Journal of KIIT. Vol. 18, No. 10, pp. 101-109, Oct. 31, 2020. pISSN 1598-8619, eISSN 2093-7571 
http://dx.doi.org/10.14801/jkiit.2020.18.10.101

# 마코프 모델을 이용한 채팅 구성원 간의 관계 분석

김소연*, Nogc Tung La**, 권기현***

# Relationship Analysis of Chat Members with Markov Model

Soyeon Kim*, Nogc Tung La**, and Gihwon Kwon***

본 연구는 2019학년도 경기대학교 학술연구비(일반연구과제) 지원에 의하여 수행되었음


## 요 약

인스턴트 메신저는 구성원간의 주요한 상호작용 도구이며, 코로나19와 같은 비대면 시대에는 더욱 그렇다. 현재까지 제안된 대다수 상호작용 분석 기법은 특정 목적 용도나, 오프라인 기준이다. 그러나 인스턴트 메신저 사용자는 온라인에서 특정 목적 없이 일상적인 대화를 나눈다. 본 논문에서는 보상 값이 있는 이산시간 마코프 모델을 이용하여 채팅 구성원 간의 관계 분석 방법을 제안한다. 대화나 댓글을 말뭉치에 따라 분류하여 하나의 상태로 나타내어 마코프 모델을 생성한 후에, 의미적인 정보 추론을 위해 질의를 확률 시제 논리로 표현한다. 모델과 질의를 받아들여 확률 모델 체킹 도구인 PRISM을 이용하여 질의의 값을 추론하고, 그 값이 갖는 의미를 해석하였다. 사례 연구로, 카카오톡 채팅 데이터를 분석하였다. 그 결과, 지금까지의 분석 기법으로 도출하기 힘들었던 "누가 이 그룹에서 대화의 주도권을 갖고 있는가?"와 같은 다양한 의미적인 정보를 도출할 수 있었다.

## Abstract

Instant Messenger plays an important tool for interactions among participants, even more so in the non face-to-face era under COVID-19. To date, most of interaction analysis techniques have some specific purposes or are based on offline. However, the majority of users who uses instant messengers applies them for everyday conversations without specific purposes, which are done online. This paper proposes an analysis method of the relationship among members in a chat group using Discrete Time Markov Model with rewards. Each chat is classified with the corpus and regarded as a state in the generated model. Based on the state, queries are formalized with probabilistic temporal logic. PRISM, probabilistic model checking tool, then takes the model and the queries, and produces the results which gives a meaningful information on relationships among members in the chat group. As a case study, actual chat data from Kakao Talk was analyzed with the proposed method. The result shows that it can derive many meaningful information such as "who takes the initiative in a group?" which was difficult to derive with the previous ones.



Keywords

chatting analysis. discrete time markov model, rewards, probabilistic model checking



* 경기대학교 컴퓨터공학과 석사과정
- ORCID: https://orcid.org/0000-0002-3181-7043
** 경기대학교 컴퓨터공학과 박사과정
- ORCID: https://orcid.org/0000-0001-7874-2782
*** 경기대학교 컴퓨터공학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-8221-4939



· Received: Apr. 30, 2020, Revised: Oct. 19, 2020, Accepted: Oct. 22, 2020


· Corresponding Author: Gihwon Kwon
Dept. of Computer Engineering, Kyonggi University, 154-42, Gwangyosan-ro, Suwon-si, Kyonggi-do, Korea.
Tel.: +82-31-249-9666, Email: khkwon@kgu.ac.kr

# Ⅰ. 서 론

정보화 기기 대중화로 인해서 인스턴트 메신저는 만 6세 이상 인터넷 이용자 95.9%가 사용하는 중요한 상호작용 수단이 되었다[1]. 인스턴트 메신저의 사용이 증가함에 따라 이를 통한 참여자 간의 상호작용에 관한 분석이 필요하다.

본 논문에서는 보상 값을 갖는 이산시간 마코프 모델을 이용하여 인스턴트 메신저를 사용하는 채팅 구성원 간의 관계를 분석한다. 마코프 모델을 이용한 이전의 분석 기법들은 오프라인에서 특정 목적을 갖는 회의 분석을 대상으로 하였다[2][3]. 본 논문에서는 온라인상의 불특정 채팅을 마코프 모델로 분석하려고 한다.

인스턴트 메신저에서 채팅은 주로 댓글로 이루어진다. 댓글이란 현재 화자의 글에 대한 다른 화자의 반응이며, 채팅 분석에서는 댓글 반응 관계 분석이 중요하다. 현재 사용 가능한 인스턴트 메신저 분석 도구는 “대화 참여 비율”, “대화 참여 시간” 등 형식적인 정보만을 제공한다. 이러한 정보로는 “어떤 화자에게 누가 가장 많이 반응하는가?”, “누가 대화의 주도권을 갖는가?” 같은 의미적인 정보를 추론하기 어렵다. 본 논문에서는 보상 값을 갖는 이산시간 마코프 모델을 사용하여 인스턴트 메신저 참여자 간의 반응 관계를 분석하여 의미적인 정보를 추론하여 기존 형식적인 정보를 보완하고자 한다.

분석 과정은 다음과 같다. 첫째, 채팅을 마코프 모델로 모델링하기 위하여 채팅을 말뭉치로 분리하여 채팅에서 이루어지는 각각의 글을 하나의 상태로 표현한다. 둘째, 상태에서 상태로 이동하는 확률을 전이 값으로 갖는 이산시간 마코프 모델을 작성한다. 셋째, 원하는 기댓값을 추론할 수 있도록, 상태 및 전이에다 보상 값을 배정한다. 넷째, 의미적인 정보 추론을 위하여 질의를 확률 시제 논리 언어로 작성한다. 마지막으로, 확률 모델 체킹 도구를 이용하여 이산시간 마코프 모델에 대하여 질의 결과를 구하고, 얻어진 결과를 해석한다.

본 논문에서는 사례 연구로 인스턴트 메신저인 ‘카카오톡’ 데이터를 사용하였다. 카카오톡 선정 이유는 국내 점유율이 가장 높기 때문이다[4]. 실험에 사용된 카카오톡 채팅 데이터를 분석한 결과, 단순히 채팅 참여도가 높은 것과 상호작용이 적극적인 것과는 차이가 있었다. 본 논문에서는 채팅 참여도뿐만 아니라 댓글 반응 관계 분석을 통하여 각 참여자의 참여 적극성을 분석할 수 있다는 점에서 의의를 가진다.

논문 구성은 다음과 같다. 2장에서 배경지식을 설명하고, 3장에서는 카카오톡 데이터로부터 마코프 모델 생성하는 방법과 질의 작성법을 설명하고, 4장에서 분석 결과를 기술한다. 마지막으로 5장에서 결론 및 향후 연구를 보인다.

# Ⅱ. 관련 연구

## 2.1 이산시간 마코프 모델

이산시간 마코프 모델은 확률 과정의 한 기법으로 특정 시점에 시스템이 어떤 상태에 있을 확률을 구한다. 마코프 모델에서 *n*번째 상태는 바로 직전인 *n*-1번째 상태에 의해서만 영향받는데, 이것을 무기억성(memoryless) 성질이라 부른다.

$$\begin{aligned}&\Pr(X_n = j | X_{n-1} = i, ..., X_1 = k, X_0 = s) \\ &= \Pr(X_n = j | X_{n-1} = i)\end{aligned} \quad (1)$$

뿐만 아니라, 시간이 흐름에 따라서 확률값이 변하지 않는 균일성(homogeneous) 성질을 갖는다.

$$\Pr(X_n = j | X_{n-1} = i) = \Pr(X_1 = j | X_0 = i) \quad (2)$$

이들 두 가지 성질 덕분에, 마코프 모델에서는 확률 과정을 상태 기반 관점으로 다룰 수 있다 (마코프 모델에 관한 상세한 설명은 [5]를 참고하기 바람). 여기서는 상태 기반 관점으로 마코프 모델을 다룰 것이며, 마코프 모델 정의는 다음과 같다.

$$M = (S, \mathbb{P}, L) \quad (3)$$

여기서 $S$는 상태 집합이며, $\mathbb{P} : S \times S \rightarrow [0,1]$ 는 상태에서 상태로의 전이 확률을 나타낸 전이확률행렬이다. 이 행렬에서는 모든 행의 합은 1이 되어야

한다. 즉, $\sum_j \mathbb{P}(i,j)=1$ 이다. 또한 $AP$를 단순 명제 집합이라고 할 때, $L:S\rightarrow 2^{AP}$은 단순 명제를 각 상태에 배정하는 함수이다.

### 2.2 마코프 모델에 관한 질의

마코프 모델을 분석하기 위해서 다양한 분석 기법이 사용되고 있으나 본 논문에서는 모델 체킹을 사용한다. 왜냐하면, 모델 체킹(Model checking)은 마코프 모델이 갖는 상태 공간을 전부 조사하는 철저한 기법이기 때문이다[6]. 또한, 모델 체킹 자동화 도구로 PRISM[7]을 사용한다. PRISM은 마코프 모델과 상태에 관한 질의를 받아들여, 질의의 만족성 여부를 판정한다. 질의는 확률 시제 논리라 불리는 pCTL(probabilistic Computational Tree Logic) 언어를 이용하여 표현하는데, 상태에 관한 질의 작성 규칙은 다음과 같다[8].

$$\Phi ::= tt \mid a \mid \neg\Phi \mid \Phi_1 \wedge \Phi_2 \mid P_{\trianglelefteq p}[\varphi] \mid S_{\trianglelefteq p}[\Phi] \tag{4}$$

여기서 $a\in AP$는 단순 명제, $\trianglelefteq \in \{\leq, <, \geq, >\}$는 비교 연산자, $p\in[0,1]$는 확률이며, $P$는 순간 상태 확률(transient state probability) 연산자, $S$는 안전 상태 확률(steady state probability) 연산자이다. 순간 상태 확률은 상태들이 연속적으로 나열된 경로 $\varphi$에 의존하는데, 경로를 작성하는 pCTL 구문 규칙은 다음과 같다.

$$\varphi ::= X\Phi \mid \Phi_1 U^{\leq k} \Phi_2 \tag{5}$$

여기서, $X$와 $U$는 각각 다음(ne**X**t)과 언틸(**U**ntil)을 나타내는 시제 연산자이다. 위의 구문 규칙에서는 최소 연산자만 사용하였다. 경로에서 미래와 항상을 나타내는 $F$, $G$ 연산자는 논리적 동치로 구할 수 있다.

$$F\Phi \equiv tt\, U\, \Phi \tag{6}$$

$$G\Phi \equiv \neg F \neg \Phi \tag{7}$$

위의 구문 규칙으로 마코프 모델 상태에 관한 질의를 논리식으로 작성하면, PRISM은 초기 상태로부터 논리식을 평가한다. 그러나, 때로는 초기 상태가 아닌, 특정 조건을 만족하는 상태들로부터 논리식을 평가할 때 필터 연산자를 사용한다.

$$filter(type, \Phi, cond) \tag{8}$$

$cond$를 만족하는 상태에서 시작하여 논리식 $\Phi$을 처리한 후에 필터 유형인 $type$을 적용한 결과를 돌려준다. 본 논문에서 사용한 필터 유형들이다.

- count: $\Phi$를 만족하는 상태의 개수를 반환한다.
- avg: $\Phi$를 만족하는 상태의 평균값을 반환한다.

### 2.3 보상 값

마코프 모델에서 보상 값은 확률값 이외에 정량적인 정보(quantitative information)를 얻을 때 사용된다[9]. 예를 들어, 채팅에서 "어떤 화자의 글에 누가 얼마나 빨리 반응하는가?" 정보를 얻기 위해서는 마코프 모델에 보상 값을 미리 배정해 놓아야 한다. 보상에 관한 질의는 식 (4)의 상태에 관한 질의를 확장한 것으로서, $R$ 연산자를 사용한다.

$$\Phi ::= \ldots \mid R_{=?}[C \leq k] \mid R_{=?}[F\Phi] \tag{9}$$

보상 질의에서, $R_{=?}[C\leq k]$는 $k$ 시간까지 보상 값의 누적 평균, 즉 기댓값이다. 반면, $R_{=?}[F\Phi]$는 $\Phi$를 만족하기 직전까지의 기댓값이다.

## III. 마코프 모델링 및 질의 작성

### 3.1 채팅 모델

본 논문에서는 카카오톡 채팅을 이산 확률 마코프 모델로 분석한다. 사례로 8명 참여자가 나눈 941개 채팅 데이터를 사용한다. 마코프 모델을 구하기 위해서, 대화를 말뭉치로 표현한다. 말뭉치는 오프라인 회의 분석 용도로 고안된 것을 참고하여[10], 채팅 분석에 맞게 수정하였다. 말뭉치는 참여자, 유형, 반응, 감정의 4-튜플로 구성된다.

- 참여자: 8명을 A~H로 표기
- 유형: 표 1과 같이 14개 유형으로 구분
- 반응: 반응(reaction), 무반응(noreaction)으로 구분
- 감정: 감정(sentiment), 무감정(nosentiment) 구분

예를 들어, <A_inf_noreaction_nosentiment> 튜플은 참여자 A가 무반응, 무감정으로 정보 제공하는 상태를 나타낸다. 또한 <C_eval_reaction_sentiment>는 참여자 C가 이전 대화와 관련하여 감정적 평가를 제시한 상태이다.

표 1. 대화 유형
Table 1. Dialog types

| Types | Meaning |
|---|---|
| back | Words other than words that convey meaning such as self-talk |
| stall | Words to draw attention |
| inf | Words to convey information |
| sug | words that elicit other people's reaction such as suggestions |
| eli | Words that ask for information such as questions |
| ans | Words that answer suggestions, questions, etc. |
| expre | words that responds to someone else's words such as admiration |
| eval | Evaluation of opinions such as consent or rejection |
| off | Additional explanations such as the reason for what you said |
| emo | Words that express feelings such as laughter and cry |
| media | Media such as pictures and videos |
| other | Words not included in the above categories |

반응을 분석하고자 유형 중에서 'sug, eli, ans'를 반응 유도형이라고 분류한다. 채팅의 각 글을 말뭉치로 분류한 4-튜플을 하나의 상태로 간주한 후에, 상태와 상태간의 전이 누적 정보를 이용하여 마코프 모델을 구한다. 뿐만 아니라, 어떤 주제의 채팅이 무한히 계속되는 것이 아니므로 채팅의 시작과 끝을 START와 STOP 특별한 상태로 나타내었다. 이러한 방식으로 위에서 설명한 카카오톡 데이터를 마코프 모델로 변환한 결과 187개 상태를 갖는 모델을 생성했다.

## 3.2 보상 값 배정

확률 추론뿐만 아니라 양적 추론을 위하여 상태 및 전이에 대하여 보상 값을 부여한다. 마코프 모델의 187 상태에 대해 보상 값을 배정할 뿐만 아니라 사용자, 대화 유형, 반응, 감정에 따른 보상 값을 추가한다. 각 분류마다 가중치를 두는 질의가 없어서 보상 값은 모두 1로 설정하였다. 사용된 보상 구조는 다음과 같다:

- <r_Steps>: 모든 전이마다 보상 값 1을 배정한다.
- <r_참여자, r_유형, r_반응, r_감정>: 해당 조건을 만족하는 상태에 도달할 시 보상 값 1을 배정한다.

## 3.3 질의 작성

보상 값을 부여한 마코프 모델에 대하여 의미 있는 결과를 추론하기 위해서는 pCTL 논리식으로 질의를 작성한다. 채팅 분석을 위하여 사용한 질의와 pCTL 표현은 다음과 같다.

Q1. 각 대화 유형이 감정에 미치는 영향
filter(avg, P=?[F<=7 sentiment], "유형")

Q2. 참여자간의 반응
filter(avg, R{"r_Steps"}=? [F "다른참여자"], "참여자")

Q3. 반응 유도형 대화에 대한 참여자간의 반응
filter(avg, R{"r_Steps"}=? [F "다른참여자" & ans],
"참여자" & (sug | eli))

Q4. 대화에 참여하는 시간
R{"r_Steps"}=?[F "참여자"]

Q5. 채팅 초반에 나오는 대화 유형
R{"r_유형"}=? [C<=50]

Q6. 반응 유도형 대화의 감정 여부에 따른 반응
filter(avg, R{"r_Steps"}=? [F ans], sentiment & (sug | eli))
filter(avg, R{"r_Steps"}=? [F ans], nosentiment & (sug | eli))

Q7. 대화 유형별 참여자 비율
S=? [“참여자” & “유형”]

Q8. 참여자 별 반응 비율
S=? [“참여자” & reaction]

Q9. 감정에 따른 참여자 별 반응
S=? [“참여자” & reaction & sentiment]
S=? [“참여자” & reaction & nosentiment]

위의 pCTL 질의에서 이중 인용 부호(“ ”)로 둘러싸인 한글은 비단말 기호여서, PRISM 실행시에는 구체 값으로 치환되어야 한다. 또한, 숫자나 pCTL 키워드 이외에 sentiment등 영문은 해당 상태를 지정해 놓은 레이블(label) 이다.

## Ⅳ. 분 석

### 4.1 모델의 유효성

카카오톡 채팅 분석을 하기에 앞서서 전 장에서 제안한 방법대로 생성된 마코프 모델이 과연 올바른지를 확인해야 한다. 이를 위해서 구글 앱스토의 카카오톡 분석 프로그램 앱인 카톡분석기를 사용하였다[11]. 카톡분석기로 참여자의 채팅 참여 비율을 구하면 ‘C-A-B-D-E-H-F-G’순 이었다. 마코프 모델을 이용한 참여 비율과 기존 분석기로 구한 참여 비율이 동일한지를 통하여 모델의 유효성을 검증한다. 마코프 모델로부터 참여 비율을 얻기 위하여 사용한 질의는 다음과 같다.

S=? [“r_참가자”]

참여 비율은 매 단계의 순간 확률보다는, 장기간 안정 상태 확률이기 때문에 S 연산자를 사용하였다. 질의에 대한 PRISM 실행 결과는 그림 1과 같이 ‘C-A-B-D-E-H-F-G’순 이었고, 이는 카톡분석기 결과와 일치한다.

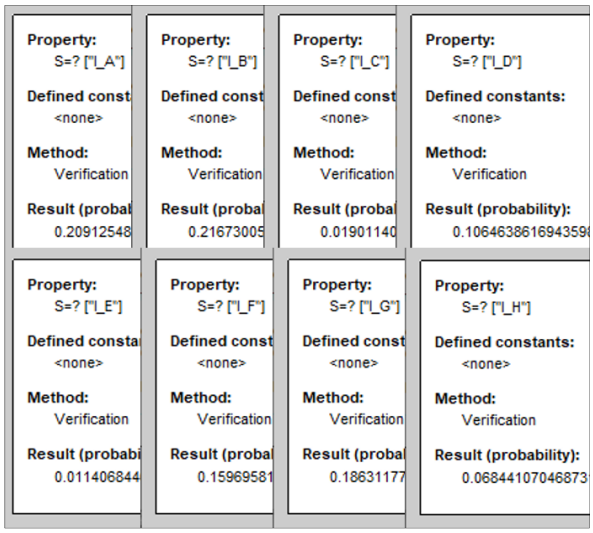



그림 1. PRISM 실행 결과
Fig. 1. Results with PRISM

### 4.2 채팅 구성원 관계 분석을 위한 질의

유효성 확인 후, 마코프 모델로부터 채팅의 의미적인 정보를 추론하고자 앞에서 설명한 9개 질의를 PRISM으로 실행하고 그 결과를 차례대로 분석한다.

1) Q1 질의를 사용하여 채팅 유형이 감정에 미치는 영향을 분석한다. 채팅에는 8명의 참여자가 있다. 현재 화자를 제외한 다른 7명의 반응을 고려하기 위하여 7번째까지라는 제한을 두고, 연산자 P를 사용하여 범위에 해당하는 확률값을 도출한다. 순간 확률보다는 전체 평균을 구하고자 필터 타입인 avg를 사용한다. 결과는 표 2와 같다.

가장 높은 확률을 가지는 채팅 유형은 emo이다. 어떤 참여자가 올린 emo 채팅에 대해서, 일곱 단계 이내에 0.90 확률로 반응한다. 대부분 채팅 유형의 비율이 비슷하다. 이는 해당 집단이 모든 채팅에 대하여 비슷한 정도의 감정을 갖는 것을 의미한다.

표 2. 질의 Q1 결과
Table 2. Results of query Q1

| back | stall | inf | sug | ans | eli |
|---|---|---|---|---|---|
| 0.74 | 0.75 | 0.86 | 0.81 | 0.81 | 0.87 |
| expre | eval | emo | media | off | other |
| 0.80 | 0.85 | 0.90 | 0.75 | 0.73 | 0.67 |

2) Q2 질의를 사용하여 각 참가자마다 다른 참여자에 대한 반응을 확인하고자 연산자 R과 전이에 따른 보상 값 r_Steps를 사용한다. 평균값을 구하고자 필터 타입에서 avg를 사용하였고, 결과는 표 3과 같다.

참여자 A(평균 6.0 단계)와 C(평균 6.6 단계)가 가장 빠르게 반응하고, 참여자 G(평균 96.4 단계)가 가장 느리게 반응한다.

표 3. 질의 2 결과
Table 3. Results of query Q2

| | A | B | C | D | E | F | G | H | Evg. |
|---|---|---|---|---|---|---|---|---|---|
| A | | 7.0 | 6.8 | 4.4 | 5.7 | 8.2 | 4.5 | 5.7 | 6.0 |
| B | 16.9 | | 14.4 | 18.1 | 19.2 | 18.8 | 19.0 | 18.6 | 17.9 |
| C | 6.2 | 4.7 | | 7.2 | 7.4 | 6.8 | 6.5 | 7.5 | 6.6 |
| D | 12.8 | 16.0 | 16.5 | | 13.0 | 14.9 | 11.6 | 13.4 | 14.0 |
| E | 30.4 | 34.9 | 33.3 | 30.0 | | 20.7 | 31.4 | 28.6 | 29.9 |
| F | 47.8 | 48.4 | 45.7 | 48.4 | 41.4 | | 49.0 | 48.4 | 47.0 |
| G | 87.5 | 101.5 | 98.9 | 87.6 | 99.3 | 101.7 | | 98.1 | 96.4 |
| H | 26.8 | 32.2 | 31.2 | 25.5 | 21.9 | 26.9 | 26.3 | | 27.3 |

3) Q3 질의를 통하여 다른 참여자의 반응 유도형 채팅에 대한 각 참여자의 반응을 확인하고자 채팅 유형 중 반응 유도형인 sug 또는 eli을 조건으로 둔다. 또한 반응 응답형 유형인 ans를 조건으로 설정한다. 전체 평균을 구하고자 하므로 avg 필터를 사용한다. 결과는 표 4와 같다.

앞의 Q2 결과 값과 비교하면, 반응 정도가 빠른 참여자 A, C는 반응 유도형 채팅에 적극적으로 참여하고, 반응 정도가 늦은 참여자 G는 반응 유도형 채팅에 소극적으로 참여한다.

표 4. 질의 3 결과
Table 4. Results of query Q3

| | A | B | C | D | E | F | G | H | Evg. |
|---|---|---|---|---|---|---|---|---|---|
| A | | 5.7 | 5.4 | 3.2 | 4.1 | 7.4 | 4.8 | 6.6 | 5.3 |
| B | 16.7 | | 13.4 | 18.6 | 19.2 | 19.1 | 17.6 | 19.4 | 17.7 |
| C | 3.7 | 2.7 | | 7.5 | 6.8 | 7.5 | 3.8 | 7.5 | 5.7 |
| D | 12.3 | 17.6 | 15.9 | | 11.5 | 11.8 | 15.0 | 12.5 | 13.8 |
| E | 34.1 | 34.3 | 34.5 | 30.7 | | 19.3 | 33.9 | 31.4 | 31.2 |
| F | 49.8 | 48.8 | 48.8 | 49.4 | 48.6 | | 47.7 | 49.8 | 49.0 |
| G | 97.5 | 100.2 | 98.5 | 79.5 | 97.6 | 100.3 | | 101.4 | 96.4 |
| H | 31.3 | 31.8 | 31.3 | 28.4 | 17.8 | 25.4 | 32.0 | | 28.3 |

4) Q4 질의를 통하여 채팅 시작 이후 각 참여자의 채팅에 참여하기까지 걸리는 정도를 구하고자 하므로 보상 연산자 R을 사용한다. 참여할 때까지 발생한 전이를 r_Steps 보상 값을 이용하여 구한다. 모델에 대하여 전체 평균값을 구하고자 하므로 필터 타입 중 avg를 사용한다. 결과는 표 5와 같다.

결과는 채팅에 처음 참여할 때까지의 전이의 평균값이다. 상호작용을 많이 하는 참여자 A, C는 빠르게 참여하며, 참여자 G는 매우 늦게 참여한다.

표 5. 질의 4 결과
Table 5. Results of query Q4

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| 5.7 | 19.3 | 7.1 | 12.5 | 22.7 | 50.5 | 94.6 | 29.2 |

5) Q5 질의를 통해 초반부에(채팅 시작 후 50단계 이내) 채팅 유형별 비율을 구하고자 하므로 보상 연산자 R을 사용한다. 또한 새로운 채팅이 반복되므로 이의 평균값을 구하고자 avg를 사용한다. 결과는 표 6과 같다.

채팅 초반에는 정보를 제공하는 inf 유형이 가장 많고, 감정을 표현하는 emo 유형이 다음으로 많다. 이는 정보 제공에 대한 '좋아요' 등의 반응이다.

표 6. 질의 5 결과
Table 6. Results of query Q5

| back | stall | inf | sug | ans | eli |
|---|---|---|---|---|---|
| 3.55 | 1.97 | 11.2 | 5.70 | 4.07 | 3.63 |
| expre | eval | emo | media | off | other |
| 4.93 | 2.56 | 6.77 | 1.47 | 2.33 | 0.51 |

6) Q6 질의를 이용하여 반응 유도형 채팅 유형에 대한 반응이 감정 포함 여부에 따라 달라지는가를 확인하기 위하여 감정의 유무와 반응 유도형 채팅인 sug, eli를 조건으로 설정한다. 반응 유도형 채팅에 대한 답을 확인해야 하므로 이에 해당하는 ans가 등장할 때까지의 전이에 대하여 확인한다. 이는 보상 기반 연산자 R과 보상 값 r_Steps에 의하여 도출된다. 결과는 표 7과 같다. 해당 질의를 통하여 평균적으로 모든 참여자가 감정이 없는 대화의 내용보다 감정이 있는 대화의 내용에 빠르게 반응하는 것을 확인 가능하다.

표 7. 질의 6 결과
Table 7. Results of query Q6

| nosentiment | sentiment |
|---|---|
| 11.27 | 13.46 |

7) Q7 질의를 통해 각 채팅 유형별 참여자 비율을 구한다. 참여자가 모든 채팅에 참여할 수 없기에 안정 상태 확률을 구한다. 표 7은 결과 값을 참여자별 비율로 변환한 것이다.

표 7. 질의 7 참여자별 결과
Table 7. Results of query Q7 with respect to members

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| back | 3% | 5% | 17% | 2% | 1% | 3% | 7% | 1% |
| stall | 1% | 4% | 3% | 0% | 4% | 0% | 0% | 1% |
| inf | 26% | 26% | 24% | 26% | 33% | 3% | 7% | 12% |
| sug | 11% | 15% | 7% | 11% | 12% | 33% | 7% | 4% |
| ans | 13% | 12% | 7% | 8% | 1% | 0% | 20% | 7% |
| eli | 13% | 12% | 7% | 8% | 1% | 0% | 20% | 7% |
| expre | 6% | 2% | 6% | 17% | 25% | 19% | 13% | 24% |
| eval | 5% | 7% | 6% | 6% | 0% | 11% | 7% | 1% |
| emo | 13% | 9% | 12% | 16% | 19% | 31% | 13% | 24% |
| media | 2% | 2% | 1% | 4% | 0% | 0% | 0% | 13% |
| off | 6% | 4% | 8% | 0% | 1% | 0% | 7% | 0% |
| other | 0% | 0% | 2% | 3% | 0% | 0% | 0% | 3% |

대부분의 참여자는 정보를 제공하는 채팅을 자주 한다. 참여자 F는 다른 사람의 반응을 이끌어내는 채팅을 주로하며, 참여자 G는 질의와 대답이 주를 이룬다. 질의의 결과 값을 채팅 유형 별 비율로 변환한 결과는 표 8과 같다.

표 8. 질의 7 대화유형별 결과
Table 8. Results of query Q7 w.r.t. dialog types

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| back | 8% | 9% | 74% | 3% | 1% | 1% | 1% | 1% |
| stall | 9% | 27% | 45% | 0% | 14% | 0% | 0% | 5% |
| inf | 24% | 16% | 34% | 12% | 10% | 0% | 0% | 3% |
| sug | 22% | 20% | 23% | 11% | 8% | 12% | 1% | 3% |
| ans | 32% | 21% | 27% | 10% | 1% | 0% | 4% | 6% |
| eli | 32% | 21% | 27% | 10% | 1% | 0% | 4% | 6% |
| expre | 14% | 3% | 19% | 19% | 18% | 7% | 2% | 17% |
| eval | 21% | 19% | 38% | 11% | 0% | 8% | 2% | 2% |
| emo | 19% | 10% | 28% | 12% | 10% | 8% | 1% | 12% |
| media | 18% | 14% | 9% | 18% | 0% | 0% | 0% | 41% |
| off | 27% | 13% | 56% | 0% | 2% | 0% | 2% | 0% |
| other | 0% | 0% | 55% | 27% | 0% | 0% | 0% | 18% |

8) Q8 질의로 안정 상태에 대하여 참여자의 반응 정도를 수치화한다. 다른 화자의 대화에 적극적으로 반응한다는 것은 채팅 집단 내에서 두 화자의 관계가 적극적이라는 것을 뜻한다. 결과를 비율로 변환하면 표 9와 같다.

다른 참여자에게 가장 많이 반응하는 참여자는 A이다. 이는 Q2의 결과와 일치한다. 다른 참여자에게 가장 적게 반응하는 참여자는 G이다. Q1 결과에서 보듯이 참여자 G는 전반적인 채팅 참여율뿐만 아니라 다른 참여자에 대한 반응도 매우 낮다.

표 9. 질의 8 결과
Table 9. Results of query Q8

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| 26% | 15% | 20% | 13% | 9% | 6% | 2% | 9% |

9) Q9 질의로 감정에 따른 참여자 별 반응을 구한다. 참여자가 항상 채팅에 참여할 수 없기에 연산자 S를 사용하여 안정 상태 확률값을 구한다. 또한 반응과 감정에 대한 상관관계 분석을 위해 이를 조건에 추가하였다. 감정을 포함하지 않는 경우 결과는 표 10과 같으며, 감정을 포함하는 경우 결과는 표 11과 같다.

대부분의 참여자는 감정을 포함하지 않는 채팅에 대하여 더 많은 반응을 보였다. 인스턴트 메신저를 통한 대화의 경우 오프라인에서의 대화에 비하여 상대적으로 다른 사람의 감정을 파악하는 것이 어렵다. 그럼에도 불구하고 인스턴트 메신저에서도 감정이 관계에 영향을 미친다.

표 10. 감정을 포함하는 않는 경우의 질의 9 결과
Table 10. Results of query Q9 with No sentiment

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| 0.08 | 0.06 | 0.20 | 0.04 | 0.024 | 0.010 | 0.00 | 0.03 |

표 11. 감정을 포함하는 경우의 질의 9 결과
Table 11. Results of query Q9 with sentiment

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| 0.04 | 0.02 | 0.03 | 0.02 | 0.02 | 0.014 | 0.00 | 0.03 |

### 4.3 질의 결과 분석

질의에 관한 결과를 기반으로 “누가 이 그룹에서 대화의 주도권을 가지는가?”, 현재 화자에게 누가 가장 많이 반응하는가?”, “누가 평균적으로 가장 많이 반응하는가?”와 같은 정보를 도출할 수 있었다.

첫째, “누가 이 그룹에서 대화의 주도권을 가지는가?”는 질의 4와 질의 7 결과를 통해서 분석된다. 기존 분석 방법으로는 대화의 양이 가장 많은 C가 주도권을 가지고 있다고 분석되었다. 하지만 제안 방법에서는 질의 7의 결과를 통해 C가 대화 참여 비율과 비교하면 상호작용이 발생하는 대화를 적게 하는 것을 확인했다. 이에 비하여 참여자 A는 참여 유도형 대화에 적극적인 것을 확인하였다. 또한 질의 4의 결과, 참여자 A가 대화에 평균적으로 가장 빠르게 참가하고 있다. 즉 A가 채팅에 가장 빨리 참여하여 정보 제공이나 정보 요청을 함으로써, 대화의 주도권을 가지고 있다고 분석되었다.

둘째, “누가 현재 화자에게 가장 많이 반응하는가?”에 해당하는 정보는 질의 2와 3을 통해 확인 가능하며, 그 결과를 정리하면 표 13과 같다. 또한 해당 결과를 통하여 대화에 가장 적극적으로 참여하는 참여자는 A임을 알 수 있다.

셋째, “누가 평균적으로 가장 많이 반응하는가?”는 질의 8, 9번을 통해 확인 가능하다. 질의 8을 통해 참여자 A가 전체 대화 데이터 내에 가장 많은 반응을 보이며, 질의 9를 통하여 감정 여부에 상관없이 참여자 A가 가장 많은 반응을 보이는 것을 알 수 있다.

표 13. 참여자간 반응 정도 결과
Table 13. Results of the degree of response

| Member | Other members |
|---|---|
| A | C |
| B | C |
| C | A |
| D | A |
| E | A |
| F | A, C |
| G | A |
| H | A |

## V. 결론 및 향후 과제

코로나19로 인한 비대면 시대에는 채팅을 통한 상호작용이 더욱 증가할 것이다. 예를 들어, 재택근무할 때 업무 배정이나 업무 모니터링이 채팅을 통해서 진행된다. 그러므로 채팅에 관한 체계적인 분석이 필요하다.

본 논문에서는 보상 값을 갖는 이산시간 마코프 모델을 이용하여 채팅 참여자들의 반응 관계 분석 기법을 제안하였고, 제안 기법의 유효성을 확인하고자 실제 카카오톡 채팅 데이터에 적용하였다. 이를 통하여 온라인 채팅에서 “누가 이 그룹에서 대화의 주도권을 갖는가?”, “누가 현재 화자에게 가장 빨리 많이 반응하는가?”와 같은 정보를 확률 모델 체킹으로 추론할 수 있었다. 우리가 알기로는, 카카오톡 채팅 분석에 확률 모델 체킹을 시도한 이전 연구는 없었다. 이번 연구를 통해서 카카오톡 채팅 데이터를 이산시간 마코프 모델로 모델링하는 방법을 제안하였고, 채팅 분석에 필요한 다양한 속성을 확률 시제 논리로 표현하는 방법을 연구하였다.

여기서는 불특정 집단의 채팅을 분석하였으나, 향후에는 소프트웨어 프로젝트 관리로 확장하고자 한다. 왜냐하면 프로젝트 관리 분석이 채팅의 댓글 분석과 유사하기 때문이다. 이를 통해서, 프로젝트 참여자들의 반응 정도 등을 추론하고자 한다.

## References

[1] M. Noh, Survey on Internet Usage, Ministry of Science and ICT, 2018.

[2] O. Andrei and G. Murray, "Interpreting Models of Social Group Interactions in Meetings with Probabilistic Model Checking", In Proc. of Group Interaction Frontiers in Technology, Oct 2018.

[3] J. Meredith, "Conversation Analysis and Online Interaction", Research on Language and Social Interaction, Vol. 52, No. 3, pp. 241-256, Aug 2019.

[4] Y. Gang, "An communication aspect of ‘Kakao-talk’ conversation and its sociolinguistic feature", Journal of Language and Literature, Vol. 92, pp.

5-37, 2017.

[5] F. Biagini and M. Campanino, Elements of Probability and Statistics, Springer International Publishing, 2016.

[6] M. Kwiatkowska, G. Norman, and D. Parker, "PRISM 4.0: Verification of Probabilistic Real-time Systems", In Proc. of Computer Aided Verification, pp. 585-591, 2011.

[7] E. M. Clarke, T. A. Henzinger, H. Veith, and R. Bloem, Handbook of Model Checking, Springer, 2018.

[8] M. Kwiatkowska, G. Norman, and D. Parker, "Stochastic Model Checking", In Proc. of International School on Formal Methods for the Design of Computer, Communication, and Software Systems, pp. 220-270, May 2007.

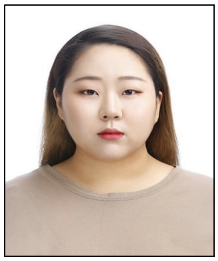
[9] G. Murray, "Markov reward models for analyzing group interaction", In Proc. of ACM International Conference on Multimodal Interaction, pp. 336-340 2017.

[10] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus", Journal of Language Resources and Evaluation, Vol. 41, No. 2, pp. 181-190, 2007.

[11] Kakao Talk Analyzer, https://play.google.com/store/apps/details?id=com.forest.kakao&hl=ko [accessed: Apr 30, 2020]

## 저자소개

### 김 소 연 (Soyeon Kim)



2020년 2월 : 경기대학교 컴퓨터공학과(공학사)
2020년 3월 ~ 현재 : 경기대학교 컴퓨터공학과(석사과정)
관심분야 : 소프트웨어 공학, 정형 검증, 소프트웨어 안전성, 시스템 안전성 분석

### Nogc Tung La

2008년: Thai Nguyen University, Vietnam(공학사)
2010년: Thai Nguyen University, Vietnam(공학석사)
2016년 ~ : 경기대학교 컴퓨터과학과 박사과정
관심분야 : 소프트웨어 공학, 소프트웨어 안전성, 시스템 안전성 분석

### 권 기 현 (Gihwon Kwon)

1985년 2월 : 경기대학교 전자계산학과(이학사)
1987년 8월 : 중앙대학교 전자계산학과(이학석사)
1991년 2월 : 중앙대학교 전자계산학과(공학박사)
1991년 2월 ~ 현재 : 경기대학교 컴퓨터공학부 교수
1999년 ~ 2000년 : 미국 카네기멜론대학 전산학과 연구교수
2006년 ~ 2007년 : 미국 카네기멜론대학 전산학과 연구교수
2014년 ~ 2016년 : 한국정보과학회 소프트웨어공학 소사이어티 회장
관심분야 : 소프트웨어 공학, 정형 검증, 소프트웨어 안전성, 시스템 안전성 분석