

# 목 차

I. 서론	1
1. 연구배경 및 필요성	1
2. 연구의 목적 및 방법	3
II. 이론 및 선행 연구의 고찰	4
1. 영화 흥행 관련연구	4
2. 영화 감정 어휘 관련연구	5
3. 의사결정나무분석 관련연구	6
4. 시각화분석 관련연구	8
III. 데이터 수집 및 정제	10
1. 영화 리뷰 데이터 수집	10
2. 감정어휘 사전 구축 및 감정어휘 데이터 생성	10
3. 데이터 특성 파악	13
IV. 통계를 활용한 예측 분석	15
1. 군집분석과 MDS시각화를 이용한 영화 장르 군집화	15
가. 군집분석의 정의	16
나. 군집분석 결과	16
다. 군집 결과 시각화	18
2. 의사결정나무분석을 활용한 영화 흥행도 예측	19
가. 의사결정 나무 분석의 정의	19
나. 전체 영화에 대한 의사결정나무분석	20
다. 군집 1 영화에 대한 의사결정나무분석	22
라. 군집 2 영화에 대한 의사결정나무분석	24
마. 군집 3 영화에 대한 의사결정나무분석	26
바. 군집 4 영화에 대한 의사결정나무분석	28
사. 의사결정나무분석 결과에 대한 종합적인 해석	30

V. 시각화 분석 및 검증	32
1. Parallel coordinates의 개념	32
2. Parallel coordinates의 기능	33
가. 번들링(Bundling)	33
나. 축(Axes)	34
다. 색상(Colour)	34
라. 기술 통계(Descriptive statistic)	34
마. 데이터 선택(Data Selection)	35
바. 제거된 데이터 표현	36
3. Parallel coordinates를 활용한 분석	37
가. 영화 장르 별 흥행도의 분포와 대표 감정 어휘의 분포	37
나. 영화의 대표 감정 어휘 사이의 상관관계	38
4. 통계분석 결과에 대한 시각화 검증	40
가. 전체 영화에 대한 의사결정나무분석 및 시각화 검증	40
나. 군집 1 영화에 대한 의사결정나무분석 및 시각화 검증	43
다. 군집 2 영화에 대한 의사결정나무분석 및 시각화 검증	46
라. 군집 3 영화에 대한 의사결정나무분석 및 시각화 검증	49
마. 군집 4 영화에 대한 의사결정나무분석 및 시각화 검증	52
바. 시각화 검증 결과에 대한 종합적인 해석	54
VI. 결론	57
참고문헌	59
Abstract	62

## 그림 목차

그림 1. 최종후 외 1명 연구의 <그림1> “의사결정나무” .....	6
그림 2. 권영란 외 1명 연구의 <그림1> “The construction of decision tree.” .....	7
그림 3. Eser Kandogan 연구의 <그림12> "Overview of 'churn' dataset, where churned customers are marked with blue (dark) color." .....	8
그림 4. Soon Tee Teoh 외 1명 연구의 <그림5> “An auxiliary display is shown on the un-utilized space at the lower left of the display.” .....	9
그림 5. 유클리디안 거리, 최장거리 연결법 수행을 수행한 수형도 .....	17
그림 6. 유클리디안 거리, 최단거리 연결법 수행을 수행한 수형도 .....	17
그림 7. MDS시각화 분석 결과 .....	18
그림 8. 전체 영화에 대한 의사 결정 나무 분석 .....	21
그림 9. 군집 1에 속한 영화에 대한 의사 결정 나무 분석 .....	23
그림 10. 군집 2에 속한 영화에 대한 의사 결정 나무 분석 .....	25
그림 11. 군집 3에 속한 영화에 대한 의사 결정 나무 분석 .....	27
그림 12. 군집 4에 속한 영화에 대한 의사 결정 나무 분석 .....	29
그림 13. 분포에 따른 Parallel coordinates .....	32
그림 14. (왼쪽) 일반적인 Parallel coordinates (오른쪽) 번들링 기능을 추가한 Parallel coordinates .....	33
그림 15. (왼쪽) 기본적인 데이터 축의 순서 (오른쪽) Happy의 데이터 변수 축 순서 변경 .....	34
그림 16. 영화의 장르별로 지정된 색상 .....	34
그림 17. 선택된 데이터 변수들의 평균값과 영화 수의 합계 .....	35
그림 18. 선택된 데이터 변수들의 평균 값과 평균 선 (굵은 line 그래프) .....	35
그림 19. 김씨 표류기(Castaway on the Moon)에 대한 하이라이트 .....	35

그림 20. 장르가 액션 & 코미디이고 상영 스크린 수가 100개 이상	36
그림 21. 선택되지 않은 데이터가 제거되는 화면	36
그림 22. 선택되지 않은 데이터를 표현하는 화면	36
그림 23. 영화의 대표 장르가 코미디인 영화에 대한 감정 어휘 분포	37
그림 24. 영화의 대표 장르가 호러인 영화에 대한 감정 어휘 분포	38
그림 25. Disgust값이 0.1이하(왼쪽), 0.1이상 0.3이하(가운데), 0.3이상(오른쪽)일 때의 분석 결과	39
그림 26. Sad값이 0.175이하(왼쪽), 0.175이상 0.35이하(가운데), 0.35이상(오른쪽)일 때의 분석 결과	39
그림 27. Surprise값이 0.2이하(왼쪽), 0.2이상 0.4이하(가운데), 0.4이상(오른쪽)일 때의 분석 결과	40
그림 28. 분할기준이 적용되기 전의 Parallel coordinates시각화	41
그림 29. Happy > 0.235이 적용된 시각화 결과	41
그림 30. Happy > 0.235 & Anger < 0.045이 적용된 시각화 결과	41
그림 31. Happy > 0.235 & Anger < 0.045 & Sad > 0.145이 적용된 시각화 결과	42
그림 32. 노드 14 에 대한 최종 Parallel coordinates	42
그림 33. 노드 14 에 최종 포함된 영화 정보	43
그림 34. 전체 영화에 대한 최종 Parallel coordinates	43
그림 35. 분할기준이 적용되기 전의 Parallel coordinates시각화	44
그림 36. Surprise > 0.175이 적용된 시각화 결과	44
그림 37. Surprise > 0.175 & Boring < 0.065이 적용된 시각화 결과	44
그림 38. Surprise > 0.175 & Boring < 0.065 & Happy > 0.125이 적용된 시각화 결과	45
그림 39. 노드 16 에 대한 최종 Parallel coordinates	45
그림 40. 노드 16 에 최종 포함된 영화 정보	46
그림 41. 군집1에 대한 최종 Parallel coordinates	46
그림 42. 분할기준이 적용되기 전의 Parallel coordinates시각화	47

그림 43. Happy > 0.505 이 적용된 시각화 결과	47
그림 44. Happy > 0.505 & Surprise > 0.195 이 적용된 시각화 결과	47
그림 45. 노드 14 에 대한 최종 Parallel coordinates	48
그림 46. 노드 14 에 최종 포함된 영화 정보	48
그림 47. 군집2에 대한 최종 Parallel coordinates	49
그림 48. 분할기준이 적용되기 전의 Parallel coordinates시각화	49
그림 49. Happy > 0.295이 적용된 시각화 결과	50
그림 50. Happy > 0.295 & Surprise > 0.275이 적용된 시각화 결과	50
그림 51. 노드 9 에 대한 최종 Parallel coordinates	51
그림 52. 노드 9 에 최종 포함된 영화 정보	51
그림 53. 군집 3에 대한 최종 Parallel coordinates	51
그림 54. 분할기준이 적용되기 전의 Parallel coordinates시각화	52
그림 55. Boring < 0.045이 적용된 시각화 결과	52
그림 56. Boring < 0.045 & Disgust < 0.035이 적용된 시각화 결과	53
그림 57. 노드 9 에 대한 최종 Parallel coordinates	53
그림 58. 노드 9 에 최종 포함된 영화 정보	53
그림 59. 군집 4에 대한 최종 Parallel coordinates	54
그림 60. 영화 The Big Swindle에 대한 시각화 결과	55
그림 61. 영화 Fast Five에 대한 시각화 결과	55

## 표 목차

표 1. 영화 흥행도 관련 논문	4
표 2. 하효지 외 2명의 연구, “최종 68개의 감정어휘”	10
표 3. 하효지 외 2명의 연구, “감정어휘의 군집화”	11
표 4. 하효지 외 2명의 연구, “최종 선정된 36개의 감정 어휘”	12
표 5. 영화 데이터의 특징	13
표 6. 의사결정나무분석 응용분야	19
표 7. 의사결정나무분석의 종류	20
표 8. 분석에 사용된 데이터의 특징(전체영화)	21
표 9. 전체 영화 집단에 대한 이익도표	22
표 10. 분석에 사용된 데이터의 특징(군집1)	22
표 11. 군집 1 영화 집단에 대한 이익도표	24
표 12. 분석에 사용된 데이터의 특징(군집2)	24
표 13. 군집 2 영화 집단에 대한 이익도표	26
표 14. 분석에 사용된 데이터의 특징(군집3)	26
표 15. 군집 3 영화 집단에 대한 이익도표	28
표 16. 분석에 사용된 데이터의 특징(군집4)	28
표 17. 군집 4 영화 집단에 대한 이익도표	30
표 18. 분석된 집단별 최종마디에 대한 결과	30
표 19. 데이터의 개수가 15개 이상인 장르 별 감정 어휘 분포(평균 값)	37
표 20. 전체 영화의 대표 감정 어휘 별 상관관계 표	38