Journal of the Korean Data & Information Science Society
2015, **26**(3), 619–630

http://dx.doi.org/10.7465/jkdi.2015.26.3.619
한국데이터정보과학회지

사회네트워크분석과 텍스트마이닝을 이용한 배구 경기력 분석[†]

강병욱[1] · 허만규[2] · 최승배[3]

[1](주)엠소프트테크놀러지 · [2]동의대학교 분자생물학과 · [3]동의대학교 데이터정보학과

접수 2015년 2월 25일, 수정 2015년 3월 12일, 게재확정 2015년 5월 18일

요 약

본 연구의 목적은 '사회네트워크분석'과 '텍스트마이닝'을 이용하여 국내 남자프로배구 구단의 공격, 패스 패턴을 찾아내고, 배구경기력과 관련된 핵심 키워드 추출하여 경기력을 평가하여 향후 구단의 경기 전력을 수립하는데 기초자료로 활용하는데 있다. 본 연구에서는 '사회네트워크분석'을 통해 도출된 그룹변수들을 '텍스트마이닝' 기법의 결과인 경기의 '승패'에 차이를 검정하기 위해 '0' 그룹 (6명)과 '1' 그룹 (11명)으로 재구성하였다. 연구의 결과로서 '사회네트워크분석'의 연결중심성과 중개중심성의 순위로 판단하면, '0' 그룹 보다 '1' 그룹이 우수한 경기력을 보였다. '사회네트워크분석'에 의해서 재구성된 '0' 그룹과 '1' 그룹에 따라서 '텍스트마이닝'에 의해서 생성된 '승패' 그룹에 대한 유의성 검정 결과 유의한 차이가 있는 것으로 나타났다 (p값: 0.001). '그룹별' 클러스터링 결과, '0' 그룹의 경우 'D' 선수와 'E' 선수가 '세트' 플레이를 통하여 정확하게 득점한다고 할 수 있다. '1' 그룹의 경우 'K' 선수가 '디그'에 의해서 '공격'을 하는 경우 실패하는 경우가 많고, 'C' 선수와 'P' 선수는 '세트' 정확한 플레이를 한 것으로 나타났다.

주요용어: 문자 군집분석, 사회네트워크분석, 웹 사이언스, 중심성 측도, 텍스트마이닝 기법.

1. 서론

인터넷의 발달로 인해 인터넷과 디지털 기기만 있으면 언제, 어디서, 누구든지 정보를 제공하고 공유할 수 있게 됨에 따라 많은 분야에서 다량의 데이터가 발생되고 있다. 기업들은 보다 많은 이익의 창출을 위해, 관공서 등 각 종 기관들은 국민들에게 보다 나은 서비스를 제공하기 위해 빅 데이터 속에서 유용한 정보를 찾고자 노력하고 있다. 이러한 경향으로 인해 빅데이터를 분석하기 위한 다양한 방법론 및 기법들이 개발되고 있다. 정형화된 데이터를 분석하는 대표적인 방법론으로 데이터마이닝이 있고, 비정형화 데이터를 분석하는 사회네트워크분석 (social network analysis)과 텍스트마이닝 (text mining) 등이 있다.

본 연구에서 관심 있는 데이터는 비정형화된 텍스트 데이터로서 사회네트워크 서비스뿐만 아니라 이메일과 메시지 등 많은 방법을 통해서 얻어질 수 있다. 이 때문에 많은 주목을 받고 있는 분야가 '사회미디어 분석'이고, '사회네트워크분석'과 '텍스트마이닝'이 대표적인 기술이다. '사회네트워크분석'은 사회네트워크 서비스를 통하여 연결된 또는 연결되지 않은 개인 (또는 기관)으로 구성된 사회적 구조 (점과 선으로 구성된 망)에 대한 사회과학적·통계학적 분석방법이다 (Huh, 2010). '사회네트워크

[†] 이 논문은 공동저자 강병욱의 석사학위 논문을 재구성한 것임.

[1] (150-010) 서울특별시 영등포구 여의도동 14-24번지, (주)엠소프트테크놀러지, 대리.
[2] (614-714) 부산광역시 부산진구 가야동 산 24번지, 동의대학교 분자생물학과, 교수.
[3] 교신저자: (614-714) 부산광역시 부산진구 가야동 산 24번지, 동의대학교 데이터정보학과, 교수.
E-mail: csb4851@deu.ac.kr

분석'을 통해 보이지 않는 인적 네트워크 관계의 유형과 개인 (또는 기관)의 역할 등을 파악하고, 관계의 유형에 따라 어떠한 효과가 나타나는 지를 알아볼 수 있다. '텍스트마이닝'은 비정형 데이터를 대상으로 텍스트 간의 암묵적인 정보를 추출하는 과정이다. 비정형화된 데이터를 활용하여 유용한 정보를 어떻게 도출할 것인가에 대한 많은 연구가 진행되고 있으며, '사회네트워크분석'과 '텍스트마이닝' 등과 같은 다양한 기법들이 활용되고 있다.

'사회네트워크분석'과 관련된 연구로 Choi 등 (2011)은 축구팀 내에서 각 선수들의 역할에 대한 수행 실태를 평가하고, 향후 경기에서 팀의 경기 전략을 수립하는데 기초자료로 활용하기 위해 '사회네트워크분석'을 수행하였다. 그리고 Choi (2013)는 D대학교의 사례를 중심으로 '사회네트워크분석'을 통하여 학과 간의 전출과 전입의 특성을 분석하였다. Cho (2012)는 '사회네트워크분석'을 이용하여 학과별 복수전공자들의 유입과 유출에 대한 특성을 분석하였고, Cho (2014)는 개인특성 변수들이 취업여부에 미치는 효과를 분석하기 위해서 '사회네트워크분석'의 연결망 구조로 시각화하였다. Kang (2010)은 사회연결망 연결중심성 (degree centrality)을 활용한 신규고객 상품추천방법의 추천 정확성 향상 방안에 대한 연구를 수행하였다. 그리고 Won과 Choi (2014)는 통계학 관련 국내 학술지들이 어떠한 수준의 영향력을 가지고 있는지를 다양한 KCI 인용지수를 이용하여 비교하였고, 상호인용 빈도를 활용하여 학술지 간 관련성을 알아보기 위해 '사회네트워크분석' 관점에서 조망하였다. 가장 최근의 연구로서 Kim과 Park (2015)는 토픽 모형 및 '사회네트워크분석'을 이용하여 한국데이터정보과학회지의 영문초록을 분석하였다. 연구뿐만 아니라 Kim (2007), Huh (2010), Son (2010) 등이 집필한 '사회네트워크분석'과 관련된 서적들이 있다.

'텍스트마이닝'과 관련된 연구로 Park와 Lee (2009)는 복합적 텍스트 분석을 이용한 포털 댓글에 관한 연구를 수행하였다. Oh 등 (2010)은 '텍스트마이닝'을 이용한 Claim Data 분석에 관한 연구를 하였고, Oh와 Jin (2012)은 '텍스트마이닝'을 이용한 쇼핑몰 구매후기 분석에 관한 연구를 수행하였다. Jung (2010)은 '텍스트마이닝'과 '네트워크분석'을 활용한 미래예측 방법 연구라는 주제로 정성적인 방법에 의해 작성된 제 3회 과학기술예측조사 (대조군)의 기술들과 정량적인 방법 (텍스트마이닝)에 의한 과학기술 미래비전 (실험군)의 기술들을 비교 분석하였다. 이와 같이 다양한 분야에서 얻어진 방대한 비정형 데이터는 다방면으로 적용이 가능하다는 것을 알 수 있다.

본 연구의 목적은 2011∼2012 V-리그 국내 남자 프로배구 경기의 문자중계 데이터인 텍스트 데이터를 '사회네트워크분석'과 '텍스트마이닝' 기법을 이용하여 공격, 패스 등의 패턴을 찾아내고, 배구 경기력과 관련된 핵심 키워드를 추출하여 이를 토대로 경기력을 평가하는데 있다. 그리고 이러한 분석 결과를 토대로 향후 국내 남자 프로배구 구단의 경기 전력을 수립하는데 기초자료로 활용하는데 있다. 본 연구의 내용을 통한 기초자료를 활용하여 배구 구단은 각 선수들에 대한 연봉 협상 및 경기 전력을 수립할 수 있을 것으로 기대한다.

본 연구는 다음과 같이 구성되어 있다. 2절에서는 '사회네트워크분석'과 '텍스트마이닝'에 대해서 간략히 소개하고, 3절에서는 본 연구에서 수행한 분석방법에 대해서 일련의 과정 (자료 수집, 분석, 결과 등)을 분석 흐름도를 통해 소개한다. 그리고 4절에서는 '사회네트워크분석'과 '텍스트마이닝'을 이용한 분석결과에 대해 보여주고, 마지막 절에서 본 연구의 한계점 및 향후 연구과제로 결론을 맺는다.

2. 사회네트워크분석과 텍스트마이닝

2.1. 사회네트워크분석

사회네트워크는 웹 사이언스 (web science)의 연구 분야 중 하나로, 웹상에서 개인 또는 집단이 하나의 개체가 되어 각 개체들 간의 상호의존적인 관계에 의해서 만들어지는 사회적 관계 구조를 말한다. 모든 개체들은 네트워크 안에 존재하는 개별적인 주체들로서 사람, 회사, 기관, 홈페이지 등이 될 수 있다.

사회네트워크에서 방향성은 누가 누구에게 정보를 제공하였는지에 대해서 알 수 있지만 비방향성은 이러한 방향성을 알 수 없는 경우이다. 이러한 사회네트워크를 통하여 각 개체들 간의 어떤 영향을 미칠 것인지, 전체 네트워크와 어떤 연관성을 가지며 어떻게 영향을 주고받는지 등을 분석하는 것이 '사회네트워크분석'이다. 즉, '사회네트워크분석'은 수많은 개체들과 그 개체들 사이의 관계망 내에서 어떤 개체들이 얼마나 중요한 역할을 하는지, 하위 그룹이 없는지 또는 연결 관계가 무엇을 의미하는지 등의 개체들 간 사이에 어떤 패턴을 찾기 위한 사회과학적 또는 통계학적 방법론이라고 할 수 있다 (Choi 등, 2011).

'사회네트워크분석'에서 중요한 척도 중에 하나인 중심성 (centrality)은 네트워크 속에서 개체 간에 미치는 영향력을 나타내는 개념으로서 어떤 노드가 중심의 역할을 하는 지, 어떤 노드가 중계 역할을 하는 지 등 각 노드들의 역할을 규명할 수 있는 척도이다. 즉, 중심성은 한 개체가 얼마나 많은 다른 개체들과 연결되어 있는 지로 측정할 수도 있고, 한 개체가 다른 모든 개체들에 도달하려면 몇 단계나 필요한 지로 측정할 수도 있다 (Cho, 2012). 중심성 척도의 종류로 어떤 특정 노드에 연결된 노드들 수를 파악하는 연결중심성, 어떤 특정 노드에 다른 노드들이 얼마나 가깝게 있는 지를 파악하는 근접중심성 (closeness centrality) 그리고 어떤 노드들이 다른 노드들에 대해서 얼마나 중계 역할을 하는 지를 파악하는 매개중심성 (betweenness centrality) 등이 있다 (Choi 등, 2011). 중심성 척도들에 대한 상세한 내용은 참고문헌을 참조해 주기 바란다.

2.2. 텍스트마이닝

'텍스트마이닝'은 비정형화된 대규모 텍스트 데이터로부터 유용한 패턴 및 관계를 발견, 추출하는 과정이다. 텍스트마이닝은 (1) '데이터 수집과정', (2) '용어 추출과정', (3) '정보 추출과정' (4) '정보 분석과정'의 4단계의 절차를 거친다. 먼저 '데이터 수집과정'은 '텍스트마이닝'의 첫 번째 단계로서 비정형 대규모 텍스트 데이터를 수집하는 단계이다.

둘째, '용어 추출과정'은 구문의 패턴이나 단어들의 연관성을 고려하는 연관성 분석에 의해 단어들을 추출하여 분석 대상 용어들의 후보를 만들고, 추출된 후보 용어에 대해서 여러 가지 통계적 방법을 통해 전체를 대표하는 용어들을 추출하는 과정이다. 즉, 첫 번째 단계에서 얻어진 분석대상 데이터를 중심으로 정보를 추출할 수 있도록 데이터를 가공하는 단계로 수집된 자료 (문서)를 기본으로 관련 키워드를 추출한다.

셋째, '정보 추출과정'은 문서 자체를 찾는 것이 아니라 문서 내에서 유용한 정보를 찾는 과정이다. 예를 들어 제품을 소개하는 문서의 경우, 제품명, 제품의 기능, 주의 사항 등과 같이 상세 정보를 포함하는 용어를 추출하기 위한 과정이다.

넷째, '정보 분석과정'은 세 번째 과정에서 얻어진 최종 키워드들에 대해서 '빈도', '분류', '클러스터링', '컨셉 링크' 기법 등을 이용하여 유용한 정보를 도출해 내는 과정이다. '빈도'는 가장 많이 얻어지는 키워드가 무엇인지에 대한 정보를 도출해 내고, '분류'는 앞서 추출된 텍스트의 내용에 따라 문서들을 범주화 시켜주는 과정이다. 즉, 주어진 신문 텍스트 문서가 스포츠 분야인지, 정치 분야인지 등을 텍스트에 따라 분류하는 것을 의미한다. '클러스터링'은 문서에 포함되어 있는 추출된 단어들을 유사도에 따라 여러 개의 텍스트 집단으로 군집화 시켜주는 과정이다. '컨셉 링크'는 어떤 특정한 키워드를 중심으로 또 다른 키워드들 간의 관계를 파악하는 기법이다.

3. 연구 방법

3.1. 분석데이터

본 연구에서 사용될 분석데이터는 국내 남자 프로배구 '2011∼2012 V-리그' 경기의 문자중계 데이터로서 총 6개 구단 중에서 하나의 A구단을 선정하여 '2011∼2012 V-리그' 정규리그에서 총 35경기 (실제 36경기 중 상대팀 기권으로 인한 1승 제외)에 대한 문자중계 데이터를 추출하였다. 문자중계 데이터에서 A구단 내에 소속된 선수들의 이름은 익명을 보장하기 위하여 'A, B, C, ⋯'와 같이 알파벳으로 대체하였고, 구단이 승리한 경기와 패한 경기를 구분하여 분석데이터를 구축하였다.

'사회네트워크분석'을 위한 데이터를 구성하기 위해서 본 연구에서 관심을 두고 있는 용어들을 크게 4가지 ('player', 'action', 'success', 'score')로 구분하여 지정된 4개의 변수에 저장하도록 하였고, 4가지 변수 이외의 용어들을 삭제하여 정제하였다. 즉, 'player'라는 변수에는 A, B와 같이 선수들이 저장된다. 그리고 from, to에 대한 정보를 추출하게 되는데 from이란 배구공의 이동에 있어서 시작점을 의미하고, to는 배구공의 종료점을 의미한다. 예를 들면, 배구공이 A선수에서 B선수 순서로 이동을 했다면, from 값은 A선수가 되고, to 값은 B선수가 된다. 보통 배구경기에서는 3단계 동작 (third step attack)을 기본으로 하고 있는데, 본 연구에서는 2단계 동작의 경우도 from, to 값에 포함시켰다. '텍스트마이닝'을 위한 데이터를 구성하기 위해서 '사회네트워크분석'에서와 마찬가지로 분석에 불필요한 자료들을 삭제하여 데이터를 정제하였다. '텍스트마이닝'을 수행하기 위하여 '공격의 성공 여부' 및 '수비의 성공 여부' 그리고 '득점여부'로 구분하여 원시데이터를 분할, 구축하였다.

3.2. 그룹변수 생성

'사회네트워크분석' 결과, 분류된 9개의 그룹 중 6명 (A, D, E, F, L, N)으로 구성된 그룹과 1명 또는 2명으로 구성된 8개의 그룹으로 얻어졌다. 본 연구에서는 6명으로 구성된 그룹과 나머지 1명 또는 2명으로 구성된 그룹들을 하나의 그룹으로 하여 2개의 그룹으로 그룹변수를 생성하였다. 본 연구에서 두 집단으로 나눈 이유는 (1) '사회네트워크분석' 결과 9개의 그룹으로 분석하기에는 무리가 있다고 판단하였고, (2) 인접한 그룹 간에는 비슷한 성격을 갖는다고 알려져 있어 1명 또는 2명으로 구성된 그룹을 하나의 그룹으로 통합하였다 (SAS Korea, 2010). 그리고 (3) 향후 '텍스트마이닝' 분석과 연계하여 분석을 고려하기 위해서이다.

3.3. 분석 과정

본 연구의 분석과정은 Figure 3.1과 같다. 본 연구의 분석과정은 먼저 (1) 분석하고자 하는 데이터를 수집하여, (2) 연구 목적과 분석 기법에 따라 정제·가공하여 '사회네트워크분석'과 '텍스트마이닝'을 위한 데이터를 구성한다. 그리고 (3) 두 분석 기법을 이용하여 분석을 실시하고, '사회네트워크분석' 결과에서 도출된 정보와 '텍스트마이닝' 기법에 적용시켜 분석한 결과를 이용하여 최종적인 결과를 도출한다.

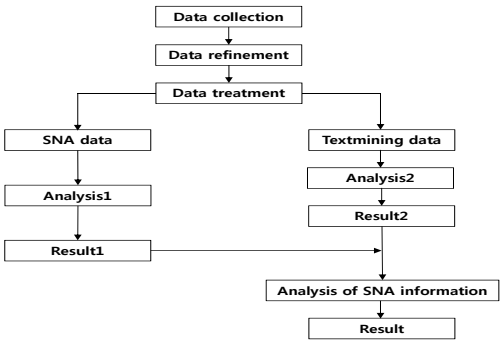

Figure 3.1 Analysis process

(1) 데이터정제

데이터 정제는 분석하기에 용이하게 정제하는 과정으로서 보다 효율적이고 정확한 연구결과 도출을 위해서 매우 중요한 단계이다. '사회네트워크분석'을 위한 원시데이터는 분석에 적합하지 않기 때문에 적절한 데이터 정제가 필요하다. 예를 들면, 동일한 from_id와 to_id인 경우 하나의 개체로 처리하고, 중복된 개체의 수만큼 해당하는 개체에 가중치 (weight)를 부여한다.

'텍스트마이닝'을 위한 데이터 정제는 2.2절에서 기술된 '텍스트마이닝' 4단계 절차에서 두 번째 단계인 '용어 추출과정'과 세 번째 단계인 '정보 추출과정'에 해당한다. 본 연구와 관련하여 예를 들면, 먼저 (1) 문자중계 내용에 포함되어 있는 불필요한 키워드들을 삭제하고, 공격 및 수비의 성공 여부와 득점 여부 등의 후보용어를 추출한다. 그리고 (2) 본 연구에서는 두 번째 단계에서 추출한 후보 키워드들 중에서 분석에 필요한 용어들인 '서브', '리시브', '블로킹', '오픈', '어택', '속공', 'C속공', '스파이크', '시간차' 등을 추출한다.

(2) 데이터 가공

'사회네트워크분석'과 '텍스트마이닝'에서 사용되는 속성들이 다르기 때문에 각각의 기법에 맞도록 데이터를 가공해야 한다. '사회네트워크분석'에 사용되는 데이터는 상호 연결된 네트워크 (from, to) 형식에 맞게 구성되어 져야 된다. Table 3.1은 '사회네트워크분석'을 위한 원시데이터를 가공하는 과정을 보여 주고 있다. 왼쪽에 있는 박스 부분은 원시데이터의 문자 중계 내용이고, 중앙에 있는 박스의 내용은 불필요한 용어들 (경기시작, 성공, 득점 등)을 삭제하여 가공된 결과를 'player', 'action', 'success', 'score'로 분할하여 지정된 변수에 저장되는 것을 보여 주고 있다. 오른쪽 박스의 내용은 from, to에 대한 정보를 추출하는 부분으로서 from이란 배구공의 이동에 있어서 시작점을 뜻하고, to는 배구공의 종료점을 뜻한다.

Table 3.1 Data treatment for social network analysis

Content of a letter relay	Player	Action	Success	Score	From	To
game start					B	A
3.A serve	A	serve	1	0		
6.B blocking success score	B	blocking	1	1	⋮	
3.A serve	A	serve	1	0		
6.B blocking	B	blocking	1	0	B	C
10.C dig fail	C	dig	0	0	C	D
5.D set	D	set	1	0		
⋮	⋮	⋮	⋮	⋮	⋮	

'텍스트마이닝'의 경우 비정형화된 텍스트 데이터를 분석하는 기법이기 때문에 문장, 단어 조합 등과 같은 텍스트 형식의 속성이 포함해야 한다. Table 3.2는 '텍스트마이닝'을 위한 원시데이터를 가공하는 과정을 보여 주고 있다. 즉, '사회네트워크분석'과 같이 문자중계 내용에 포함되어 있는 불필요한 자료들을 삭제하고, 세부적인 '텍스트마이닝'을 위해 공격 및 수비의 성공 여부를 알기 위한 변수로 'Victory'를 생성시켰다. 본 연구에서는 소셜 네트워크 분석을 위해 SAS의 Optgraph procedure for SNA를 사용하였고, 텍스트마이닝 기법은 SAS의 SAS Text Miner를 사용하였다.

Table 3.2 Data treatment for text mining

Content of a letter relay	Text	Victory
game start	A serve	Win
3.A serve	B blocking success score	Win
6.B blocking success score	A serve	Win
3.A serve	B effective blocking	Win
6.B effective blocking	C dig fail	Loss
10.C dig fail	D set	Loss
5.D set		
⋮	⋮	⋮

4. 분석결과

4.1. 사회네트워크분석

4.1.1. 중심성 측도 결과

Table 4.1은 구단 내에 존재하는 총 17명의 선수들에 대한 중심성 척도들에 대한 기술통계량을 보여주고 있다. 선수들 간 평균 연결 정도는 약 26번이라고 할 수 있고, 평균 근접 정도는 0.868로 1에 매우 가까운 결과를 보이고 있음을 알 수 있다.

그리고 Table 4.1에 주어져 있지 않지만 본 연구에서 재구성한 두 그룹 (그룹 '0'과 그룹 '1')은 세 가지 중심성 척도들 모두에 대해서 유의수준 0.01하에서 유의한 차이를 보였다 (p <0.001).

Table 4.1 Result of centrality measurement

N	Measurement	Mean ± Standard error	Min	Max
17	Degree	26.471 ± 5.768	9	32
	Closeness	0.868 ± 0.108	0.582	1.000
	Betweenness	0.012 ± 0.010	0	0.036

4.1.2. 그룹별 중심성 순위

Table 4.2는 팀 내의 선수들의 연결중심성과 매개중심성의 순위를 나타낸 표이다. 연결중심성과 매개중심성의 순위로 판단할 때, 그룹 '1'이 그룹 '0' 보다 우수한 경기력을 보이는 그룹인 것으로 나타났다.

보다 세부적으로 해석하면, 연결중심성과 매개중심성 모두에서 'O' 선수가 각각 32와 0.036의 값으로 나타나 가장 많은 매개체 역할을 한 선수인 것으로 나타났다. 그룹별로는 '0' 그룹은 'N' 선수, '1' 그룹은 'O' 선수가 가장 활발한 역할을 했다고 할 수 있다. 연결성과 중개성 두 척도측면에서 볼 때, '1' 그룹에 속한 선수들이 '0' 그룹에 속한 선수들보다 높은 순위에 위치함을 알 수 있다. 그리고 연결중심성과 매개중심성의 순위들이 다른 선수들과 상대적으로 모두 낮은 'G'와 'D' 선수는 후보일 가능성이 높다고 할 수 있다.

4.1.3. 사회네트워크 구조

Figure 4.1은 선수들 간의 연관성을 알아보기 위해 작성된 사회네트워크 구조이다. 사회네트워크 구조에서 노드의 크기는 해당 노드의 Degree를 적용하여 나타낸 것이고 연결선의 굵기는 앞서 언급한 가중치를 적용하여 나타낸 것이다. 그룹의 종류는 노드의 모양으로 구분을 하였다. 사회네트워크 구조를 보면, 많은 연결선이 존재하여 선수들 상호 간에 많은 상호작용을 하고 있음을 알 수 있다 (▲: '0' 그룹,

●: '1' 그룹). 특히, 'D' 선수와 'G' 선수에게 연결된 연결선이 다른 선수들에 비해 상대적으로 작기 때문에 경기에서 활약이 적었음을 알 수 있다.

Table 4.2 Rank of centrality measurement for player

Rank	Player_code	Degree	group	Rank	Player_code	Between	group
1	O	32	1	1	O	0.036	1
2	I	31	1	2	I	0.026	1
3	N	30	0	3	M	0.022	1
3	P	30	1	3	C	0.022	1
3	K	30	1	5	P	0.017	1
3	B	30	1	6	E	0.012	0
7	M	29	1	7	N	0.011	0
7	C	29	1	7	K	0.011	1
7	Q	29	1	7	B	0.011	1
10	A	28	0	10	Q	0.009	1
10	L	28	0	11	A	0.008	0
12	E	27	0	12	L	0.005	0
13	J	24	1	13	F	0.002	0
14	F	23	0	13	J	0.002	1
14	H	23	1	15	H	0.001	1
16	D	18	0	16	D	0.000	0
17	G	9	1	16	G	0.000	1

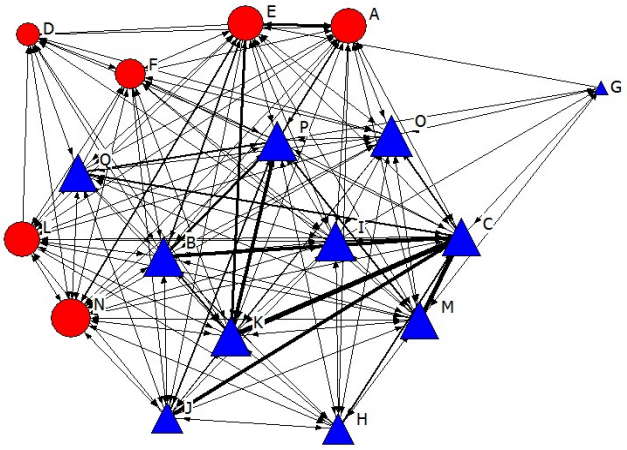

Figure 4.1 Structure of social network analysis

4.2. 텍스트마이닝 분석

4.2.1. 추출된 텍스트의 빈도

Table 4.3는 각 단어들의 출현 정도를 알아보기 위한 단어들의 빈도에 대한 결과이다. '세트'라는 단어의 빈도가 3,574로 가장 많이 나온 것을 알 수 있고, 선수 중에서는 빈도가 2,809로 'K' 선수의 이름이 가장 많이 나타난 것을 알 수 있다. 즉, 해당 팀에서는 세트 공격을 위주로 경기를 이끌어 나가며 'K' 선수가 많은 움직임이 있었고, '포히트', '오버네트'의 '파울'은 거의 없었음을 알 수 있다.

Table 4.3 Frequency of drawn words

Term	Frequency	Term	Frequency	Term	Frequency	Term	Frequency
set	3,574	M	1,551	C quick attack	647	F	77
accuracy	3,310	B	1,295	spike	612	D	34
serve	3,029	open	1,295	quick attack	604	line	27
K	2,809	P	1,240	N	571	over	27
receive	2,661	fail	1,231	H	452	touch net	26
gain	2,268	Q	1,127	L	367	fault	42
success	2,268	E	898	in	362	dribble	14
C	2,165	attack	834	out	356	catch ball	6
blocking	2,025	I	776	assist	256	G	5
J	1,713	A	714	caught net	188	four hit	10
dig	1,616	O	656	time lag	328	over net	3

4.2.2. 컨셉 링크

Figure 4.2는 여러 단어들 사이의 연관성을 시각적으로 보기 위해 수행한 컨셉 링크 (concept link)의 결과이다. 컨셉 링크에서 선의 굵기는 득점과 각 단어들과의 관련된 횟수로 표현되며, 선의 굵기가 굵을수록 관련된 단어들과 '득점'의 상호 관련성이 높다는 것을 의미한다. 컨셉 링크의 결과, 득점과 관련이 높은 선수들은 'I', 'J', 'K', 'Q' 선수들로 나타났다. '득점'과 관련성이 높은 공격 유형 (단어)은 전기한 '오픈' 유형 이외에도 '성공', 'C 속공', '시간차', '속공', '공격'임을 알 수 있다.

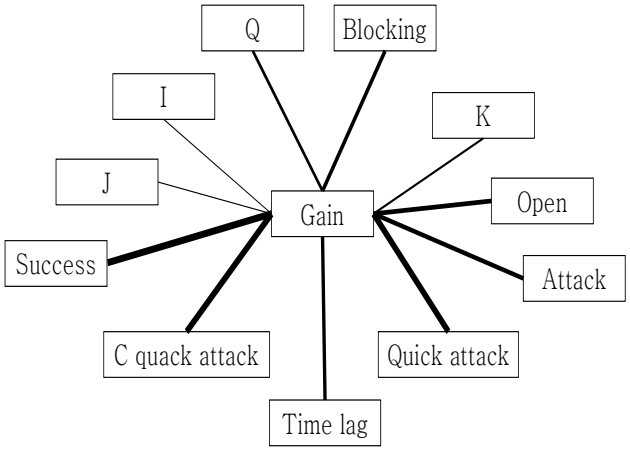

Figure 4.2 Concept link for a 'gain' letter

4.2.3. 집단 간 차이 검정

Table 4.4는 '사회네트워크분석'을 통해 도출된 그룹변수에 따라 '텍스트마이닝' 기법의 결과 '승패'에 대한 차이가 있는지를 알아보기 위해 교차분석을 수행한 결과이다. 즉, 두 그룹에 따라 경기의 '승패'에 차이를 보이는지를 알아보기 위한 분석결과이다. 교차분석결과 p값이 0.001로 나타나 유의수준 0.01하에서 두 집단에 있어서 경기의 '승패' 여부에 유의한 차이가 있는 것으로 나타났다. 빈도 (비율)로 보았을 때, '0' 그룹에 비해 '1' 그룹이 상대적으로 선수들 간 팀워크와 경기력이 좋다고 할 수 있

다.

Table 4.4 Difference test by network group

Items		Frequency (%)		χ^2	p
		loss	win		
group	0	2305 (86.62)	356 (13.38)	1768.392	<.0001
	1	5,806 (42.11)	7,983 (57.89)		

4.2.4. 클러스터링 분석 결과

(1) 전체데이터에 대한 클러스터링

Table 4.5는 단어들에 대한 클러스터링 결과이다. 클러스터링 결과, 팀은 '오픈' 유형의 공격이 주 득점 요인이라 할 수 있다. 그리고 'C' 선수와 'E' 선수는 '세트' 공격을 시도했을 때 정확한 플레이를 구사하며, 'A' 선수와 'B' 선수의 '디그' 플레이는 다소 실책성이 많았음을 알 수 있다.

Table 4.5 Text clustering (total)

Term	Frequency	%
success, gain, open, arrack	4,554	28%
C, set, perfect, E	3,481	21%
spike, I, serve, Q	2,929	18%
receive, N, M, A	2,658	16%
B, fail, dig, A	1,424	9%
Q, fail, blocking, assist	1,404	9%

(2) 승패에 따른 클러스터링

Table 4.6은 경기 결과의 '승패'에 따른 클러스터링 결과이다. '승패' 경기에 대한 클러스터링 결과에서 '승리'한 경기의 경우 공격 패턴의 시작이 되는 '리시브'가 정확하다고 할 수 있다. '승리'한 경기의 특성을 세부적으로 살펴보면, 'J'와 'Q' 선수는 '디그'에 실패가 많고, 'C'와 'P' 선수는 '세트' 플레이가 정확하다고 할 수 있다. 'I'와 'Q' 선수는 '서브'와 '스파이크'를 잘하는 선수이며, 'K' 선수는 'C 속공'의 '성공'에 의한 득점을 하는 선수라는 것을 알 수 있다. 그리고 'A'와 'M' 선수는 '리시브'가 '정확'하고, 'K' 선수는 또한 '오픈' '공격'을 할 경우 '아웃'되는 경향이 있는 것으로 나타났다.

Table 4.6 Text clustering by outcome (left: win, right: loss)

Win			Loss		
Term	Frequency	%	Term	frequency	%
fail, J, Q, dig	2,141	26	accuracy, E, set, C	1,681	21
accuracy, C, P, set	1,652	20	open, sucess, gain, attack	1,406	17
serve, spike, Q, I	1,551	19	Q, O, serve, spike	1,378	17
success, gain, C quick attack, K	1,243	15	B, accuracy, L, A	1,352	17
A, M, accuracy, receive	1,124	13	I, fail, blocking, Q	1,347	17
K, attack, open, out	628	8	fail, dig, B, P	947	12

'패배'한 경기의 경우, '브로킹'과 '디그'의 '실패'가 많은 것으로 나타났다. '패배'한 경기의 특성을 세부적으로 살펴보면, 'C'와 'E' 선수는 '세트' 플레이가 정확하며, '오픈'과 '공격'에 의한 성공률이 높고 득점을 하는 것으로 나타났다. 'O'와 'Q' 선수는 '서브'와 '스파이크'를 했으며, 'A'와 'B' 선수, 그

리고 'L' 선수는 정확성이 높음을 알 수 있다. 그러나 'I'와 'Q' 선수는 '블로킹'의 실수가 있었으며, 'B'와 'P' 선수는 '디그'의 실수가 있었음을 알 수 있다.

(3) 네트워크 그룹에 따른 클러스터링

Table 4.7은 네트워크 그룹별 클러스터링 결과이다. '0' 그룹의 경우 'D'와 'E' 선수가 '세트' 플레이를 통하여 정확하게 득점한다고 할 수 있다. 'L'와 'A' 선수가 주로 정확한 플레이를 구사하며, 'C'와 'N' 선수가 '리시브', '속공', '블록킹'이 뛰어나다고 할 수 있다. 그리고 'A' 선수는 '서브'와 '스파이크'에 능하나 볼이 아웃시킬 가능성이 높으며, '스파이크 서브'와 '디그'에 의해서 공격이 실패하는 경향이 있는 것으로 나타났다. 'L' 선수는 '오픈', '시간차' 유형의 공격을 통해 득점을 하는 경향이 있다.

'1' 그룹의 경우 'K' 선수가 '디그'에 의해서 '공격'을 하는 경우 실패하는 경우가 많고, 'C'와 'P' 선수는 '세트' 플레이가 정확하다고 할 수 있다. 'I'와 'Q' 선수는 '서브'와 '스파이크'에 강하고, 'K' 선수는 'C 속공'의 성공률이 높음을 알 수 있다. 그리고 'B' 와 'J' 선수, 'M' 선수는 정확성이 높음을 알 수 있고, '어시스트', '블로킹', '속공'은 '실패'할 확률이 높음을 알 수 있다.

Table 4.7 Text clustering by network group (left: 0 group, right: 1 group)

0 group			1 group		
Term	Frequency	%	Term	Frequency	%
accuracy, E, set, D	735	28	fail, dig, attack, K	3,385	25
L, A, accuracy, receive	498	19	C, P, set, accuracy	2,621	19
N, receive, C quick attack, blocking	486	18	serve, spike, Q, I	2,390	17
serve, spike, out, A	433	16	C quick attack, K, success	2,048	15
L, open, time difference, gain	343	13	B, accuracy, J, M	1,937	14
fail, dig, A, blocking	166	6	fail, assist, blocking, quick attack	1,408	10

5. 결론 및 향후 연구과제

본 연구에서는 2011~2012 V-리그 국내 남자프로배구 경기의 문자중계 데이터를 이용하여 국내 남자프로배구 구단의 공격, 패스 패턴을 찾아내고 배구경기력과 관련된 핵심 키워드 추출하였다. 이를 토대로 선수들의 경기력을 평가하기 위해서 '사회네트워크분석'과 '텍스트마이닝' 기법을 이용하여 분석을 수행하였다.

분석결과는 다음과 같다. '사회네트워크 분석'의 결과 '0'그룹 (6명)과 '1'그룹 (11명)의 네트워크 그룹을 생성시켰다. 그룹 내에서 활발한 경기력을 보여준 선수는 'O' 선수와 'N' 선수인 것을 알 수 있었다. 컨셉 링크의 결과, 득점과 관련성이 높은 선수는 'I', 'J', 'K', 'Q' 선수들로 나타났다. '득점'과 관련성이 높은 공격 유형은 '오픈', '성공', 'C 속공', '시간차', '속공', '공격'으로 나타났다. '사회네트워크분석'을 통해 도출된 그룹변수에 따라 '텍스트마이닝' 기법의 결과인 경기의 '승패'에 차이는 p값이 0.001로 나타나 유의수준 0.01하에서 유의한 차이가 있는 것으로 나타났다. 빈도 (비율)로 보았을 때, '0' 그룹에 비해 '1' 그룹이 상대적으로 선수들 간 팀워크와 경기력이 좋다고 할 수 있다.

전체 데이터에 대한 클러스터링 결과, 팀은 '오픈' 유형의 공격이 주 득점 요인이라 할 수 있다. 그리고 'C'와 'E' 선수는 '세트' 공격을 시도했을 때 정확한 플레이를 구사하며, 'A'와 'B' 선수의 '디그' 플레이는 다소 실책성이 많았음을 알 수 있다. '승패' 경기에 대한 단어들의 클러스터링 결과, '승리'한 경기의 경우 공격 패턴의 시작이 되는 '리시브'가 정확하다고 할 수 있다. '패배'한 경기의 경우, '브로킹'과 '디그'의 '실패'가 많은 것으로 나타났다. 네트워크 그룹별 클러스터링 결과, '0' 그룹의 경우 'D'와 'E' 선수가 '세트' 플레이를 통하여 정확하게 득점한다고 할 수 있다. '1' 그룹의 경우 'K' 선수가

'디그'에 의해서 '공격'을 하는 경우 실패하는 경우가 많고, 'C'와 'P' 선수는 '세트' 플레이가 정확하다고 할 수 있다.

본 연구에서는 배구 구단의 경기력을 평가하는데 있어서 한 구단의 문자중계 데이터를 기준으로 하여 분석하였다. 배구 경기에서의 경기력은 해당 구단은 물론 다른 구단의 경기력 등에 적지 않은 영향을 미칠 수 있다. 본 연구는 한 시즌 (2011∼2012) 경기 결과를 이용하였기 때문에 구단 경기력을 판단하는데 기초자료로 활용하는데 한계가 있다. 따라서 향후에는 여러 시즌에 걸쳐 전체 배구 구단의 경기 문자중계 데이터를 활용하여 배구 경기력을 평가함으로써 보다 정밀한 구단 전략을 수립하는데 충분한 기초자료로 활용해야 할 것이다. 그리고 '사회네트워크분석'을 수행하는데 있어서 선수들 간에 각종 액션들의 복잡한 상호 관련성으로 인해 방향성을 고려하지 않았기 때문에 보다 더 정밀한 분석을 위해서는 방향성을 고려한 분석이 필요하다. 더 나아가 '사회네트워크분석'과 '텍스트마이닝' 분석 이외에 공간 정보를 이용한 공간통계학적 분석과 선수들의 바이오리듬 정보를 결합하여 분석한다면 보다 정밀한 정보를 얻을 수 있을 것으로 기대된다.

References

Cho, J. S. (2012). Inflow and outflow analysis of double majors using social network analysis. *Journal of the Korean Data & Information Science Society*, **23**, 693-701.

Cho, J. S. (2014). Analysis of employee's characteristic using data visualization. *Journal of the Korean Data & Information Science Society*, **25**, 727-736.

Choi, S. B., Kang, C. W., Choi, H. J. and Kang, B. W. (2011). Social network analysis for a soccer game. *Journal of the Korean Data & Information Science Society*, **22**, 1053-1063.

Choi, S. B. (2013). Characteristic analysis for moving in and moving out of departments - Focused on the D university example -. *Journal of the Korean Data & Information Science Society*, **24**, 105-115.

Huh, M. H. (2010). *Introduction to social network analysis utilizing R*, Free Academy, Paju.

Jung, K. H. (2010). *A study of foresight method based on text mining and complexity network analysis*, Korea Institute of S&T Evaluation and Planning, Seoul.

Kang, B. S. (2010). Performance improvement methods for new customer recommendations using degree centrality of social network. *Journal of the Korean Data Analysis*, **12**, 1511-1521.

Kim, G. H. and Park, C. (2015). Analysis of English abstracts in Journal of the Korean Data & Information Science Society using topic models and social network analysis. *Journal of the Korean Data & Information Science Society*, **26**, 151-159.

Kim, Y. H. (2007). *Social network analysis*, Parkyeungsa, Seoul.

Oh, H. S., Cho, S. K., Kang, C. W. and Lim, D. S. (2010). Fashion company's claim data analysis using text mining. *Journal of the Korean Data Analysis Society*, **12**, 297-305.

Oh, S. W. and Jin, S. H. (2012). A study on analysis of internet shopping mall customers' reviews by text mining. *Journal of the Korean Data Analysis Society*, **14**, 125-137.

Park, H. W. and Lee, Y. O. (2009). A mixed text analysis of user comments on a portal site : The 'BBK scandal' in the 2007 presidential election of south Korea. *Journal of the Korean Data Analysis Society*, **11**, 731-744.

SAS Korea. (2010). *Optgraphic procedure for SNA*, SAS Software Korea Ltd.

Son, D. W. (2010). *Social network analysis*, Parkyeungsa, Seoul.

Won, D. K. and Choi, K. H. (2014). Network analysis and comparing citation index of statistics journals. *Journal of the Korean Data Analysis Society*, **25**, 317-325.

Journal of the Korean Data & Information Science Society
2015, **26**(3), 619–630

http://dx.doi.org/10.7465/jkdi.2015.26.3.619
한국데이터정보과학회지

Performance analysis of volleyball games using the social network and text mining techniques[†]

Byounguk Kang[1] · Mankyu Huh[2] · Seungbae Choi[3]

[1]EnSOFTechnology Inc.
[2]Department of Molecular Biology, Dongeui University
[3]Department of Data Information Science, Dongeui University

Received 25 February 2015, revised 12 March 2015, accepted 18 May 2015

Abstract

The purpose of this study is to provide basic information to develop a game strategy plan of a team in a future by identifying the patterns of attack and pass of national men's professional volleyball teams and extracting core key words related with volleyball game performance to evaluate game performance using 'social network analysis' and 'text mining'. As for the analysis result of 'social network analysis' with the whole data, group '0' (6 players) and group '1' (11 players) were partitioned. A point of view the degree centrality and betweenness centrality in 'social network analysis' results, we can know that the group '1' more active game performance than the group '0'. The significant result for two group (win and loss) obtained by 'text mining' according to two groups ('0' and '1') obtained by 'social network analysis' showed significant difference (p-value: 0.001). As for clustering of each network, group '0' had the tendency to score points through set player D and E. In group '1', the player K had the tendency to fail if he attack through 'dig'; players C and D have a good performance through 'set' play.

Keywords: Centrality measurement, social network analysis, text clustering analysis, text mining technique, web science.

[†] This work was reconstructed from a master's degree thesis of Byounguk Kang who is coauthor.

[1] Assistant manager, EnSOFTechnology Inc., Seoul 150-010, Korea.
[2] Professor, Department of Department of Molecular Biology, Dongeui University, Busan 614-714, Korea.
[3] Corresponding author: Professor, Department of Data Information Science, Dongeui University, Busan 614-714, Korea. E-mail: csb4851@deu.ac.kr