양한 종류의 데이터를 포함하는 빅데이터가 떠오르고 있다. 정형, 반정형, 비정형 데이터와 같은 예전에 보관만 되어있는 수많은 데이터가 다양한 프로그램이나 소프트웨어에 의해서 분석을 하면, 유의미한 결과를 나타낼 수 있다[4-5].

초기에는 3V라고 정의되어 Volume, Velocity, Variety 순으로 크기, 속도, 다양성의 의미를 많이 사용되었지만, 최근 기술과 데이터의 발전으로 인해 5V로 정의되어 Value, Veracity로 가치와 정확성의 의미가 추가되었다.

2. 크롤링

무수히 많은 웹사이트에 저장 및 개시되어있는 뉴스 기사나 문서를 수집하여 검색 대상의 Text 문서를 수집하는 기술이다. 대표적으로 BeautifulSoup 라이브러리를 사용하며, 웹사이트의 태그를 조작하여 수집된다.

스파이더(Spider), 봇(Bot), 지능 에이전트라고도 불리며, 방대한 자료를 검색하는 특징은 있으나 역이용하여 순위를 조작하거나 검색을 피할 수 있는 단점이 존재하지만, 범용성이 좋아 네이버, 구글 등 다양한 서비스를 제공하는 업체에서도 빼놓을 수 없는 기술이다.

3. 셀레늄

셀레늄(Selenium)은 웹 애플리케이션 테스트를 위한 포터블 프레임워크다. 셀레늄은 테스트 스크립트 언어를 학습할 필요 없이 기능 테스트를 만들기 위한 플레이백 도구를 제공한다. (셀레늄 IDE) C 샤프, 그루비, 자바, 펄, PHP, 파이썬, 루비, 스칼라 등 수많은 유명 프로그래밍 언어들에서 테스트를 작성하기 위한 테스트 도메인 특화 언어(Selenese)를 제공한다.

4. 하이브

하이브는 Apache에서 개발한 SQL 형식의 병렬 처리 데이터베이스를 의미한다. 하둡에 저장되어있는 데이터를 Map-Reduce를 직접 구현하여 데이터를 분석하기엔 개발 능력과 경험이 필요하기 때문에 이를 이해하고 구현하는 시간과 노력을 줄이고자 개발되었다.

HiveQL이라고 불리는 SQL 형식의 문법을 사용하기 때문에 기존에 사용자들이 편하게 사용할 수 있으며, 하둡을 기반으로 동작하기 때문에 하둡이 설치가 되어 있어야 한다[6-7].

5. R 프로그래밍

1993년 오클랜드 대학에서 통계 분석과 결과로 그래프를 표현하기 위해 개발된 인터프리터 프로그래밍 언어이다. R 프로그래밍은 오픈소스로 이루어져 있기 때문에 범용성이 넓어 사용자가 편리하게 사용할 수 있으며, 수많은 통계 관련 패키지가 존재 하기 때문에 편리하게 사용할 수 있다.

최근에는 웹 어플리케이션 개발 프레임워크인 Shiny의 발전으로 통계 또는 머신러닝과 관련된 모델을 웹과 연동하여 표현할 수 있기 때문에 사용자가 시각적으로 느낄 수 있는 분석 프로그램 중에 많이 활용되고 있다.

III. 데이터 수집 및 처리

이번 장에서는 크롤링 및 셀레늄 주요 코드를 소개하

고, 분석에 사용된 데이터 소개와 처리 과정에 대하여 기술 한다. 그림 1은 특정 사이트의 뉴스 기사를 수집하는 크롤링과 관련된 코드를 보여준다.

1. 특정 프로야구단 기사 데이터 수집 및 처리

```
    from bs4 import BeautifulSoup

    def get_text(URL, output_file):
        source_code_from_url = urllib.request.urlopen(URL)
        soup = BeautifulSoup(source_code_from_url, 'lxml',
            from_encoding='utf-8')
        content_of_article = soup.select('div.article_txt')
        for item in content_of_article:
            string_item = str(item.find_all(text=True))
            output_file.write(string_item)

    for title in soup.find_all('a', {'class': "go_naver"}):
        now += str(title['href'])
        source_code_from_sub =
urllib.request.urlopen(title['href'])
        soup_sub = BeautifulSoup(source_code_from_sub,
'lxml', from_encoding='utf-8')
        content = soup_sub.find_all('div',{"id":
"newsEndContents"})

    shutil.copy('C:/Users/ihc/Downloads/multiTimeline.csv',
SAVE_FILE_PATH + '/' + keyword)
```

그림 1. 데이터 수집을 위한 크롤링 코드
Fig. 1. Crawling code for data collection

BeautifulSoup 라이브러리를 이용하여 데이터를 수집하였으며, BeatifulSoup 라이브러리는 HTML에서 데이터를 추출하고, 구문 분석된 페이지에 대한 구문 분석 트리를 작성하기 때문에 웹 스크래핑에 유용하다.

라이브러리 아래 코드는 크롤링에 핵심적인 코드를 설명하며, 페이지를 변수에 저장하고 해당 페이지에 포함된 모든 기사의 주소를 모두 가져오고, 기사 본문과 URL은 파일 입출력을 통해 저장한다. 페이지 검색은 마지막 페이지 변수를 통해 마지막 페이지가 나올 때까지 계속 반복해서 돌리며, 마지막 페이지에 도달했다면 스크랩 작업을 중지한다.

마지막으로 수집된 데이터를 로컬 데스크탑에 저장하기 위해 코드를 추가하여 크롤링을 실행하였다.

```
def getCSV(teams, keyword, startdate, enddate):
    print('\n----------' + keyword + ' Google
Trends Data Download Started!----------')
    binary =
'C:/Users/ihc/AppData/Local/Programs/Python/Python36-3
2/Lib/site-packages/selenium/webdriver/chrome/chromedri
ver.exe'
    driver = webdriver.Chrome(binary)
        #driver = webdriver.Firefox()
```

그림 2. 데이터 수집을 위한 셀레늄 코드
Fig. 2. Selenium code for data collection

그림 2는 구글 사이트에서 제공하는 특정 프로야구단의 트렌드 관련 CSV 파일을 다운로드 받기 위한 핵심 코드를 의미한다. Chrome에 WebDriver를 설치하여 구글 트렌드와 관련된 데이터를 수집하였다.

그림 3. 크롤링 실행 결과
Fig. 3. Crawling execution reseult

그림 3은 그림1의 데이터를 이용하여 크롤링을 실행한 결과를 보여준다. 특정 프로야구단에 대해 관련된 기사를 계속 다운로드 받게 된다.

그림 4. 크롤링을 통해 수집된 특정 프로야구단 기사
Fig. 4. Certain professional baseball articles through crawling

크롤링 과정을 통해 수집된 데이터의 모습을 보여준다. 그림3의 결과이며, 특정 프로야구단을 선택해서 크롤링과 셀레늄을 하였으며, 수집된 기사의 년도는 2012년도부터 2017년도까지 데이터를 수집되었다.

그림 5는 크롤링과 셀레늄을 통해 수집된 기사 데이터의 처리하는 하이브의 주요 코드를 보여준다. 하이브는 하둡위에서 동작하며, 병렬 처리를 하기 때문에 일반 DBMS에 비해서 빠른 속도로 처리가 가능하다.

```
Create table word(keyword string, count int)
row format delimited
fields terminated by '\t'
lines terminated by '\n';

load data local input
'/mnt/share/Crawling/result/특정프로야구단/part-r-00000'
overwrite into table word;

insert overwrite local directory
'/mnt/share/Crawling/result/특정프로야구단/lasall_data'
row format delimited fields terminated by '\t\'

select * from word where keyword not in ('수', '있다', '는', '한',
'있는', '첫', '중', '더', '큰', '그', '또', '될', '하지만', …
```

그림 5. 데이터 처리를 위한 하이브 코드
Fig. 5. Hive code for data processing

word 테이블을 생성하고, 키워드를 저장할 컬럼, 그리고 기사의 개수를 확인할 수 있는 카운트 컬럼을 생성한다. 생성 후 수집되어 저장된 데이터를 테이블에 입력하고, 탭을 기준으로 기사 데이터를 저장한다. 마지막에 select 구문은 단어 필터링을 위한 '수', '있다', '는' 과 같은 불용어를 포함 시키지 않기 위한 구문이다.

2. 프로야구 성적과 관중 데이터 수집 및 처리

프로야구 성적과 관중 수 데이터를 수집하기 위해서 KBO 사이트에서 데이터를 추출하였으며, 데이터를 수집한 방법과 처리한 방법을 소개한다.

그림 6은 프로야구 성적과 관중 수 데이터를 수집하기 위해서 KBO 사이트에서 보유하고 있는 년도별, 월별, 구단별 성적 데이터의 모습을 보여준다. 데이터 수집이 완료된 후 그림 7은 KBO 사이트에서 추출한 데이터를 CSV 파일 형태로 변환 후 액셀을 이용하여 데이터를 정형화하여 처리한 모습을 보여준다.

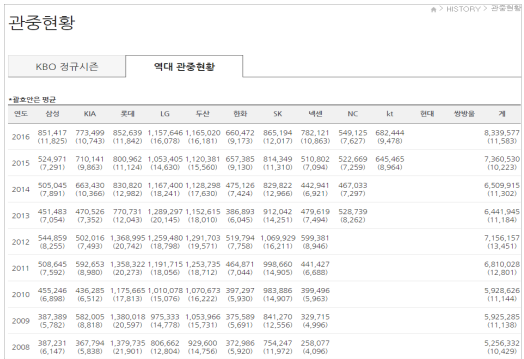


그림 6. KBO 사이트 데이터 현황
Fig. 6. KBO site data status

연도	삼성	KIA	롯데	LG	두산	한화	SK	넥센	NC	KT	계
2016	851417	773499	852639	1157646	1165020	660472	865194	782121	549125	682444	8339577
2015	524971	710141	800962	1053405	1120381	657385	814349	510802	522669	645465	7360530
2014	505045	663430	830320	1167300	1128298	475126	829822	442941	467033	0	6509915
2013	451483	470526	770731	1289297	1152615	386893	912042	479619	528739	0	6441945
2012	544859	502016	1368995	1259480	1291703	519794	1069929	599381	0	0	7156157
2011	508645	592653	1358322	1191715	1253735	464871	998660	441427	0	0	6810028
2010	455246	436285	1175665	1010078	1070673	397297	983886	399496	0	0	5928626
2009	387389	582005	1380018	975333	1053966	375589	841270	329715	0	0	5925285
2008	387231	367794	1379735	806662	929600	372986	754247	258077	0	0	5256332

그림 7. 데이터 처리를 완료한 모습
Fig. 7. Data processing completed

3. 프로야구 타자 타율과 출루율 성적 수집 및 처리

프로야구 타자 타율과 출루율 성적 데이터를 수집하기 위해서 Batman 사이트에서 데이터를 추출하였으며, 데이터를 수집한 방법과 처리한 방법을 소개한다.

순위	선수	팀명	경기수	타율	타수	득점	안타			홈런	타점	도루	볼넷	삼진	병살	출루율	장타율
							총합	2루타	3루타								
1	김선빈	KIA	69	0.378	233	41	88	20	0	2	41	3	21	26	5	0.427	0.489
2	서건창	넥센	67	0.365	260	46	95	16	2	4	45	9	35	23	6	0.436	0.488
3	나성범	NC	51	0.361	202	46	73	17	1	11	41	9	14	47	2	0.412	0.619
4	이대호	롯데	67	0.356	250	33	89	6	0	12	44	0	21	37	10	0.419	0.524
5	김태균	한화	50	0.349	189	27	66	12	0	7	43	0	26	29	4	0.434	0.524
6	김재환	두산	67	0.341	270	47	92	17	1	15	45	2	32	56	4	0.419	0.578
6	이명기	KIA	58	0.341	232	35	79	10	3	3	37	2	13	29	1	0.378	0.448
8	최형우	KIA	68	0.339	242	46	82	18	2	16	52	0	51	34	8	0.458	0.628
9	최주환	두산	62	0.338	201	37	68	7	3	5	35	3	21	23	5	0.409	0.478
9	박용택	LG	64	0.338	234	36	79	12	1	3	39	3	29	43	6	0.407	0.436
11	안치홍	KIA	63	0.335	233	50	78	13	2	9	43	4	23	29	9	0.401	0.524
12	이정후	넥센	70	0.331	248	56	82	15	3	2	24	5	27	31	6	0.399	0.440
12	손아섭	롯데	69	0.331	272	53	90	16	3	7	33	10	39	41	4	0.419	0.489
14	양의지	두산	57	0.330	188	30	62	12	0	9	44	1	23	24	8	0.417	0.537
15	로사리오	한화	58	0.322	230	50	74	14	1	18	55	5	23	31	9	0.397	0.626
16	모창민	NC	67	0.319	251	34	80	15	3	8	51	3	15	45	3	0.356	0.498
16	윤석민	넥센	69	0.319	263	42	84	12	1	6	41	0	18	36	16	0.371	0.441
16	솔광민	한화	65	0.319	238	37	76	16	0	5	42	2	15	47	5	0.356	0.450
19	민병헌	두산	66	0.318	258	42	82	12	0	7	37	1	22	44	6	0.388	0.446

그림 8. Batman 사이트 데이터 현황
Fig. 8. Batman site data status

rank	player	team	match	batavg	bat	score	total_hit	2B	3B	HR	RBI	run	BB	3s	DP	OBP
1	조윤준	LG	2	1	1	1	1	0	0	0	0	0	0	0	0	1
2	박광열	NC	13	0.4	15	0	6	0	0	0	0	0	0	3	1	0.4
3	나주환	SK	24	0.385965	57	13	22	2	0	5	12	0	2	6	4	0.435484
4	최형우	KIA	138	0.375723	519	99	195	46	2	31	144	2	83	83	12	0.464401
5	정진호	두산	31	0.375	24	9	9	0	1	0	3	0	6	1	0	0.53125
6	김태균	한화	144	0.364839	529	94	193	39	0	23	136	1	108	97	11	0.47546
7	이용규	한화	113	0.35177	452	98	159	20	4	3	41	21	63	29	7	0.437859
8	김태완	넥센	24	0.35	20	3	7	1	1	0	0	0	3	8	0	0.458333
9	김주찬	KIA	130	0.34638	511	97	177	37	3	23	101	9	30	68	13	0.386282
10	박용택	LG	138	0.345776	509	84	176	24	0	11	90	6	58	71	13	0.411765
11	구자욱	삼성	108	0.343458	428	105	147	19	13	14	77	10	55	68	4	0.419878
12	박민우	NC	121	0.342529	435	84	149	16	6	3	55	20	55	70	9	0.419802
13	유한준	kt	110	0.335784	408	70	137	22	0	14	64	4	47	40	10	0.404348
14	조영훈	NC	109	0.335196	179	29	60	14	0	5	35	3	21	43	2	0.407767
15	박건우	두산	132	0.334711	484	95	162	36	4	20	83	17	38	86	9	0.389925
16	윤석민	넥센	92	0.334311	341	72	114	15	0	19	80	2	46	50	16	0.420398
17	고종욱	넥센	133	0.333966	527	92	176	22	9	8	72	28	28	103	10	0.369912
18	최재원	LG	28	0.333333	81	20	27	3	0	4	16	3	17	17	2	0.456311
18	백창수	LG	10	0.333333	18	4	6	2	0	1	3	0	2	6	1	0.428571
20	이진영	kt	115	0.331536	371	49	123	23	1	10	72	2	45	44	13	0.403302
21	모창민	NC	63	0.330827	133	16	44	8	0	5	20	3	9	24	7	0.369863
22	최정민	KIA	88	0.329032	155	29	51	4	1	0	11	8	8	48	1	0.385965
23	최진행	한화	28	0.328571	70	4	23	2	0	1	9	0	11	16	1	0.433735

그림 9 CSV형태로 데이터를 변환한 모습
Fig. 9. Data converted into CSV format

그림 8은 프로야구 성적과 관중 수 데이터를 수집하기 위해서 Batman사이트에서 보유하고 있는 데이터의 형식을 보여준다. 타율을 기준으로 정렬되어 있으며, 관련된 선수의 선수명과 팀명, 타율, 출루율 등 모습을 보여준다.

그림 9는 분석에 필요한 데이터를 CSV형태로 변환한 모습을 보여주며, 이중에서 타율과 출루율 컬럼을 활용한다.

Ⅳ. 분석 결과

본 장에서는 앞서 수집하고 처리한 데이터의 분석 그래프를 보여주며, 분석 결과의 순서는 특정 프로야구단의 기사 관련 분석, 프로야구 성적과 관중 수 관계 분석, 타자 타율과 출루율 성적 분석 순으로 진행된다.

1. 특정 프로야구단 기사 분석

```
tbl <- read.csv("graphdata.csv")

tbl

plot(tbl$검색.횟수, type="l", col="navy", lwd=3, ylim=c(0,110),
axes=FALSE, xlab="DAY", ylab="SEARCH", font.lab=2,
font.lab=2)

axis(1, at=1:260, lab=tbl$날짜)
axis(2, ylim=c(0,80000))

title(main="DOOSAN BEARS TREND", col.main="dark slate
gray", font.main=10, cex.main=3)
```

그림 10. 특정 프로야구단 기사 빈도 코드
Fig. 10. Specific professional baseball article frequency code

그림 10은 특정 프로야구단의 기사에 관련된 빈도가 얼마나 나왔는지 분석 코드를 보여준다. 처리가 완료된 csv파일을 tbl 변수에 불러온 뒤 plot 함수로 그래프로 표현하였다. 분석 결과는 그림 11과 같다.

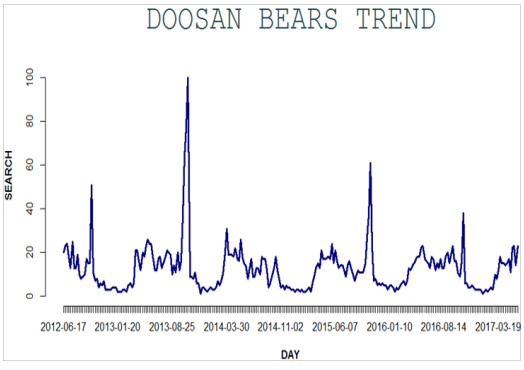


그림 11. 특정 프로야구단 기사 빈도 그래프 결과
Fig. 11. Specific professional baseball article frequency graph results

그림 11을 보면 2012년 중반부터 2017년도 초반까지의 특정 프로야구단에 관련된 기사 빈도수를 보여주며, 2013년 후반부에 높은 빈도의 기사가 검출된 것을 볼 수 있다. 2013년도에 관련된 특정 프로야구단의 이슈를 확인해보니 논문에 언급하기 어렵지만, 관련된 이슈가 많이 올라온 것을 확인할 수 있었다.

```
library(KoNLP)
library(wordcloud)

f <- file("sub2016-10-16~2016-10-30.txt", encoding =
"UTF-8" ,blocking=F)
txtLines <- readLines(f)
nouns <- sapply(txtLines,extractNoun,USE.NAMES=F)
head(unlist(nouns), 20)

write(unlist(nouns),"sub2016-10-16~2016-10-30.txt")
revised <- read.table("sub2016-10-16~2016-10-30.txt")
nrow(revised)
wordcount <- table(revised)
length(wordcount)
head(sort(wordcount, decreasing=T),10)

library(RColorBrewer)
pal <- brewer.pal(9,"Set1")
wordcloud(names(wordcount),freq=wordcount,scale=c(5,1),rot.
per=0.25,min.freq=1,random.order=F,random.color=T,colors=
pal)
```

그림 12. 특정 프로야구단 워드클라우드 코드
Fig. 12. Specific professional baseball wordcloud code

그림 12는 특정 프로야구단의 워드클라우드 그래프와 관련된 코드의 모습을 보여준다. KoNLP 형태소 분석과 표현하기 위한 wordcloud 패키지를 설치하고, f 변수에 수집된 기사를 삽입한 후 한줄씩 텍스트라인으로 파일을 읽어 들인다. 원본 파일을 복제본으로 두고, RColorBrewer 패키지를 설치하여 적용한 후 wordcloud함수를 이용하여 그래프로 출력한다.

그림 13. 특정 프로야구단 워드클라우드 그래프 결과
Fig. 13. Specific professional baseball wordcloud graph results

그림 13은 그림 12의 코드를 이용하여 특정 프로야구단의 기사를 워드클라우드 그래프로 출력한 결과를 보여준다. 특정 프로야구단의 구장 이름과 한국시리즈와 관련되어 있는 것을 확인할 수 있으며, 그 외 관련된 주요 키워드를 볼 수 있다.

2. 프로야구 성적과 관중 수 분석

```
mat = matrix(nrow=12, ncol=11)
mat <- read.csv("count.csv", head=TRUE)

for(j in 2:12){
  for(i in 1:11){
    mat[i, j-1] = (count[i,j]-count[i+1,j])/count[i+1,j] * 100
  }
}
Write1.csv(mat, "mat.csv", row.names=Ture)

mat2 = matrix(nrow=12, ncol=11)
mat <- read.csv("coun2t.csv", head=TRUE)

for(j in 2:12){
  for(i in 1:11){
    mat[i, j-1] = (count[i,j]-count[i+1,j])/count[i+1,j] * 100
  }
}
Write2.csv(mat2, "ma2t.csv", row.names=Ture)
```

그림 14. 년도별 및 월별 관중 증감률 계산 코드
Fig. 14. Year and Monthly Audience Change Calculation Code

그림 14는 구단의 년도 별 관중 수와 월별 관중 수를 R에 로드 하여 메모리에 업로드하고, 년도 별 및 원별로 구단 관중의 증감률을 계산한 코드이다. 계산 후 Write1.csv 에는 년도 별 관중 수의 증감률이 결과로 나오고, Write2.csv 에는 월별 관중 수의 증감률을 결과로 가지고 있는 파일이다.

그림 15는 특정 프로야구단의 년도 별 순위를 그래프로 나타내기 위한 코드를 보여준다. rank_lotte 변수에 특정 프로야구단의 년도 별 순위 데이터를 대입한 후 barplot 함수를 이용한 코드이다.

```
rank_lotte <- read.csv("rank_lotte.csv", head=TRUE)
rank_lotte

barplot(rank_lotte,beside=T,names=c("08","09","10","11","12","
13","14","15","16"),col=c("red","green","yellow","blue","purple","
gray","pink","brown","orange"))

title(main="롯데",col.main="blue",font.main=4)
title(xlab="연도", col.lab="black")
title(ylab="순위", col.lab="black")
```

그림 15. 특정 프로야구단 년도 별 순위 코드
Fig. 15. Specific professional baseball annual ranking code

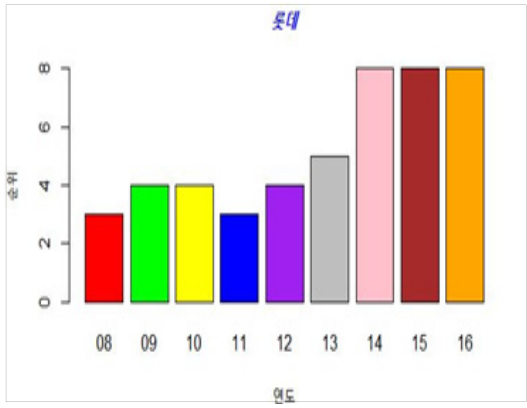


그림 16. 특정 프로야구단 년도 별 순위 그래프
Fig. 16. Specific professional baseball annual ranking graph

그림 16은 코드에 따른 결과 그래프를 의미한다. 특정 프로야구단의 년도 별 순위가 어떻게 구성되었는지 확인할 수 있다. 후반부 연도에 낮은 등수를 기록한 것을 볼 수 있다.

```
v1 <- c(mat[1])
plot(v1,type='o', col='red', ylim=c(-50,100), axes=FALSE,
ann=FALSE)
axis(1,at=1:9, lab=c("08","09","10","11","12","13","14","15","16"))
axis(2,ylim=c(0,10))

title(main="롯데",col.main="blue",font.main=4)
title(xlab="연도", col.lab="black")
title(ylab="관중 증감률", col.lab="black")
```

그림 17. 특정 프로야구단 년도 별 관중 증감률 코드
Fig. 17. Specific professional baseball annual crowd code

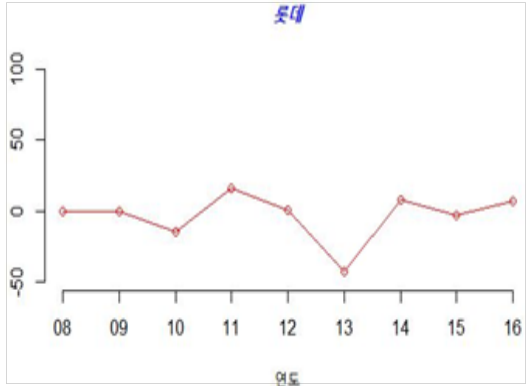


그림 18. 특정 프로야구단 년도 별 관중 증감률 그래프
Fig. 18. Specific professional baseball annual crowd growth graph

그림 17는 특정 프로야구단의 년도 별 관중 증감률을 그래프로 나타내기 위한 코드를 보여준다. 그림 14에서 계산한 증감률의 mat 변수를 불러와 plot 함수 이용하여 그래프로 표현한 코드이다.

그림 18은 코드에 따른 결과 그래프를 의미한다. 특정 프로야구단의 년도 별 순위가 어떻게 구성되었는지 확인할 수 있다. 2010년도, 2013년도에 관중이 급격히 감소한 것을 볼 수 있으며, 2014 ~ 2016년도의 순위가 낮지만, 관중의 증감률은 변화가 없는걸로 보아 즉, 년도 별 순위와 관중 수의 결과와는 상관없는 것을 확인할 수 있다.

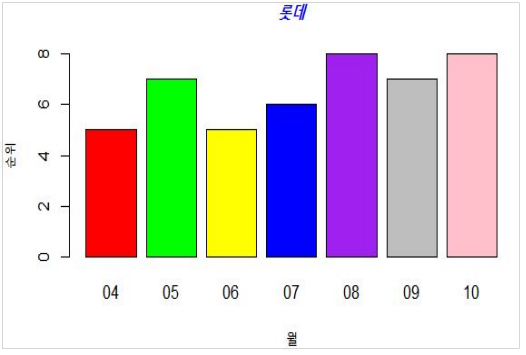


그림 19. 특정 프로야구단 월별 순위 그래프
Fig. 19. Specific professional baseball monthly ranking graph

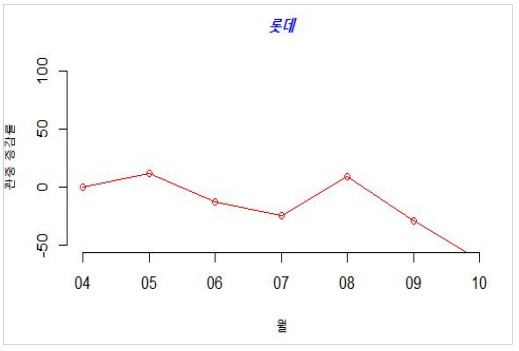


그림 20. 특정 프로야구단 월별 관중 증감률 그래프
Fig. 20. Specific professional baseball monthly crowd growth graph

그림 19와 그림 20은 특정 프로야구단의 월별 순위 그래프와 관중 증감률 그래프를 보여준다. 세부 코드는 년도 별 그래프를 그리기 위한 코드에서 데이터만 변경하였기 때문에 코드는 생략하였다.

특정 프로야구단의 월별 순위와 월별 관중의 증감률을 확인하면, 5월, 8월, 9월, 10월에 낮은 등수를 기록한 것을 볼 수 있고, 관중 수 증감률은 6월, 7월, 9월, 10월에 많이 감소 한 것을 확인할 수 있다.

3. 프로야구 타자 타율과 출루율 성적 분석

```
KBO16 <- read.csv("16batting.cvs", header=T)
KBO17 <- read.csv("17batting.cvs", header= T)

f1 <- data.frame(c=factor(sample(rep(KBO16$batavg))))
t1 <- table(f1$c)

barplot((sort(t1,decreasing = TRUE)[1:10]),xlab="선수 이름",
ylab="타율", col="lightgreen", border="white" , main = "TOP")
barplot((sort(t1,decreasing = TRUE)[1:10]),xlab="선수 이름",
ylab="타율", col="lightgreen", border="white" , main = "TOP")
```

그림 21. 2016, 2017년도 KBO 타자 타율 성적 코드
Fig. 21. KBO hitter's batting average code for 2016, 2017 years

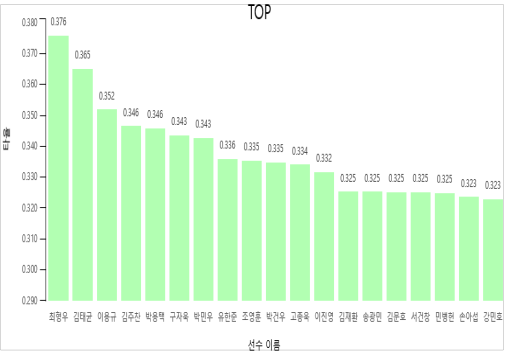


그림 22. 2016년도 KBO 타자 타율 성적 그래프
Fig. 22. KBO hitter's batting average graph for 2016 years

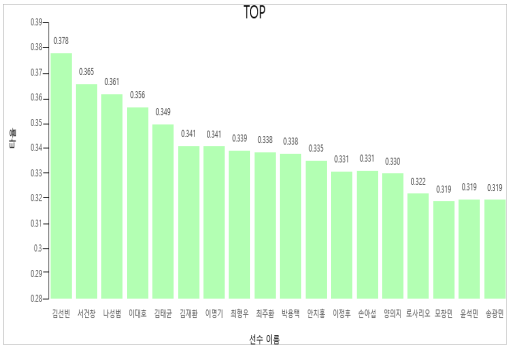


그림 23. 2017년도 KBO 타자 타율 성적 그래프
Fig. 23. KBO hitter's batting average graph for 2017 years

그림 21은 2016년도, 2017년도의 타자 타율 성적을 분석하기 위해 사용한 코드의 모습을 보여주며, KBO16, KBO17 변수에 데이터를 읽고, t1 변수에 데이터프레임을 구성한 후 barplot 함수로 그래프로 표현하였다.

그림 22와 그림 23은 2016년도, 2017년도 KBO 타자의 타율 성적을 그래프로 보여준다. 2016년도에는 최형우, 김태균 순으로 타율 성적이 좋고, 2017년도에는 김선빈, 서건창 순으로 타율 성적이 좋았다.

```
KBO16_2 <- read.csv("16obp.csv", header=T)
KBO17_2 <- read.csv("17obp.csv", header= T)

f2<- data.frame(c=factor(sample(rep(KBO16$OBP)))
t2 <- table(f1$c)

barplot((sort(t2,decreasing = TRUE)[1:10]),xlab="선수 이름",
ylab="OBP", col="lightgreen", border="white" , main = "OBP")

barplot((sort(t2,decreasing = TRUE)[1:10]),xlab="선수 이름",
ylab="OBP", col="lightgreen", border="white" , main = "OBP")
```

그림 24. 2016, 2017년도 KBO 타자 출루율(OBP) 성적 코드
Fig. 24. KBO hitter's OBP code for 2016, 2017 years

그림 25. 2016년도 KBO 타자 출루율(OBP) 성적 그래프
Fig. 25. KBO hitter's OBP graph for 2016 years

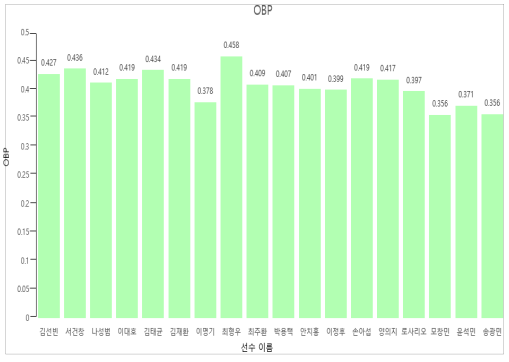


그림 26. 2017년도 KBO 타자 출루율(OBP) 성적 그래프
Fig. 26. KBO hitter's OBP graph for 2017 years

그림 25와 그림 26은 2016년도, 2017년도의 타자 출루율 성적을 분석하기 위해 사용한 코드의 모습을 보여주며, 그림 21과 코드가 유사하기에 코드 설명은 생략한다.

그림 25와 그림 27은 2016년도, 2017년도 KBO 타자의 출루율 성적을 그래프로 보여준다. 2016년도에는 김태균, 최형우 순으로 타율 성적이 좋고, 2017년도에는 최형우, 서건창 순으로 타율 성적이 좋았다.

V. 결 론

본 논문에서는 프로야구에서 생성되는 데이터를 이용하여 특정 프로야구단과 관련된 기사 분석, 순위와 관중률 분석, 타자의 타율 및 출루율 분석 연구를 수행하였다.

다양한 데이터가 존재하지만, 직접 데이터를 수집하는 방법을 사용하기 위해 크롤링 및 셀레늄 기법을 사용하였고, 다른 데이터는 외부 포탈사이트에 존재하는 데이터를 사용하였다.

데이터 처리에는 Hive-QL을 지원하는 하이브를 사용하여 정제 하였으며, 직접 코딩이 어려운 부분은 액셀 프로그램을 이용하여 데이터를 정형화 하였다. 분석 결과로는 특정 프로야구단의 기사 빈도 횟수와 어떤 단어가 가장 많이 이슈되고 있는지 알아보았으며, 야구단의 순위는 관중률의 증가 및 감소에는 영향이 없는 것으로 확인되었다.

향후 연구과제로는 트렌드를 정확하게 분석하기 위해서 SNS에서 사람들이 작성하는 글과 댓글을 같이 크롤링하여 감정 분석을 할 예정이며, 타율과 출루율의 상관관계를 추가로 분석할 예정이다.

References

[1] Crawling & Selenium Tutorial Reference Site, http://www.marinamele.com/selenium-tutorial-web-scraping-with-selenium-and-python

[2] KBO Ranking & Crowd Growth Reference Site, https://www.koreabaseball.com/Record/Player/HitterBasic

[3] Betman Batting Average & OBP Reference Site, http://www.betman.co.kr/sportsMain.so?method=inquireMain&item=BS&fromCode=top

[4] Jeong-Joon Kim, Kwang-Jin Kwak, Don-Hee Lee, Yong-Soo Lee, "Study of Trust Bigdata Platform," Journal of The Institute of Internet, Broadcasting and Communication, Vol. 16, No. 6, pp. 225-230, Dec, 2016.
DOI: https://doi.org/10.7236/JIIBC.2016.16.6.225

[5] Dong-Jin Shin, Jong-Min Eun, Ho-Geun Lee, Myoung Gyun Lee, Jeong-Min Park, Jeong-Joon Kim, "Big

Data-based Log Collection and Analysis in IoT Environments," Journal of Engineering and Applied Sciences, Vol. 13, No. 5, pp. 1064-1072, May 2018.
DOI: http://dx.doi.org/10.3923/jeasci.2018.1064.1072

[6] Dong-Jin Shin, Ji-Hun Park, Ju-Ho Kim, Kwang-Jin Kwak, Jeong-Min Park, Jeong-Joon Kim, "Big Data-based Sensor Data Processing and Analysis for IoT Environment," Journal of The Institute of Internet, Broadcasting and Communication, Vol. 19, No. 1, pp. 117-126, Feb, 2019.
DOI: https://doi.org/10.7236/JIIBC.2019.19.1.117

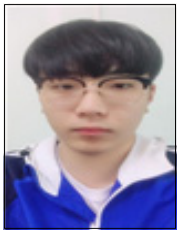
[7] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, Raghotham Murthy, "Hive - A Warehousing Solution Over a Map-Reduce Framework", Proceedings of the VLDB Endowment, Vol. 2, No. 2, pp. 1626-1629, 2009.

저 자 소 개

신 동 진(준회원)

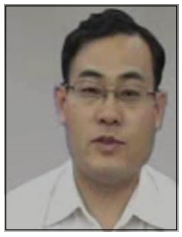
• Dong-Jin Shin received BS in Department of Computer Science and MS in Department of Smart Manufacturing Engineering at the Korea Polytechnic University in 2018 and 2020. He is currently studying PhD in Department of Computer Science at AnYang University. His research interests include Big Data, Internet of Things (IoT), Network&System security.

황 승 연(준회원)



• Seung-Yeon Hwang is received his BS in Department of Computer Science at Korea Polytechnic University in 2019. He is currently studying MS in Department of Computer Science at Anyang University. His research interests include Database System, Big Data, Data Analysis, Machine Learning, etc.

이 돈 희(정회원)



• Don Hee Lee received his M.S degree in Computer Engineering from Yonsei University, Korea, in 2005, and a Ph.D. in Computer and Information Communication Engineering from Konkuk University, Korea, in 2016. He has been working for SK holdings Ltd, Bundang, Korea, from 2002 to present. His research interests include databases, big data, spatio-temporal index in wireless communications, location-based services, and information system audit.

문 진 용(정회원)

• Jin Yong Moon received his MS in Computer Science at Konkuk University in 1998. Then he received PhD from Suwon University in 2001. He is currently a professor in the department of Visual Broadcasting Media at Gangdong University. His research interests include Database Systems, Web Science, Geographic Information Systems (GIS) and Multimedia Systems, etc.

김 정 준(정회원)

• Jeong-Joon Kim received his BS and MS in Computer Science at Konkuk University in 2003 and 2005, respectively. In 2010, he received his PhD in at Konkuk University. He is currently a professor at the department of ICT Convergence Engineering at Anyang University, His research interests include Database Systems, Big Data, Semantic Web, Geographic Information Systems (GIS) and Ubiquitous Sensor Network (USN), etc.