제36회 한국정보처리학회 추계학술발표대회 논문집 제18권 제2호 (2011. 11)

야구선수 기록과 텍스트, 오피니언 마이닝을 이용한 야구선수 특성 분석

*이태훈, 김이준, 임지연, 김응모
성균관대학교 정보통신공학부
e-mail : cmt2059@skku.edu

*Tae-Hun Lee, iee-joon Kim, ji-yeon Lim Ung-mo Kim

School of Information and Communication Engineering

Sungkyunkwan University

요 약

출범 30주년을 맞은 국내 프로야구가 600만 관중을 넘어섰다. 국내 프로야구의 인기가 그만큼 높으며 식을 줄을 모른다.[7] 하지만 많은 선수와 복잡한 규칙은 새로운 야구팬을 주춤하게 만든다. 본 논문을 통하여 필자는 새로운 방식의 야구선수 분석 방법을 제안하고자 한다. 새로운 방식의 야구선수 분석 방법은 쉽고 간단하게 야구선수들의 특성을 분석하여 표현할 것이며, 야구를 좋아하고자 하는 사람들은 조금 더 쉽게 다가갈 수 있을 것이다.

1. 서론

야구는 9명 또는 10명으로 구성된 두 팀이 방망이와 공을 사용하여 겨루는 구기 종목의 하나이다.[5] 역사가 100년 이상된 오래된 구기종목이며, 우리나라의 프로야구는 81년 출범이후 올해로 출범 30주년을 맞았다. 야구관중도 평균 1만명이 넘었으며 최근에는 600만 관중돌파, 천만 관중을 바라보고 있는 스포츠이다.[6][7] 그만큼 인기 있는 스포츠라고 볼 수 있다. 하지만 특색이 다른 8개의 구단, 각 팀당 80명이 넘는 선수, 수많은 선수 기록 등이 야구를 좋아하려고 하는 사람들에게 큰 장벽으로 느껴질 것이다.

뉴스기사에서 각 구단 및 선수에 대하여 전망을 내놓아 보다 이해하기 쉽도록 정보를 제공하고 있지만 하지만 야구에 관한 지식이 부족한 사람들은 이마저도 이해하기 힘들기 때문에 뉴스기사와 기록표를 분석해서 좀 더 가시성을 높이고 쉽게 이해할 수 있도록 정리 해주는 시스템을 제안하고자 한다.

본 논문으로 현재 야구를 좋아하거나 또는 앞으로 야구를 좋아할 많은 사람들을 위하여 야구를 좋아하게 되는데 장벽이 되는 정보들을 쉽고 간단하게 보여주는 시스템을 제안할 것이다.

그러기 위해서 현재 활발히 연구가 진행되고 있는 오피니언 마이닝과 텍스트 마이닝의 장점을 각각 이용하여 새로운 평가시스템을 제안하며 각 선수의 평가를 정수화 시켜 등급을 매길 것이다.

2. 관련연구

2.1. 오피니언 마이닝 (Opinion Mining) [1]

오피니언 마이닝은 특정 주제에 대한 글쓴이나화자의 태도를 찾아내는 것을 말한다. 웹 2.0의 발전으로 사용자들이 더 많은 의견을 표현함에 따라 오피니언 마이닝이 더욱 주목 받고 있다. 오피니언 마이닝의 주요 주제를 언어학적자원을 개발하고 발전시키는 것, 의견의 의미극성을 판단하는 등 의견을 요약하는 것, 텍스트로부터 의견이 표현된 부분을 추출하는 것으로 나눈다. 오피니언 마이닝 초기에는 자연어 처리 기법을 많이 이용하였다. 하지만 이러한 방식은 실제 적용에 있어 많은 한계점들을 보여주었다. 이러한 한계점을 극복하기위해 최근의 연구들은 통계적 분석을 기존의 자연어처리 기법과 함께 사용하여 좋은 성과를 거두고 있다.

2.2. 어휘의 의미 극성 판단 [1]

어휘의 의미 극성 정보는 오피니언 마이닝 분야에서 중요한 언어학적 자원 중 하나이다. 어휘의 의미 극성을 전산언어학적인 방법으로 접근하려는 시도는 90년대 말에 시작되었다. 이는 'and'로 연결된 형용사들은 비슷한 극성을 가질 것이라는 가정을 바탕으로 문제를 해결하고자 하였다. Turney 등은 확률론에 기반을 둔 PMI(Point-wise Mutual Information)를 사용하여 어휘의 의미 극성을 판별하였다. PMI는 비교적 간단함에도 불구하고 좋은 결과를 얻을 수 있다. PMI에 대한 자세한 소개는 [1]에서 하고자 한다. 우리말의 경우, 어휘의 체계를 구축하고자 하는 연구는 많이 진행되었지만, 우리말 어휘의 의미 극성을

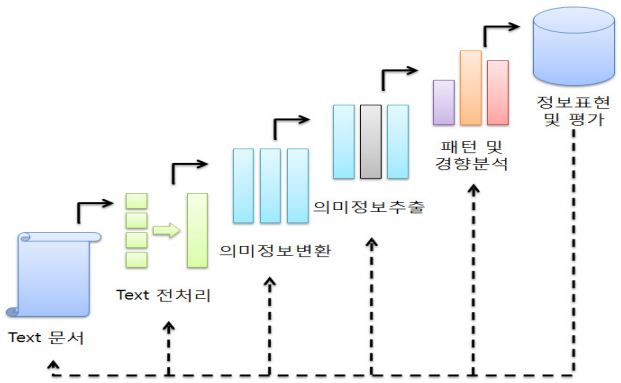
판단하는 연구는 아직 미흡한 실정이다.

우리말 어휘의 극성을 판단하기 위해 고려할 수 있는 방법은 앞에서 제시된 것처럼 다양하다. 최근의 경향은 어휘망을 이용하는 것이다. 하지만 공개 API 등을 제공하여 접근 및 사용이 용이한 우리말 어휘 망이 없고, 실제 상품평에서 자주 쓰이는 '이쁘다', '이뿌다'와 같은 비표준어는 어휘 망에 포함되지 않아 비표준어의 의미 극성을 판단할 수 없다는 단점이 있다. PMI에 기반을 둔 방법은 웹 검색 서비스나 시스템을 사용하여 쉽게 적용할 수 있다는 장점이 있고 성능 또한 우수하다.

2.3 텍스트 마이닝 (Text Mining) [2]

대량의 정보를 효과적으로 다룰 수 있는 방법에 대한 연구는 이미 활발히 진행되고 있다. DB 에 저장된 자료와 같이 정형화된 데이터로부터 정보를 추출, 가공하는 데이타마이닝 (Data Mining) 은 이미 실용성을 갖추고 많은 분야에서 널리 활용되고 있다. 그러나 디지털 정보의 대부분은 비정형 데이터로서, Text Mining 은 이러한 비/반정형 데이터에 대하여 자연어처리 (Natural Language Processing) 기술과 문서처리 기술을 적용하여 유용한 정보를 추출, 가공하는 것을 목적으로 하는 기술이다. 문서요약 (summarization), 특성추출 (feature extraction) 등이 text mining 의 핵심 연구분야며 그 응용 분야는 매우 다양하다.

Data mining 관점에서 문서로부터 구조화된 정보를 추출하여 database 화 시키거나 규칙을 찾아내는 것은 가장 일반적인 응용이며, 사용자가 Web 상에서 문서를 찾는 것을 도와주거나 사용자 profile 의 생성 및 분석, 문서에 쓰인 자연언어 식별, 대량 DB에서 문서의 분류 및 군집화, 문서분류 (Text Categorization) 정보를 이용한 문서 재해석, 신문/논문/보고서 요약, 문서 번역, 시계열 (time series) 정보의 획득을 통한 시장 및 위험도 분석, 문서 색인, 문서 여과 (filtering) 및 추천 (recommendation), 대표적 키워드나 토픽 (topic) 의 추출, 질의응답 시스템 (Question Answering System), 대규모 문서에서의 탐색 등이 가장 대표적인 응용분야라 할 수 있다.



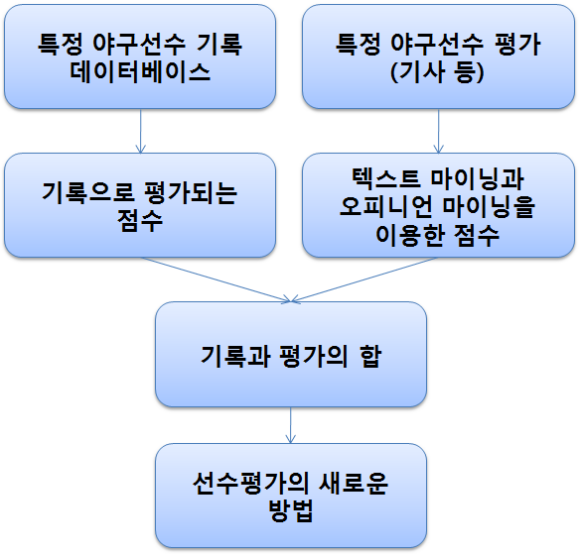
[그림 1]

3. 제안 시스템

3.1 전체 시스템

본 논문의 제안시스템은 야구선수의 기록과 야구선수들을 평가한 기사들의 두 가지 정보의 점수가 합을 이루어 새로운 선수 평가 방법을 제안한다. 먼저 야구선수가 가지고 있는 기록을 이용하여 점수로 환산한다. 3.2에서 추가로 설명할 예정이지만 경기당 점수 시스템을 이용할 것이다. 두 번째로는 뉴스기사를 통한 야구선수의 평가 및 전망을 텍스트마이닝 기법에 오피니언마이닝의 극성 판단을 통하여 관련 점수로 나타낼 것이다. 이 두 가지 점수를 합하여 새로운 선수 평가의 방법을 도출해 내는 것이다.

시스템의 전체흐름은 다음 [그림2]와 같다.



[그림 2]

3.2 야구선수 기록 분석

	연도	경기	타수	득점	안타	2타	3타	홈런	타점
최형우	2009	113	415	70	118	24	0	23	83
이대호	2009	133	478	73	140	28	1	28	100

	도루	볼넷	삼진	병살	실책	총 점수	경기당 점수
최형우	1	53	60	12	3	297	2.63
이대호	0	51	65	13	15	328	2.47

[그림 3]

야구선수는 한 시즌동안 많은 기록을 낸다. 그 중 주요 기록들을 가지고 점수로 환산 할 것이며 [그림 3]에서와 같이 총 13개의 기록을 가지고 점수를 낼 것이다. 타수, 득점, 안타, 2타, 3타, 홈런, 타점, 도루, 볼넷은 선수개인의 기량으로 낸 좋은 성적이므로 (+)의 성적이라고 볼 수 있고 삼진, 병살, 실책은 선수개인의 좋지 않은 성적이므로 (-)의 성적이라고 볼 수 있다. 따라서 '총 점수'는 각 점수를 모두 더한 성적이고 '경기당 점수'는 총 점수에서 경기수를 나눠준 값이 된다.

다음의 수식으로 간단히 표현해 보았다.

$$\text{총점수} = \text{타수} + \text{득점} + \text{안타} + 2\text{타} + 3\text{타} + \text{홈런} + \text{타점} + \text{도루} + \text{볼넷} - (\text{삼진} + \text{병살} + \text{실책})$$

$$\text{경기당 점수} = \frac{\text{총점수}}{\text{경기수}}$$

3.3 야구선수 뉴스기사 분석

특정 야구선수에 대한 전망이나 기대에 대한 평가를 내는 기사는 많다. 본 논문에서 제안하는 시스템은 단순히 텍스트 마이닝을 통하여 기사를 분석할 수도 있지만 여기서 멈추지 않고 오피니언 마이닝의 극성 판단을 통하여 관련 기사를 하나의 점수로 표현 하고자 한다. 하나의 점수로 표현한다면 3.2에서 분석한 점수와 합산하여 새로운 점수를 나타낼 수 있기 때문에 위와 같은 방법을 선택하게 되었다.

삼성은 거포로 자리매김할 가능성을 보여준 박석민과 최형우에 기대를 걸기로 했다.

[그림 4]

위의 [그림4] 에서 보면 거포나 기대라는 단어는 최형우 선수에게 좋은 점으로 작용할 수 있기 때문에 (+) 점수를 줄 수 있을 것이다. 하지만 부진, 2군 등의 단어는 선수에게 좋지 못한 (-)의 점수가 될 것이다. 각 기사는 수많은 단어를 포함하고 있을 것이다. 각각의 단어가 의미하고 있는 바를 극성 판단을 통하여 분석할 것이다.

긍정/부정	단어	횟수	점수
긍정	기대	5회	+5
	거포	1회	+1
	상당히	1회	+1
	뛰어난	2회	+2
부정	실망	1회	-1
	부상	3회	-3
합계			+5

[그림 5]

다음은 “이영욱·박석민·최형우 삼성 타선의 삼두마차”[3]의 뉴스기사를 오피니언 마이닝의 극성 판단을 통하여 분석 해본 결과이다. 긍정과 부정단어가 각각의 [그림 5]와 같이 검색되었으며 긍정의 단어는 +1점씩 부정의 단어는 -1점씩 부여하여 이 기사의 총 점수는 +5점으로 분석되었다. [그림 5]는 이를 정리한 표이다.

긍정/부정	단어	횟수	점수
긍정	기대	1회	+1
	활약	1회	+1
	자신감	1회	+1
	만회	1회	+1
부정	못했	2회	-2
	부족	2회	-2
합계			0

[그림 6]

다음은 “2010 ‘빅맨’ 이대호의 도전… 3할·30홈런·골든글러브”[4] 의 기사를 분석 한 것이다. 긍정 및 부정의 단어가 위의 [그림 6]과 같이 나타났으며 오피니언 마이닝의 극성판단을 통한 분석을 한 결과 0 점으로 분석 되었다. [그림 6]에서 분석결과를 확인 할 수 있다.

3.4 선수평가의 새로운 방법

3.2에서의 기록점수와 3.3에서의 분석된 자료의 점수를 합산한 결과 최형우 선수의 점수는 기록점수 2.63과 뉴스분석점수 5점을 얻어 총 7.63점의 결과를 얻을 수 있다. 반면 이대호 선수는 뉴스기사 분석에서 0점을 얻었으므로 기록점수 2.47점의 점수가 그대로 유지된다. 단순히 점수를 비교하였을 때에는 내년시즌에 최형우 선수가 조금 더 나은 기록을 낼 것이라고 예측할 수 있을 것이다. 비록 본 논문의 실험은 각 선수 당 하나의 뉴스기사의 분석하지 않았지만 다수의 뉴스기사를 분석하여 점수로 낼 수 있다면 좀 더 신뢰도 높은 결과를 얻을 수 있을 것이다.

4. 결론 및 향후 연구 방향

본 논문에서는 야구선수의 기록과 뉴스기사를 통한 평가를 통하여 야구선수의 새로운 평가방법을 제시하였다. 이전에 없던 시스템이니 만큼 부족한 부분도 많을 것이다. 하지만 이러한 제안 시스템을 바탕으로 보다 나은 시스템이 나오기를 바라며, 추가적으로 야구선수뿐만 아니라 야구구단에 대한 평가 시스템을 통하여 야구를 좋아하는 많은 사람들에게 보다 나은 정보를 제공할 수 있었으면 하는 바람이다.

참고문헌

[1] 송상일 이동주 이상구 “PMI를 이용한 우리말 어휘의 의미 극성 판단” 한국컴퓨터종합학술대회 논문집 Vol37, No1(C).

[2] 안태성 서형국 이경일 “텍스트마이닝 기반 고정밀 검색시스템” 정보처리학회지 제11권 제2호 2004.3.

[3] 채정민 “이영욱·박석민·최형우 삼성 타선의 삼두마차” http://www.imaeil.com/sub_news/sub_news_view.php?news_id=52321&yy=2009

[4] 이원만 “2010 ‘빅맨’ 이대호의 도전… 3할·30홈런·골든

글 러 브 ”
http://sportsworldi.segye.com/Articles/Sports/BaseBall/Article.asp?aid=20091214004388&subctg1=05&subctg2=00
[5] 위키백과 “야구”
http://ko.wikipedia.org/wiki/%EC%95%BC%EA%B5%AC
[6] 위키백과 “한국프로야구”
http://ko.wikipedia.org/wiki/%ED%95%9C%EA%B5%AD%ED%94%84%EB%A1%9C%EC%95%BC%EA%B5%AC
[7] 최민규 “프로야구 600만 관중 … 한국인 삶 접수하다”
http://joongang.joinsmsn.com/article/aid/2011/09/14/5832828.html?cloc=olink|article|default