[Spring 2023] DATA MINING THEORY AND APPLICATION (IIE 4102)

Practice-HW5
Social Network Analysis

논문참조

강세정, 김규현, 신재욱, 장현우

Introduction

01. Research Objective

- 삼국지 내에서의 시간의 흐름에 따른 등장인물들 간의 인간관계의 변화와 국가의 흥망성쇠를 시각적으로 확인해보고자 함
- 텍스트 분석 기반의 SNA를 수행하고자 함

02. Data

데이터는 삼국지 영문판(C. H. Brewitt-Taylor 번역본) txt 파일을 다운로드해 사용

삼국지의 Chapter 1~120 까지를 총 4개의 season으로 구분해 사용

① data

```
****** Romance of the Three Kingdoms ******
**** Chapter 1 ****
The world under heaven, after a long period of division, tends to unite; after
a long period of union, tends to divide. This has been so since antiquity. When
the rule of the Zhou Dynasty weakened, seven contending kingdoms sprang up,
warring one with another until the kingdom of Qin prevailed and possessed the
empire. But when Qinâ□□s destiny had been fulfilled, arose two opposing
kingdoms, Chu and Han, to fight for the mastery. And Han was the victor.
The rise of the fortunes of Han began when Liu Bang the Supreme Ancestor slew a
white serpent to raise the banners of uprising, which only ended when the whole
empire belonged to Han (BC 202). This magnificent heritage was handed down in
successive Han emperors for two hundred years, till the rebellion of Wang Mang
caused a disruption. But soon Liu Xiu the Latter Han Founder restored the
empire, and Han emperors continued their rule for another two hundred years
till the days of Emperor Xian, which were doomed to see the beginning of the
empireâ□□s division into three parts, known to history as The Three Kingdoms.
But the descent into misrule hastened in the reigns of the two predecessors of
Emperor Xian â□□-Emperors Huan and Ling â□□-who sat in the dragon throne about
the middle of the second century.
```

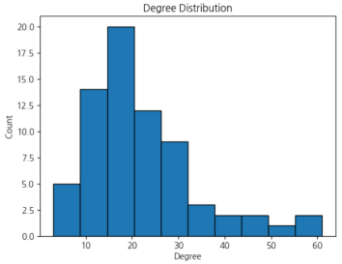
② 텍스트 전처리 : 여러 함수들을 생성해 전처리 수행

- `wordTokens()` : 소문자화 및 토큰화
- `char_tuple_f()` : 이름 list를 tuple로 변환
- `indices_dic()` : 텍스트에서 각 이름이 언급된 위치를 찾아 리스트로 저장
- `links_dic_f()` : 두 등장인물 간의 거리가 threshold 이하인 횟수를 세어 연관도로 저장
- `remove_zero_link_chars()` : interaction이 없는 등장인물 제거
- `edge_tuples_f()` : node-edge 관계로 변환
- `convert_to_korean()` : 영어로 된 등장인물 이름 한글로 변환

③ Node, Edge : undirected, weighted, edge list 방식으로 생성

- Node : 삼국지 등장 인물
- Edge : 인물들 간 인접 등장 횟수
- 70명의 등장인물이, 평균 22(766 x 2 / 70)번의 관계를 가짐

```
Number of nodes: 70
Number of nodes: 766
```



Body

Contents

Social Network Analysis

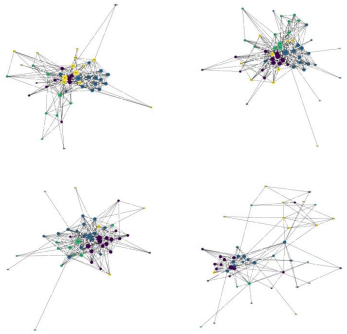
가중치 계산 방법

Edge attribute (weight)

- 두 등장인물 간의 등장 거리가 threshold(15 단어) 이내인 횟수

그래프 생성

- Undirected, Weighted 그래프
- 시즌 별로 interaction이 없는 등장인물(Weight가 0인) 노드 삭제
- 등장인물이 속한 나라별 노드 색 지정



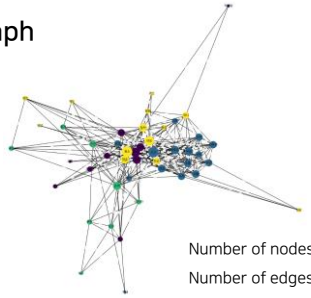
Centrality 합을 통한 국가의 영향력 파악

- 각 국가에 소속된 인물들의 DGC, BTC, CLC, EVC 각 총합을 파악하여 시기별 각 국가의 국제 정세와 영향력 측정
- 이를 실제 역사의 세력도와 비교하여, SNA의 효용성을 평가해보고자 함

SNA

시즌 1

(1) Graph

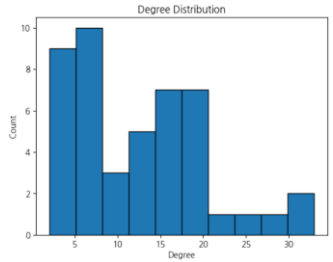


Number of nodes : 46

Number of edges : 289

(2) Attribute

- Edge attribute(weight) : 각 edge의 가중치 지정
- Node attribute : 각 인물이 속한 나라 지정(촉=0 / 위=1 / 오=2 / 기타=3)



(3) Social Network Analysis

- Structure

diameter : 4

density : 0.2792270531400966

transitivity : 0.5493046776232617

reciprocity : 0.0

<해석>

▶ 소설의 초반부는, 원소가 eigenvector centrality를 제외한 모든 centrality에서 높음. 정말 중요한 핵심 인물로서 등장함

▶ 조조는 eigenvector centrality 값이 가장 높음. 핵심 인물인 원소, 여포 등과 함께 등장함. 원소의 라이벌이었던 조조이기에 역사와 일치하는 결과임

▶ 여포, 유비, 하후돈은 원소와 조조 다음으로 핵심 인물로 소설의 초반부에 등장함

▶ 손책과 동탁은 between centrality 값만이 높음. 즉, 다른 인물들과의 연결다리 역할을 수행함

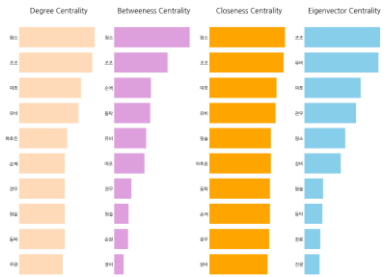
- Centrality : 상위 5명 출력

degree centrality : 원소=0.73 / 조조=0.71 / 여포=0.60 / 유비=0.58 / 하후돈=0.47

betweenness centrality : 원소=0.16 / 조조=0.11 / 손책=0.08 / 동탁: 0.08 / 유비=0.07

closeness centrality : 원소=0.79 / 조조=0.78 / 여포=0.70 / 유비=0.69 / 하후돈=0.64

eigenvector centrality : 조조=0.29 / 원소=0.28 / 여포=0.26 / 유비=0.24 / 하후돈=0.23



(4) 영향력 비교: 소속 인물의 각 centrality 합 / 후한 말 세력의 잔재, 위나라 강세

▶ 초반부는 위, 촉, 오 나라의 인물이 아닌 원소 등의 인물이 아직까지 핵심 인물로 작용하는 부분임. 이에 region 3가 다른 시즌에 비해 큰 값을 가짐. 특히, DGC, BTC에서 큰 값을 가지는 것으로 보아, 직접적인 이웃도 많으며 연결 다리로서의 역할도 수행함

▶ 위나라는 초반부터 핵심 인물이 다수 존재하는, 세계관에 큰 영향력을 끼치는 국가. degree centrality, closeness centrality에서 특히 강세를 보임

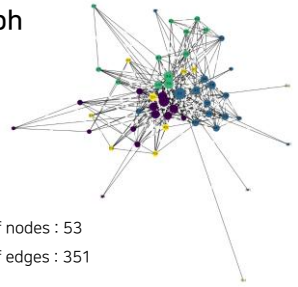
region	DGC	BTC	CLC	EVC
0	2.533333	0.162132	5.387982	1.267696
1	5.133333	0.222126	8.674548	1.103092
2	1.622222	0.149308	4.102808	0.082406
3	3.555556	0.364415	7.236793	1.153263

Body

Contents

시즌 2

(1) Graph

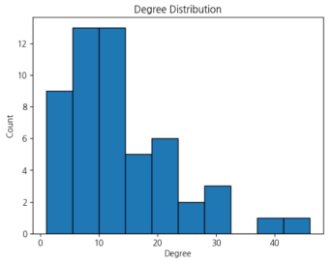


Number of nodes : 53

Number of edges : 351

(2) Attribute

- Edge attribute(weight) : 각 edge의 가중치 지정
- Node attribute : 각 인물이 속한 나라 지정(촉=0 / 위=1 / 오=2 / 기타=3)



(3) Social Network Analysis

- Structure

diameter : 3

density : 0.25471698113207547

transitivity : 0.49463683726631935

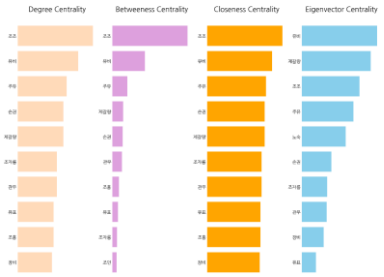
reciprocity : 0.0

<해석>

▶ 조조가 전반적으로 높은 값을 가지며 특히 BTC의 값은 2등보다 2배 이상 높음. 즉 그는 적과 동맹국 모두에게 주요 관심사임

▶ 유비는 조조 다음으로 높은 EVC 값을 보였으며, 1장에 비해서 촉나라의 강세가 강해졌고 더욱 확고한 영주가 되었음을 의미함

▶ 주유 또한 유비와 같이 전반적으로 고른 centrality값을 보이며 다양한 인물들과 많은 교류를 하고 있음



- Centrality : 상위 5명 출력

degree centrality : 조조=0.88 / 유비=0.71 / 주유=0.58 / 손권=0.54 / 제갈량=0.54

betweenness centrality : 조조=0.30 / 유비=0.13 / 주유=0.06 / 제갈량: 0.04 / 손권=0.04

closeness centrality : 조조=0.90 / 유비=0.78 / 주유=0.70 / 손권: 0.68 / 제갈량=0.68

eigenvector centrality : 조조=0.30 / 유비=0.28 / 주유=0.24 / 제갈량=0.23 / 손권=0.23

(4) 영향력 비교: 소속 인물의 각 centrality 합 / 위나라 강세, 촉·오나라 성장

▶ 조조에 대항하는 촉나라와 오나라의 인물들이 서로 상호작용함. 이로 인해서 그들의 EVC 값이 높게 나타남

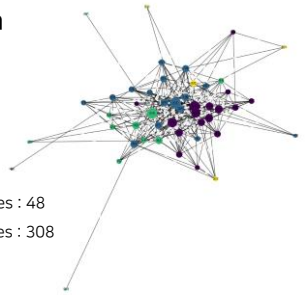
▶ 1장에 비해서 촉나라와 오나라의 영향력이 강해짐

▶ 1장에서는 거의 언급되지 않았던 제갈량이 EVC에서 매우 높은 값을 보임. 즉, 그는 다른 많은 핵심적인 등장 인물들에게 높은 영향력을 미침

region	DGC	BTC	CLC	EVC
0	4.250000	0.286822	8.174287	1.727264
1	4.538462	0.384360	9.792777	0.755107
2	3.057692	0.126983	6.825486	0.991669
3	1.653846	0.023102	5.334044	0.231436

시즌 3

(1) Graph

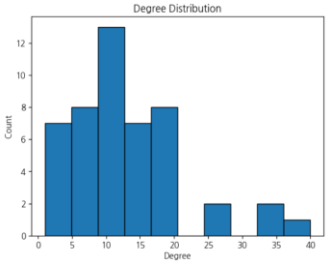


Number of nodes : 48

Number of edges : 308

(2) Attribute

- Edge attribute(weight) : 각 edge의 가중치 지정
- Node attribute : 각 인물이 속한 나라 지정(촉=0 / 위=1 / 오=2 / 기타=3)



(3) Social Network Analysis

- Structure

diameter : 4

density : 0.2730496453900709

transitivity : 0.4835820895522388

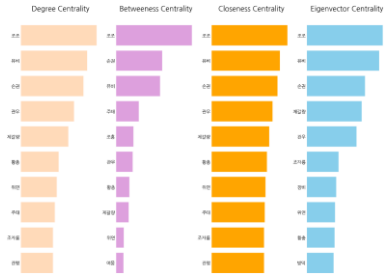
reciprocity : 0.0

<해석>

▶ 조조가 여전히 가장 높은 값을 가지고 있으며 가장 중요하고 영향력있는 인물로 묘사됨

▶ 유비와 손권이 그 다음으로 높은 centrality 값을 보임

▶ 관우가 그 다음으로 높은 수치를 보임. 삼국지의 해당 부분에서 관우는 유비에게 충성을 다하는 인물로 묘사됨과 동시에 조조와 끊임없이 합을 겨누기 때문에 전반적인 수치가 높게 나타남



- Centrality : 상위 5명 출력

degree centrality : 조조=0.85 / 유비=0.74 / 손권=0.70 / 관우=0.60 / 제갈량=0.53

betweenness centrality : 조조=0.20 / 손권=0.12 / 유비=0.12 / 주태=0.06 / 조홍=0.04

closeness centrality : 조조=0.87 / 유비: 0.78 / 손권: 0.76 / 관우=0.70 / 제갈량=0.66

eigenvector centrality : 조조=0.31 / 유비: 0.29 / 손권: 0.27 / 관우=0.26 / 제갈량=0.23

(4) 영향력 비교: 소속 인물의 각 centrality 합 / 위나라의 여전한 영향력·촉나라의 강세 증가

▶ 2장에서 높은 값을 가지고 있던 주유가 3장에서는 등장하지 않는데, 그 이유는 2장의 끝부분에서 주유가 사망했기 때문임

▶ 2장에서 주유의 영향력이 3장에서 손권의 영향력으로 대체

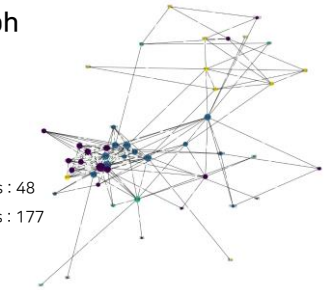
region	DGC	BTC	CLC	EVC
0	5.319149	0.290584	9.496098	2.235704
1	4.872340	0.386102	9.662263	1.257087
2	2.404255	0.160774	5.939317	0.831526
3	0.510638	0.008979	2.029028	0.108735

Body

Contents

시즌 4

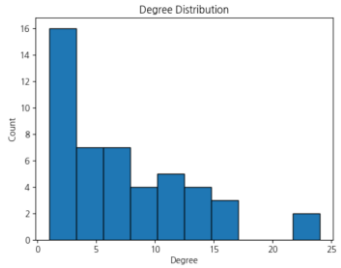
(1) Graph



Number of nodes : 48
Number of edges : 177

(2) Attribute

- Edge attribute(weight) : 각 edge의 가중치 지정
- Node attribute : 각 인물이 속한 나라 지정(촉=0 / 위=1 / 오=2 / 기타=3)



(3) Social Network Analysis

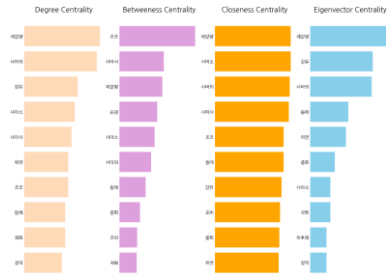
- Structure

diameter : 5
density : 0.15691489361702127
transitivity : 0.5100316789862724
reciprocity : 0.0

<해석>
▶ 제갈량을 제외한 다른 주요 등장인물들은 3장 후반부~4장 초반부에서 모두 사망했기 때문에 centrality 값이 낮아짐
▶ 따라서 이전과는 다르게 제갈량과 사마의가 가장 높은 centrality 값을 보임

- Centrality : 상위 5명 출력

degree centrality : 제갈량=0.51 / 사마의=0.49 / 강유=0.36 / 사마소=0.34 / 사마사=0.32
betweenness centrality : 조조=0.24 / 사마사=0.14 / 제갈량=0.13 / 손권=0.12 / 사마소=0.11
closeness centrality : 사마소=0.59 / 제갈량=0.59 / 사마의=0.58 / 사마사=0.57 / 조조=0.53
eigenvector centrality : 사마의=0.34 / 제갈량=0.33 / 강유=0.29 / 위연=0.26 / 사마소=0.25



(4) 영향력 비교: 소속 인물의 각 centrality 합 / 촉나라 쇠퇴, 위나라 강성

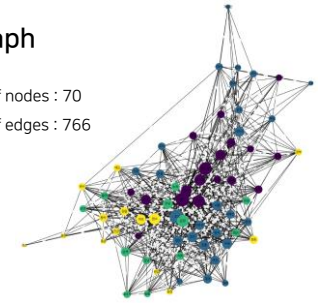
region	DGC	BTC	CLC	EVC
0	2.851064	0.285885	6.977238	1.531835
1	3.276596	0.874473	7.527919	1.826410
2	0.531915	0.166453	2.350687	0.052004
3	0.872340	0.140348	3.725761	0.085113

▶ 2~3장의 주요 인물 중 유일하게 제갈량이 4장까지 살아남아 큰 영향력을 미침. 실제 삼국지에서 그는 이전 세대와 다음 세대의 연결고리가 되는데 이러한 부분이 수치로도 반영됨
▶ 특이점은 조조가 여전히 높은 BTC값을 보이고 있다는 것인데, 해당 시점에서 조조는 이미 사망한 상태임에도 영향력이 유지되었다는 것을 의미함

시즌 전체

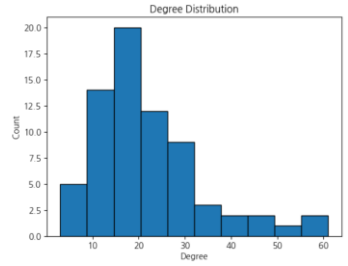
(1) Graph

Number of nodes : 70
Number of edges : 766



(2) Attribute

- Edge attribute(weight) : 각 edge의 가중치 지정
- Node attribute : 각 인물이 속한 나라 지정(촉=0 / 위=1 / 오=2 / 기타=3)



(3) Social Network Analysis

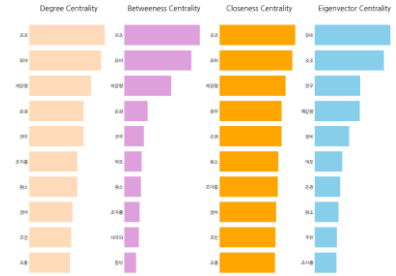
- Structure

diameter : 3
density : 0.31718426501035196
transitivity : 0.5279175537062125
reciprocity : 0.0

<해석 >
▶ 조조, 유비, 제갈량, 관우, 손권이 거의 순서대로 각 중심성에서 핵심적 면모를 보임
▶ 각 인물에 대한 해석은 결론 부분 참조

- Centrality : 상위 5명 출력

degree centrality : 조조=0.88 / 유비=0.84 / 제갈량=0.72 / 관우=0.64 / 손권=0.64
betweenness centrality : 조조=0.11 / 유비=0.10 / 제갈량=0.07 / 손권=0.04 / 관우=0.03
closeness centrality : 조조=0.90 / 유비=0.86 / 제갈량=0.78 / 관우=0.73 / 손권=0.73
eigenvector centrality : 조조=0.25 / 유비=0.24 / 제갈량=0.22 / 관우=0.21/ 손권=0.20



(4) 영향력 비교: 소속 인물의 각 centrality 합

region	DGC	BTC	CLC	EVC
0	6.855072	0.300140	11.189627	1.991115
1	8.304348	0.273735	15.446598	1.264027
2	3.695652	0.074331	7.048061	0.643278
3	3.347826	0.087515	7.784154	0.695748

▶ 우리는 흔히 유비·관우·장비로 대표되는 촉나라가 삼국지의 주인공이라고 생각하지만, 실제로는 처음부터 끝까지 위나라가 소설 내에서 가장 높은 영향력을 행사하였음
▶ 그러나 EVC만큼은 촉나라가 위나라보다 훨씬 높은 값을 보이는데, 이는 위나라는 조조라는 독보적인 인물이 중심이 되었던 반면에 촉나라는 유비를 중심으로 능력 있고 비중이 높은 여러 인물들이 고르게 영향력을 가지고 있었다고 해석할 수 있음

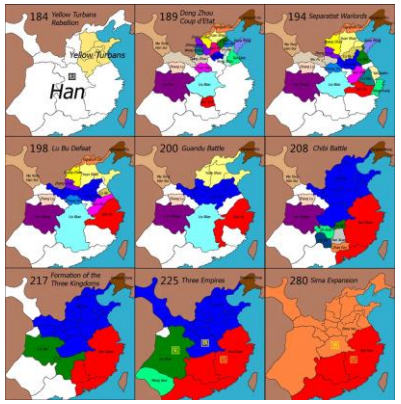
Conclusion

01. Result

- SNA를 통한 삼국지의 4부작을 분석함으로써, 등장인물의 흥망성쇠를 파악할 수 있음

- 각 국가의 인물의 centrality 합과 역사의 세력도 비교를 통해, SNA의 효용성을 확인할 수 있음

- 소설에서 가장 영향력 있는 인물은 네트워크와 centrality을 보면 조조임이 분명함



- 유비는 두 번째로, 특히 높은 Eigenvector centrality 에서 알 수 있듯 가장 중요한 연결을 가진 인물임

- 손권은 다른 centrality에 비해 Eigenvector centrality이 낮음. 핵심 인물과의 교류는 적은 인물임

- 관우는 다른 centrality에 비해 Eigenvector centrality이 높음. 실제로 핵심 인물인 조조, 유비와의 교류가 잦았음

- 각 인물의 시즌 별 centrality 변화 파악을 통해 삼국지의 흐름을 파악할 수 있음

02. Limitation

(1) Data 관련 한계

- threshold 이내에 동일 인물이 여러 번 반복하여 언급되는 경우 중복으로 세어지는 문제점이 있음
(e.g. 유비는 느꼈다. 조조가 유비를 견제하고 있다는 것을. → 유비-조조 관계가 2번 세어짐)

(2) SNA 관련 한계

- 등장 횟수가 해당 인물의 소설 내 영향력이라고 가정했지만, 많이 등장한다고 해서 꼭 영향력이 높다고 할 수는 없음

03. Development

(1) Data 관련 개선점

- 삼국지에 등장하는 인물은 1000명이 넘는데 그 중 주요 인물 70명만 다루었기 때문에, 주변 인물들도 포함하면 더 방대한 결과 분석이 가능할 것이라 예상됨

(2) SNA 관련 개선점

- 해당 코드에서는 인물 간의 관련성을 계산할 때 threshold 미만의 거리에 해당 두 인물이 등장하는지로 판단했는데, 문장 단위로 판단한다면 더 명확한 관련성 계산이 가능할 것이라 예상됨

Data Source

[데이터 출처]

(1) The Romance of Three Kingdoms (Translated by C. H. Brewitt-Taylor)
- http://anthony.sogang.ac.kr/ThreeKingdoms/index.htm

[참고 코드]

(1) Github 코드
- https://github.com/dmanolidis/three-kingdoms