creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사 학위논문

나이브 베이지안 분류를 이용한 감성 기반 검색시스템

2022년 2월

서울대학교 대학원

협동과정 바이오엔지니어링전공

김 현 식

나이브 베이지안 분류를 이용한 감성 기반 검색시스템

지도 교수 최 진 욱

이 논문을 공학석사 학위논문으로 제출함
2022년 2월

서울대학교 대학원
협동과정 바이오엔지니어링전공
김 현 식

김현식의 공학석사 학위논문을 인준함
2022년 1월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

초 록

본 논문은 감성이 반영된 알고리즘을 기반으로 검색 순위를 재조정하여 좀 더 개인의 감정에 맞는 검색결과를 추천하는 모델에 관한 연구이다. 2019년 12월 이후로 세계적인 감염증인 코로나-19의 대유행에 의해 OTT(Over-the-top media service)플랫폼 시청 시간이 늘어났지만, 영화 선택을 위해 한참을 고민한다는 넷플릭스 증후군이라 현상이 발생하였다. 따라서, 본 연구는 위와 같은 문제를 해결하기 위해 검색시스템을 이용하는 사용자의 검색 정확도와 만족도를 높이기 위한 감성분석을 통한 사용자의 감성이 반영한 영화검색시스템을 개발하였다.

감성분석이란 자연어처리 기술로써 화남, 우울함, 활기찬, 걱정스러움, 부드러움의 5개의 감성으로 본 연구에서 새롭게 정의하였다. 이러한 정의를 나이브 베이지안 분류의 머신러닝 기법으로 사용하여 기상예보와 영화리뷰 데이터를 감성분석을 하였다.

나이브 베이지안 분류 기법은 단어의 특징들이 모두 독립이라는 가정 아래, 추론 대상의 사전 확률에 새로 들어온 추가적인 정보를 기반으로 사후 확률을 추론하는 통계적 기법이다. 이러한 원리를 이용하여 머신러닝의 학습과 시험 통해 모델을 평가한 결과 영화리뷰와 일기예보 각각 0.84와 0.81의 정확도 성능을 확인할 수 있었다.

나이브 베이지안 분류를 통해 생성된 감성예측점수는 감성벡터 생성모델로 전달되어 감성예측점수에 순위를 매겨 감성벡터로 변환되었다. 변환된 일기예보와 영화리뷰의 감성벡터는 본 연구에서 고안한 검색엔진과 함께 계산되어 감성이 반영된 검색결과를 추천한다. 이때 알고리즘에 감성을 반영하는 정도를 조절하는 감성지수(α)를 도입하여 추후 최적의 성능을 확인

하는 실험을 하였다.

또한 본 연구에서 정보검색에 특화된 루씬 라이브러리를 활용하여 감성이 반영된 검색엔진을 제작하였다. 그리고 실험참여자에게 직접검색에 참여할 수 있도록 영화검색 인터페이스를 포함한 영화검색 프로그램을 개발하였다.

본 연구에서 감성지수를 변경해가며 감성지수에 따른 검색시스템의 검색결과의 영향력에 대한 비교 실험을 하였다. 실험참가자 50명을 대상으로 일반검색엔진 $\alpha = 0$과 감성이 $\alpha = 0.3$과 $\alpha = 0.7$ 으로 반영된 검색엔진의 검색 만족도를 통해 성능을 비교하는 실험을 진행하였다. 본 연구의 실험 결과는 $\alpha = 0.7$일 때, 평균 3.53점, $\alpha = 0.3$일 때, 평균 3.63점, $\alpha = 0.7$일 때, 평균 3.89점의 검색 만족도를 기록하였다. 일반검색엔진과 서로 다른 감성지수를 반영한 감성검색엔진과의 T-test 검정을 통해, 감성지수가 $\alpha = 0.7$로 반영하는 감성검색엔진이 통계적으로 유의한 검색 만족도를 나타냄을 확인하였다.

본 연구에서 개발한 감성을 반영한 검색엔진은 기존에 접근하지 못한 감성의 영역을 기계적인 모델을 사용하여 접근할 수 있을 것이다. 이를 통해 음악, 영화, 쇼핑과 같은 검색플랫폼 등에서 상업적으로 활용될 시 유용할 것으로 기대된다.

또한 본 연구를 통해 검색시스템을 이용하는 사람의 감정상태를 기계적으로 파악할 수 있어 의료분야에서 활용할 가능성이 있을 것이라 예상된다.

주요어 : 자연어처리, 감성분석, 머신러닝, 나이브 베이지안 분류, 정보검색

학 번 : 2019-29690

목 차

1. 서 론 1

1.1. 연구의 배경 및 필요성 1

1.2. 감성분석 5

1.3. 나이브 베이지안 분류 알고리즘 9

2. 연구 목적 11

3. 연구 내용 12

3.1. 전체 시스템의 구성 12

3.2. 데이터 세트 15

3.3. 나이브 베이지안 분류기 20

3.4. 감성이 반영된 검색시스템 27

3.5. 영화 검색 프로그램 33

3.6. 실험 구성 42

4. 결 과 46

4.1. 감성지수의 영향력에 대한 실험 결과 46

4.2. T-test 유의성 검정 48

5. 고찰 및 결론 49

참고문헌 51

Abstract 53

표 목차

표 3.1 실험에 사용된 데이터 세트 16

표 3.2 검색을 위한 데이터베이스 17

표 3.3 나이브 베이지안 분류 성능평가를 위한 데이터 세트 26

표 3.4 나이브 베이지안 분류 성능 21

표 3.5 Task 구성과 내용 43

표 4.1 감성지수에 따른 검색만족도 표 47

표 4.2 감성 지수에 따른 검색성능 만족도 그래프.....47

그림 목차

그림 1.1 Todd Atkins의 감성분석법을 기반으로 만든 5가지 감성 정의 6
그림 1.2 지도학습을 적용하기 위한 데이터 사전작업 예시 7
그림 3.1 전체 시스템 구성 개략도 13
그림 3.2 데이터 전처리 과정 18
그림 3.3 감성분석을 위한 나이브 베이지안 분류와 감성벡터 생성 모델.....21
그림 3.4 감성벡터 생성모델 개략도 24
그림 3.5 감성예측 점수를 감성벡터로 변환 모델.....25
그림 3.6 루씬 라이브러리를 사용한 검색시스템 개략도.....29
그림 3.7 순위 재정렬 시스템 개략도 30
그림 3.8 감성기반 영화검색시스템의 시작화면.....35
그림 3.9 영화검색 시스템의 메인 검색창 36
그림 3.10 영화추천 검색결과 38
그림 3.11 영화 1편당 3개의 리뷰 탭 38
그림 3.12 영화 추천에 대한 실험참가자 만족도 평가.....39
그림 3.13 감성을 반영한 일기에보 선택지 40
그림 3.14 감성을 반영하기위한 일기예보 시청각 자료.....41
그림 3.15 전체 실험구성에 대한 도식도 45

1. 서 론

1.1. 연구의 배경 및 필요성

2019년 12월 전 세계적으로 SARS-CoV-2의 감염증인 코로나-19의 대유행에 의해 사람들은 사회적인 거리두기를 시행하였다[1]. 그로 인해 가정에서 시간을 보낼 수 있는 넷플릭스, 왓챠, 웨이브 등과 같은 OTT(over-the-top media service, 온라인 동영상서비스)플렛폼 이용량이 늘고 있다. 그러나, 사용자가 흥미를 끌 수 있는 작품들이 추천되지 않아 시간을 낭비하거나 흥미가 떨어져서 해당 플렛폼을 종료한다는 신조어인 넷플릭스 증후군(netflix syndrome)을 겪는 사용자들이 늘고 있다[2]. 그리고 방대한 콘텐츠가 오히려 심리적 부담감이나 피로감을 유발한다.

본 연구는 일반적인 검색시스템에서 사용자들의 주변여건에 따른 검색기능을 가능하게 하여 검색 시 원하는 콘텐츠를 더 편하게 검색할 수 있는 기능을 제공하는 것에 초점을 맞추었다.

일반적인 검색 시스템은 사용자가 필요로 하는 정보를 수집하여 내용을 분석한 뒤 찾기 쉬운 형태로 보관하였다가, 정보에 대한 검색요청이 발생할 경우 해당 정보를 찾아 제공한다. 그러나 정보의 양이 기하급수적으로 증가하면서 적합한 정보를 찾는 것이 점차 어려워지고 있다.

수많은 정보 중에서 사용자의 수요를 파악해 필요로 하는 정보를 찾기 위하여 다음과 같은 검색기법들이 사용되고 있다.

- **클릭률을 이용한 추천 알고리즘**

추천 알고리즘은 사용자의 이전 클릭률(click through data)을 기반으로 검색 기록(search history)을 사용하여 사용자가 좋아할 거라고 예측되는 항목을 추천한다[3]. 그러나 이러한 검색 추천 시스템은 이전 검색 기록에만 의존하기 때문에 새로운 장르나 다양한 특징의 컨텐츠를 반영하지 못할 뿐만 아니라 검색 사용자의 수요를 실시간으로 파악하지 못할 수 있다. 따라서 우리는 감성적인 요소가 가미된 검색시스템이 필요하다고 판단하여 연구를 시작하게 되었다.

- **PageRank 알고리즘의 이용**

페이지랭크(PageRank)는 검색엔진에 의해 찾아지는 수많은 웹페이지 간의 연결관계에 대한 중요도를 기반으로 유용한 정보를 제공해 주는 정밀한 알고리즘이다[4]. 중요도는 다른 페이지에서 연결되는 링크의 숫자를 정규화(normalize)하여 확률적으로 인용관계의 웹페이지를 수치로 계산한다. 예를 들면, 신뢰를 얻고 있는 웹페이지에 링크를 걸면 페이지랭크 수치가 올라가기 때문에 영향력이 있는 페이지를 인용할수록 검색 순위가 올라간다. 페이지랭크 알고리즘을 통해 수많은 관련된 웹페이지 중에서 사용자들이 진정 관심있는 적합한 정보를 찾을 수 있다.

- **관련 연구**

1997년 구글의 페이지랭크라는 혁신적인 알고리즘의 등장으로 웹 검색과 함께 검색분야에서 엄청난 패러다임이 일어났다. 검색엔진의 속도와 정확도가 급속도로 발전하여 검색은 어디에서나 존재하였다. 그러나 검색의 패러다임 덕분에 인터넷 정보는 셀 수 없을 정도로 급증하였다. 사용자가 능동적으로 수많은 정보 중에서 키워드를 구체적으로 미리 알고 검색을 한다는 것은 쉽지 않았다. 반면, 추천은 '협업 필터링(collaborative filtering)' 과 '콘텐츠 기반 필터링(content based filtering)과 같은 사용자의 이전 행위에 대한 기록을 바탕으로 사용자에게 관심이 있을 것으로 예측되는 정보를 제공해 준다[5, 6].

- 감성기반 검색시스템의 필요성

최근 기술 혁신으로 인해 방대한 양의 빅데이터가 생성되고 있다. 그러나, 그 많은 정보들 중에서 진정한 가치 있는 정보를 가려내는 것은 쉽지 않다. 따라서 효율적인 검색을 위해 사용자의 검색 요구를 잘 이해할 수 있는 정밀한 검색모델이 필요하다. 위에 언급한 클릭률과 PageRank 알고리즘 등은 사용자의 요구, 주변 조건에 따른 요구 변화를 잘 반영하기 힘들다. 주변 조건에 따라서 찾고 싶은 대상의 우선 순위가 달라질 수 있다. 예를 들어 김연경 선수가 일본을 이겼을 때와 흐린 날씨가 계속되어 기분이 가라앉아 있을 때 보고싶은 영화는 동일한 키워드를 넣는다 하더라도 추천이 달라야 한다는 것이 이 연구의 motivation이다. 우리는 이러한 아이디어로 감성기반 시스템을 설계하고, 사용자 주변의 조건, 예를 들면 날씨, 즐거운 뉴스, 주변 분위기에 따라 검색 순위를 조정해 주는 검색모델을 연구하게 되었다.

의료에서는 감성기반 검색시스템의 자료를 활용하여 사용자의 감정상태를 알아내는 것도 가능할 것이다. 예를 들어 김연경 선수가 우승을 한 상황에서도 계속해서 다른 사람의 선택과 달리 우울한 영화를 주로 선택한다든가, 다른 사람들이 매우 기쁠 때 주로 선택하는 영화들을 평소에도 많이 선택한다면 위의 두 경우의 사용자의 감정상태는 한쪽으로 지나치게 치우쳐 있는 것을 알 수 있다. 이와 같이 감성을 이용한 검색시스템을 개발하면 그 간 기계적 모델을 이용해서 검색을 이용하는 모델에 비해서 접근하지 못했던 부분에 대해서 정보를 활용할 수 있을 것이다.

1.2. 감성분석

감성 분석(sentiment analysis)이란 텍스트에 들어있는 의견이나 감성, 평가, 태도 등의 주관적인 정보를 컴퓨터를 통해 분석하는 과정이다. 자연어 데이터에 들어있는 감성을 분석하는 일은 오래 전부터 연구되었지만, 언어가 가지고 있는 모호성 때문에 쉽지 않다[7]. 그러한 모호함을 극복하기 위해서 본 연구에서 개발한 나이브 베이지안 분류(naïve bayesian classification)모델을 이용하여 감성을 분석하였다. 이 분류 모델은 머신러닝(machine learning)의 일종인 지도학습법(supervised learning)을 기반으로 동작한다.

감성분석 기법에는 여러가지 기법이 존재한다. 첫번째로 긍정 또는 부정으로 이진 분류하는 극성 분석법과 두번째로 긍정과 부정 사이에 아라비아 숫자를 이용해 감성의 정도를 수치로 나타내는 척도분석법 있다. 마지막으로 분노, 기쁨, 놀람 등으로 감성의 상태를 미리 정의하여 감성을 분석하는 방법인 상태분석법을 사용하여 본 연구의 감성분석을 하였다.

- **본 연구에서 사용한 감성분류 기법**

본 연구를 위해서 심리치료사 Todd Atkins의 6가지 감성에 대한 카테고리 분류법을 활용하였다[8]. 이러한 분류법은 앞서 설명한 감성의 상태를 미리 정의하여 감성을 분류하는 상태분석법을 사용하였고, 인간의 감성을 균등하게 표현하기 위해 세부 감성을 포괄할 수 있는 대표감성 6가지로 감성의 카테고리를 정의하였다. 이때, 감성을 대표하는 감성은 행복한, 흥분된, 부드러운, 두려운, 화남, 슬픔(happy, excited, tender, scared, angry, sad)의 6가지 감성으로 정의하였다.

이를 바탕으로 본 연구에서 일기예보와 영화리뷰의 데이터 특성에 맞게 수정하여 인간의 감성을 5가지 화남, 우울함, 활기찬, 걱정됨, 부드러움(angry, depressed, energetic, nervous, tender)의 감성을 대표하는 카테고리로 정의하였다(그림1.1).

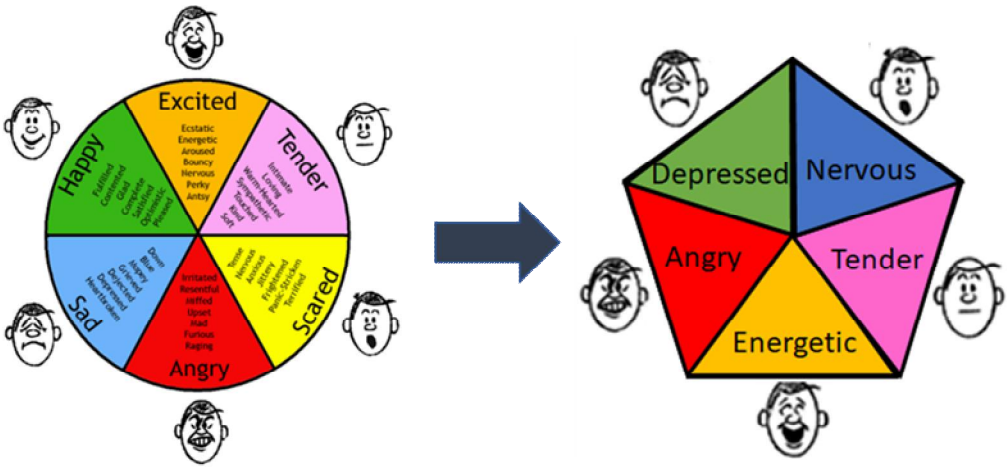


그림 1.1 Todd Atkins의 감성분석법을 기반으로 만든 5가지 감성 정의

본 연구에서는 머신러닝의 지도학습 방법 적용하여 감성의 상태를 미리 정의하였다. 지도학습은 학습모델에 새로운 데이터가 입력되면 이전에 학습한 데이터의 특징을 파악한다. 그리고 미리 정의된 감성을 기반으로 새로운 데이터가 어느 감성 카테고리에 포함되어 있는지 특징을 파악하여 분류한다. 우리는 이러한 접근법을 사용하기 위해 사전에 아래와 같은 단계로 데이터 사전 작업하였다.

본 연구에서는 감성분석을 위해 3단계 사전 작업을 하였다. 첫번째 단계는 기상청에서 제공하는 일기예보와 IMDb 영화검색 사이트에서 데이터 정보를 '수집' 하였다. 두번째 단계는 텍스트 문서에 감성을 나타내는 단어를 찾고 5가지 카테고리로 '감성사전'을 제작하였다. 세번째 단계는 준비된 텍스트 문서 읽고 본 연구에서 정의한 5가지 감성으로 문서를 '분류' 하였다(그림 1.2).

위와 같은 단계를 통해 주관성이 포함된 영화리뷰와 날씨에 대한 감성을 느낄 수 있는 단어가 포함된 일기예보를 우리가 정의한 카테고리로 분석할 수 있다.

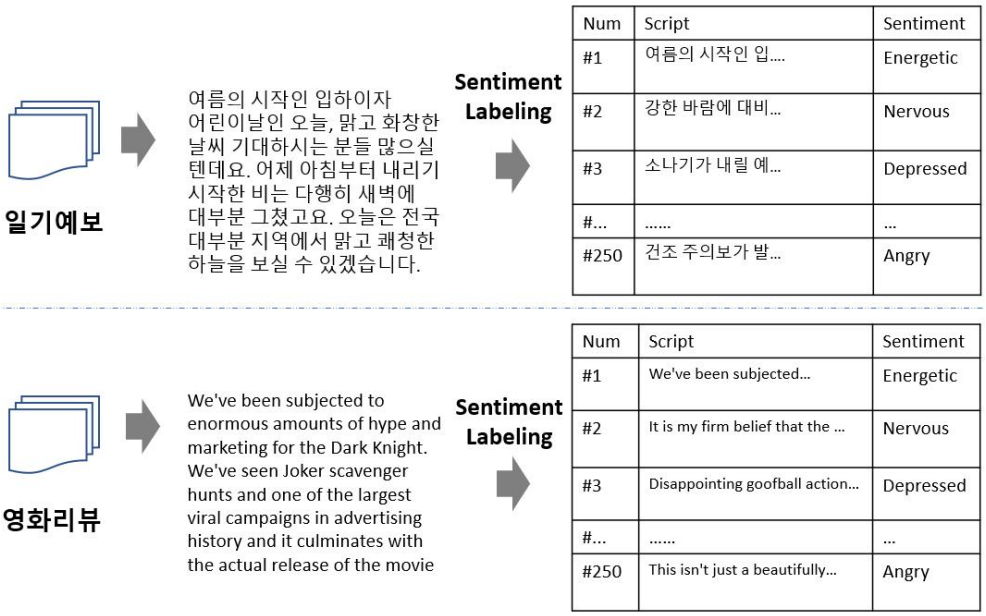


그림 1.2 지도학습을 적용하기 위한 데이터 사전작업 예시

1.3. 나이브 베이지안 분류 알고리즘

나이브 베이지안 분류 알고리즘은 추론 대상의 사전 확률과 추가적인 정보를 기반으로 해당 대상의 사후 확률을 추론하는 통계적 방법이다. 이때 모든 사건은 독립사건으로 가정하여 조건부 확률에 대한 베이즈 정리(baye's theorem)가 사용되고 수식 1.1은 다음과 같다.

$$\boldsymbol{P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B)}} \qquad 1.1$$

위 수식(1.1)에서 $P(A_k)$는 사전확률(prior probability)로 이미 알고 있는 사건의 확률이며, $P(B|A_K)$ 는 A_k 사건이 일어났을 때 B 가 발생할 확률인 조건부확률이다. $P(A_K|B)$ 는 사후확률로 사전확률과 조건부확률을 통해서 알게 되는 사후확률을 의미한다. 따라서 베이즈 정리는 사전확률과 조건부확률을 알 수 있다면 다른 사건이 발생했을 때 일어날 수 있는 확률인 사후확률을 알 수 있다는 것을 의미한다[9].

나이브 베이지안 분류는 모든 사건을 독립이라는 순진한(Naïve)한 가정하에 수식과 의미적으로 단순화하여 계산을 용이하게 만드는 알고리즘이다. 따라서 조건부 독립성은 수식적으로 다음 수식1.2 과 같이 나타낼 수 있다

($i \neq j, k, l$ 라고 가정한다).

$$\begin{aligned} &P(b_i|A_k, b_j) = P(b_i|A_k), \\ &P(b_i|A_k, b_j, b_k) = P(b_i|A_k), \\ &P(b_i|A_k, b_j, b_k, b_l) = P(b_i|A_k) \end{aligned} \qquad 1.2$$

위 수식1.2 와 같이 독립을 가정한다면, 수식 1.3과 같은 조건부확률의 복잡한 연쇄법칙을 계산을 수식1.4 과 같이 확률에 대한 곱으로 간단하게 계산할 수 있다[10].

$$
\begin{aligned}
P(A_k, b_1, \dots, b_n) &= P(A_k)P(b_1, \dots, b_n|A_k) \\
&= P(A_k)P(b_1|A_k)P(b_2, \dots, b_n|A_k, b_1) \\
&= P(A_k)P(b_1|A_k)P(b_2|A_k, b_1)P(b_2, \dots, b_n|A_k, b_1, b_2) \\
&= P(A_k)P(b_1|A_k)P(b_2|A_k, b_1) \dots P(b_2|A_k, b_1, b_2, b_3, \dots, b_{n-1})
\end{aligned} \tag{1.3}
$$

$$
\begin{aligned}
P(A_k, b_1, \dots, b_n) &\propto P(A_k, \dots, b_n) \\
&P(A_k)P(b_1|A_k)P(b_2|A_k)P(b_3|A_k) \dots \\
&P(A_k)\prod_{i=1}^{n} P(b_i|A_k)
\end{aligned} \tag{1.4}
$$

따라서 순진한(Naïve)한 독립성은 복잡한 문제 해결에서 예측과 추론에 대한 강점을 갖는다. 머신러닝의 기법 중에 하나인 나이브 베이지안 분류는 지도학습(supervised learning)모델로서, 특히 텍스트 분류에 강점을 갖는다.

2. 연구 목적

본 연구는 감성이 반영된 알고리즘을 이용하여 검색 순위를 재조정하는 검색 방식의 방법의 효과에 대해서 연구하였다. 주변의 여건에 따라서 검색순위가 다르게 재조절 될 수 있다면 그때의 주변 사건, 날씨 등 개인의 감성에 영향을 주는 사실들을 검색에 반영하여 검색을 진행하면 보다 개인의 감정에 맞는 검색결과를 보여줄 수 있을 것이라는 아이디어에 기초하였다[11].

이를 위하여 나이브 베이지안 분류모델을 이용하여 감성분류를 진행하고, 검색모델에서 이를 주변조건에 맞는 감성에 대한 영향력을 반영하여 기존의 검색엔진에서 나온 결과에 대해서 재정렬하는 방식으로 결과를 재조정하고 이를 실제 사용자들을 이용하여 평가를 진행하였다.

3. 연구 내용

3.1. 전체 시스템 구성

이 장에서는 본 연구 시스템의 전체 구성에 관해 기술한다. 본 연구에서 제안하는 나이브 베이지안 분류를 이용한 감성 기반 검색시스템은 개략적으로 그림 3.1과 같다.

전체 시스템 크게 나이브 베이지안 분류기와 검색엔진으로 구성되어 있다. 나이브 베이지안 분류기는 일기예보와 영화 데이터를 입력 받아 감성분석의 결과물로 감성벡터를 생성하였다. 생성된 일기예보 감성벡터와 영화리뷰 감성벡터는 그림 3.1의 두 화살표와 같이 감성분석 단계를 빠져나와 벡터를 전달한다. 일기예보 벡터는 사용자의 날씨에 대한 감성을 반영하기 위해서 사용되었고, 영화리뷰 벡터는 검색엔진이 검색에 사용될 데이터베이스에 추가되었다. 검색엔진 단계에서, 사용자에 의해 검색 쿼리가 입력되면 검색엔진에 의해 영화순위와 감성벡터가 순위 재정렬 시스템으로 전달된다. 사용자의 감성을 반영하는 오늘의 날씨를 선택하고 검색 쿼리를 입력하면 영화순위와 날씨감성벡터 그리고 영화감성벡터들이 본 연구에서 고안한 알고리즘에 의해 순위가 재정렬 된다. 따라서 사용자 인터페이스에 재정렬된 순위에 따른 영화가 추천되어 보인다.

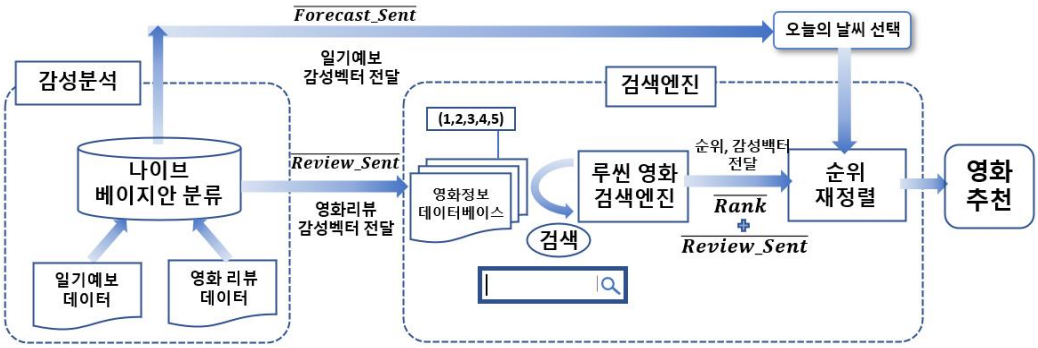


그림 3.1 전체 시스템 구성 개략도

- 나이브 베이지안 분류기 구성

본 연구에서 개발한 나이브 베이지안 분류기는 감성을 반영하기 위해 다음과 같이 구성하였다. 본 연구에서 고안한 나이브 베이지안 분류 알고리즘을 사용하여 일기예보와 영화리뷰 데이터를 5가지 감성으로 감성 예측 점수를 매긴다. 이러한 점수에 대한 순위를 정해 1위부터 5위까지 예측 순위를 정해 5차원 감성벡터로 변환한다. 영화리뷰 감성벡터 $\overline{\boldsymbol{Review_Sent}}$는 검색엔진에 의해 검색될 데이터 베이스에 전달되고, 일기예보 감성벡터 $\overline{\boldsymbol{Forcast_Sent}}$는 사용자 인터페이스에서 사용자의 감성을 반영하기 위해 전달된다.

- 검색엔진 구성

본 연구에서 개발한 검색엔진은 감성을 반영하기 위해 다음과 같이 구성하였다. 나이브 베이지안 분류기를 통해 생성된 영화리뷰 감성벡터 $\overline{\boldsymbol{Review_Sent}}$를 영화정보 데이터 베이스에 연결한다. 이러한 데이터 베이스를 빠르게 검색하기 위해 특별한 형태로 저장하는 역색인(inverted index)구조 변환하여 데이터베이스를 준비한다. 사용자에 의해 검색엔진에 쿼리가 입력되면, 검색엔진에 의해 영화정보와 $\overline{\boldsymbol{Review_Sent}}$를 찾게 되고 검색순위$\overline{\boldsymbol{Rank}}$ 와 함께 순위 재정렬 시스템으로 전달된다. 실험을 위한 영화 검색프로그램에서 사용자의 감성을 반영하는 날씨를 입력하도록 설계하였다. 이를 통해 일기예보에 대한 감성벡터 $\overline{\boldsymbol{Forcast_Sent}}$ 또한 순위 재정렬 시스템으로 전달되어 본 연구에서 고안한 알고리즘에 의해 순위가 재정렬 되어 사용자의 감성을 반영한 영화가 추천된다.

3.2 데이터 세트

본 연구에 사용된 데이터 세트는 크게 감성분석을 위한 데이터 세트와 검색시스템 구축을 위한 데이터베이스로 구성되었다. 그리고 감성분석을 위한 데이터 세트는 실험에 직접 사용될 데이터 세트와 나이브 베이지안 분류 모델의 성능을 평가하기 위한 데이터 세트로 나뉘었다.

감성을 분석을 위한 데이터는 2020년 대한민국 기상청에서 제공하는 일기예보의 텍스트 형식 문서를 수집하였다[12]. 이 문서는 감성분석을 할 때 중요하지 않는 기상청 프로그램에 대한 소개와 아나운서 인사말 같은 내용은 미리 제거하였다. 그리고 IMDb 영화검색 사이트에서 선정한 2020년 기준 인기 있는 영화의 리뷰를 수집하였다[13].

검색을 위한 데이터베이스 또한 IMDb 영화검색 사이트에서 인기있는 영화 1,100개를 선정하여 영화정보 데이터 세트를 수집 하였다.

3.2.1. 나이브 베이지안 모델을 위한 데이터 세트

본 연구에서 감성분석을 위해 실험에 사용될 데이터는 255건의 일기예보 문서와 3,550건의 IMDB 영화리뷰 문서이다.

여기서 일기예보의 250건은 나이브 베이지안 분류 모델의 학습에 사용되었고 5건은 실험에서 감성이 반영하는 오늘의 날씨로 사용된다. 5개의 일기예보는 실험참가자가 오늘의 날씨를 봄, 여름, 가을, 겨울, 재난의 중에 선택할 수 있도록 구성하였다. 이러한 선택지는 검색엔진에 날씨에 대한 감성을 반영하기 위해 사용되어진다.

영화리뷰의 250건은 나이브 베이지안 분류 모델의 학습에 사용되고 3,300건은 감성분석 통해 검색을 위한 데이터베이스로 사용된다. 여기서 3,300개의 영화리뷰는 동일한 영화의 1편당 3개의 서로 다른 리뷰이며 1,100개의 영화로 구성 되어있다(표 3.1). 감성분석 과정 통해 생성된 영화리뷰에 대한 감성벡터는 IMDB의 영화정보 데이터베이스와 합쳐져서 검색시스템을 위해 사용되었다.

문서 종류	데이터 세트	문서 수	문서 길이
영화 리뷰	훈련세트	250	350
	실험세트	3,300	350
일기예보	훈련세트	250	1,800
	실험세트	5	1,800

표 3.1 실험에 사용된 데이터 세트

3.2.2. 검색시스템 구축을 위한 데이터베이스

검색엔진은 검색을 수행하기 위해 데이터베이스가 필요하다. IMDb 영화검색 사이트에서 수집한 영화정보 데이터는 다음 표 3.2와 같이 구성하였다. 그리고 앞서 설명한 3,300개의 영화리뷰와 영화리뷰에 대한 감성벡터 3,300건이 데이터베이스에 추가되었다. 영화리뷰에 대한 감성벡터 3,300건은 감성분석을 통해 생성된 결과물이다.

IMDb 영화정보		
데이터 필드	**설명**	**데이터 수**
title	제목	1,100
startYear	상연 연도	1,100
runtimeMinutes	상영 시간	1,100
genres	장르	1,100
p_direc_name	감독 이름	1,100
p_actor	주연배우 이름	1,100
averageRating	평균 평점	1,100

추가

영화 1편당 3개의 리뷰		
Sentiment_vector	**영화리뷰 감성벡터**	3,300
Review	영화 리뷰	3,300

표 3.2 검색을 위한 데이터베이스

결과적으로 앞서 설명한 과정을 통해 만들어진 표 3.2는 검색엔진이 검색을 위한 데이터 베이스로 사용되었다.

3.2.3. 텍스트 전처리

감성분석을 위한 데이터의 텍스트는 다음 그림 3.2과 같이 전처리 작업으로 처리하였다. 전처리 작업에 사용한 데이터는 일기예보와 영화리뷰 데이터에 해당한다. 두 데이터 모두 구어체로 작성되어 어휘의 생략이나 축약하는 경우가 많았으나, 특히 영화리뷰는 문자가 아닌 것들과 이모티콘 그리고 인터넷 주소 등이 포함되어 있어 전처리를 하였다.

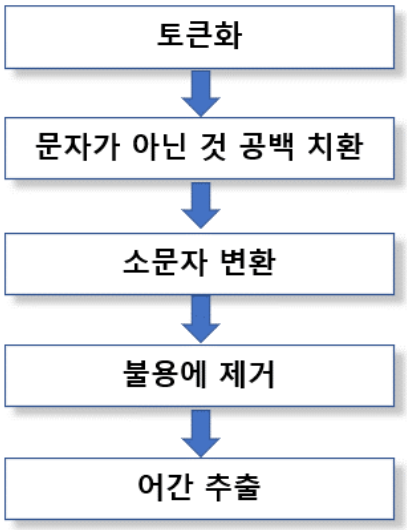


그림 3.2 데이터 전처리 과정

토큰화: 나이브 베이지안 분류 모델은 단어의 출현 빈도수를 기반으로 계산이 하기 때문에 unigram을 사용하여 단어를 토큰화 하였다.

문자가 아닌 것을 공백으로 치환: 컴퓨터가 텍스트를 이해할 수 있도록 정규표현식으로 영문자나 한글문자가 아닌 문자는 모두 공백으로 치환하였다.

소문자 변환: 본 연구에서 대문자와 소문자의 구별로 인해 서로 다른 문자로 인식되는 것을 방지하기 위해 영어의 모든 대문자는 소문자로 변환하였다.

불용어 제거: 일반적으로 텍스트 문서에서 자주 나타나는 단어는 학습 모델로서 학습이나 예측 실제로 기여하지 않아 제거하였다.

어간추출: 텍스트의 문서에 포함된 단어의 수를 줄일 수 있는 기법인 어간추출을 하였다. 굴절되거나 변형된 단어를 일반화를 한다면, 단어의 빈도수를 기반으로 하는 나이브 베이지안 모델의 성능을 높일 수 있었다.

3.3. 나이브 베이지안 분류기

다음 그림 3.3은 본 연구에서 고안한 나이브 베이지안 분류기와 감성벡터 생성모델을 통해 일기예보와 영화리뷰의 문서에 대한 감성분석의 전체과정을 나타낸다.

나이브베이지안 분류 모델은 그림3.3과 같이 학습(training)과 시험(test)과정으로 구성되어 있는 머신러닝의 지도학습 방식을 따른다.

모델 훈련 단계에서 나이브 베이지안 분류기는 단어의 출현 빈도수를 카운트하여 훈련데이터의 특징을 파악하고, 시험 단계에서 새로 입력된 데이터의 감성예측 점수를 매겨 수치적으로 분석한다. 그리고 분석된 점수는 감성 벡터 생성모델을 통해 1위부터 5위까지 순위를 매겨 5차원 벡터 형태로 변환한다. 이렇게 생성된 감성벡터는 순위 재정렬 시스템으로 전달되어 본 연구에서 고안한 알고리즘에 의해 검색순위를 재정렬 하게 된다.

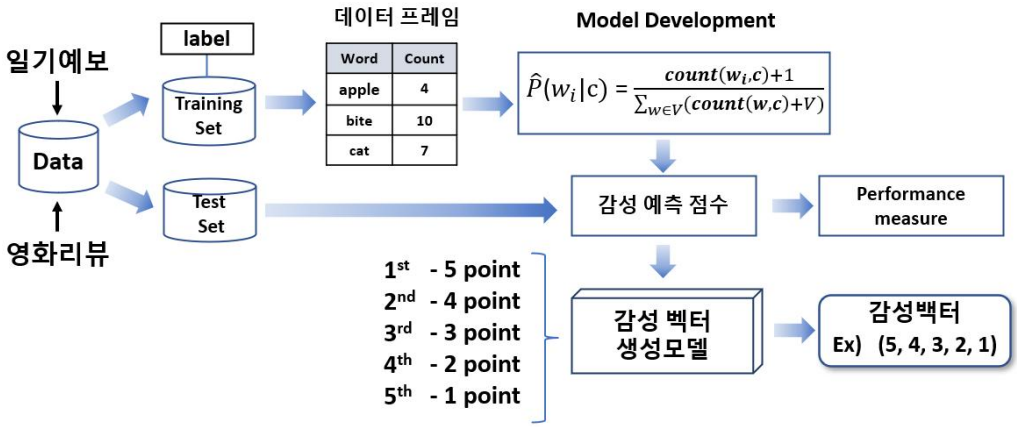


그림 3.3 감성분석을 위한 나이브 베이지안 분류와 감성벡터 생성 모델

3.3.1. 나이브 베이지안 분류 알고리즘

나이브 베이지안 분류 과정은 추론 대상의 사전 확률과 추가적인 정보를 기반으로 해당 대상의 사후 확률을 추론하는 머신러닝의 지도학습 방식을 따른다. 따라서, 데이터의 특징을 사전에 학습하는 방법으로 사전확률을 계산해야 한다. 5개의 감성 카테고리로 지정된 문서를 단어의 출현 빈도수를 카운트하여 변수로 저장하여 모델을 학습시켰다. 단어를 세고 저장할 때, 단어의 빈도수 기반의 확률 모델을 사용하기 위해 bag of word 방식을 사용하였다. 이를 통해 단어와 출현 빈도를 새고 저장할 수 있는 데이터 프레임을 제작하였다[14].

다음 수식 3.1~3.3은 본 연구에서 사용한 나이브 베이지안 분류 모델 알고리즘이다.

$$\hat{P}(c) = \frac{N_c}{N} \tag{3.1}$$

$$\hat{P}(w_i|c) = \frac{count(w_i,c)}{\sum_{w \in V} count(w,c)} \tag{3.2}$$

$$\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum_{w \in V}(count(w,c)+V)} \tag{3.3}$$

수식(3.2)는 앞선 1.3절에서 설명한 수식(1.4)에서 유도된 알고리즘이다. 그러나, 훈련 데이터에 없던 값이 입력될 때 그에 대한 $\hat{P}(w_i|c)$가 '0'이 되어 정상적인 분류가 되지 않는 zeros problem이 발생하였다.

이를 방지하기 위해 수식 3.3 과 같이 라플라스 스무딩(laplace smoothing)기법을 사용하였다. 이것은 훈련 데이터에서 한번도 출현하지 않은 단어(unknown word)를 추가로 삽입 하는 방법으로 zeros problem 문제를 해결 할 수 있었다. 여기서 변수 V 는 훈련 데이터에서 나오고 중복이 없는 전체 단어의 수 이다. 이어서, 문서의 감성을 예측하기 위해 $\hat{P}(c) \times \hat{P}(w_i|c)$의 값을 계산하면 각 감성의 확률은 1 보다 작은 값이 나오게 된다. 이 현상은 곱셈의 항목이 많아지면 소수점 자리가 너무 작아져서 감성의 구별이 어렵거나 컴퓨터의 계산 범위를 넘어서는 언더플로우(underflow)문제가 발생한다. 우리는 $\hat{P}(c) \times \hat{P}(w_i|c)$ 의 값에 Log 를 취해 곱셈계산식에서 덧셈계산식으로 변환하여 언더플로우를 방지하였다(수식 3.4 ~ 3.5) [14].

$$Log\big(P(c|w_i)\big) = Log(\hat{P}(c) \times \hat{P}(w_i|c)) \qquad 3.4$$

$$= Log\left(\hat{P}(c)\right) + Log\left(\hat{P}(w_1|c)\right) + \cdots + Log(\hat{P}(w_n|c)) \qquad 3.5$$

최종적으로 완성된 나이브 베이지안 알고리즘으로 일기예보와 영화리뷰의 데이터를 사용하여 감성을 예측하는 확률을 구했다. 이 예측값은 Log를 사용해 언더플로우를 방지하였고 이를 통해 데이터 문서마다 5가지 감성에 대한 예측 점수를 도출 할 수 있었다. 감성에 대한 예측 점수는 다음 절에서 제안한 감성벡터 생성모델로 전달되었다.

3.3.2. 감성벡터 생성모델

감성벡터 생성모델 과정은 이전 나이브 베이지안 분류 모델에서 전달받은 감성에 대한 예측점수 값을 전 받아 감성벡터를 생성한다(그림 3.4). 5차원 감성벡터로 변환하는 이유는 순위 재정렬 시스템에서 알고리즘 계산의 용이함을 위해서 변환한다.

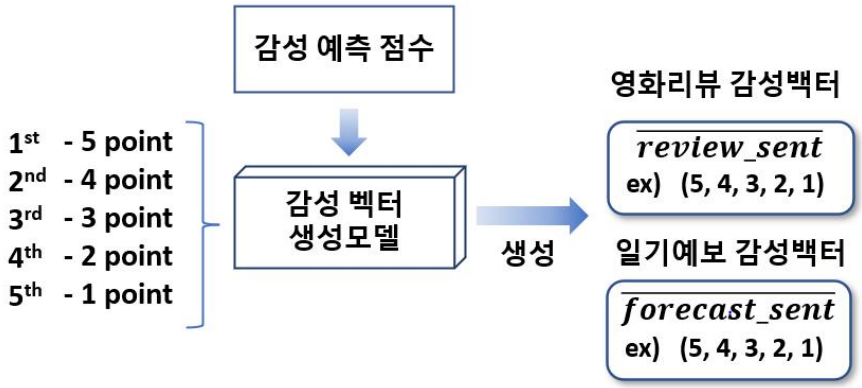
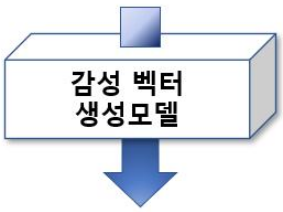


그림 3.4 감성벡터 생성모델 개략도

이러한 감성예측 점수 값은 문서 한 건당 5개의 감성에 대한 수치를 나타내었고, 그림 3.5과 같이 5가지 감성을 angry, depressed, energetic, nervous, tender 순으로 구성하였다. 이때 언더플로우 방지를 위해 Log를 적용했기 때문에 음수로 결과값이 나오고 0에 가까울수록 예측한 감성 카테고리에 근접하다고 해석할 수 있다.

Angry	Depressed	Energetic	Nervous	Tender
-0.12	-0.45	-0.97	-0.54	-0.27



Angry	Depressed	Energetic	Nervous	Tender
5	3	1	2	4

그림 3.5 감성예측 점수를 감성벡터로 변환

예측점수가 큰 값을 기준으로 1등은 5점, 2등은 4점, 3등은 3점, 4등은 2점, 5등은 1점을 적용해 감성벡터를 생성하였다(그림 3.5).

이렇게 생성된 일기예보에 대한 감성벡터는 사용자의 감성을 날씨로 반영하기 위해 사용되었다. 영화리뷰에 대한 감성벡터는 검색시스템의 데이터 베이스로 추가되어 검색엔진에 의해 사용되었다.

3.3.3. 나이브 베이지안 분류기의 성능평가

나이브 베이지안 분류 모델의 성능을 평가하기 위해 표 3.3와 같이 데이터 세트를 준비하였다. 영화리뷰와 일기예보 데이터는 훈련데이터 250개와 실험데이터 100개로 구성하였다(표3.3).

문서 종류	데이터셋	문서 수	문서 길이
영화 리뷰	훈련세트	250	350
	시험세트	100	350
일기예보	훈련세트	250	1,800
	시험세트	100	1,800

표 3.3 나이브 베이지안 분류기의 성능평가를 위한 데이터 세트

나이브 베이지안 분류 모델의 성능 평가하기 위해 머신러닝 분류모델의 대표적인 평가 지표인 accuracy(정확도), precision(정밀도), recall(재현율), 그리고 recall의 조화평균인 F_1 $score$를 이용하였다. 성능평가는 표 3.4에 정리되어 있다.

Data set	Accuracy	Precision	Recall	F1 Score
영화리뷰	0.84	0.74	0.77	0.75
일기예보	0.81	0.71	0.75	0.72

표 3.4 나이브 베이지안 분류 성능

3.4. 감성이 반영된 검색시스템

- **루씬(Lucene) 정보검색 라이브러리**

본 연구에서 검색시스템 개발을 위해 루씬(lucene)이라는 고성능 정보검색(IR, information retrieval)라이브러리를 사용하였다. 루씬을 선택한 이유는 루씬 그 자체로 독립적인 프로그램이 아닌 단순한 소프트웨어 라이브러리이기 때문에 이를 활용해 검색시스템을 개발해야 한다[15]. 루씬은 색인(index)기능과 검색(searching)기능을 프로그램에 간편하게 추가할 수 있도록 구성 되어있다. 색인은 텍스트를 빠르게 검색하기 위해 특별한 형태로 저장하는 것을 의미한다. 실제로 루씬은 역색인(inverted index)구조로 키워드에 문서들의 primary key(주소, 파일명)값을 매핑(mapping)하여 저장한다. 이것은 질의가 발생했을 때 문서안에 포함된 단어를 찾는 것이 아니라 역으로 단어가 포함된 문서를 찾는다. 따라서 빠른 검색과 디스크 공간을 효율적으로 사용할 수 있다는 장점이 있다. 또한 루씬은 full-text 검색 방식을 채택하여 사용자는 좀더 직관적이고 편리한 검색을 할 수 있을 뿐 아니라 대용량의 텍스트를 빠르고 효율적으로 검색할 수 있다.

- **검색시스템 구조**

다음은 본 연구에서 루씬 라이브러리를 사용하여 개발한 검색시스템의 개략적인 도식도이다(그림 3.6).

사용자 인터페이스를 통해 사용자가 질의와 검색결과를 주고받기 전에, 루씬 검색엔진은 미리 정보를 수집하여 색인을 해야 한다. 그림3.6은 사용자 인터페이스와 루씬 검색엔진과의 상호작용을 보여준다. 검색엔진은 문서의 텍스트를 빠르게 검색하기 위해 특별한 형태로 저장된 색인 중에서 입력된 쿼리에 관련된 정보들을 찾아서 보여준다. 색인 되어진 정보는 앞선 절에서 나이브 베이지안 분류를 이용한 감성분석을 통해 5차원 영화리뷰에 대한 감성백터가 추가되었던 영화정보 데이터베이스이다. 이를 통해 사용자가 질의를 전송하였을 때, 검색엔진에서 질의에 해당되는 검색순위 $\overline{\boldsymbol{Rank}}$를 찾는 것과 동시에 영화정보에 함께 연결된 5차원 감성벡터가 함께 검색되어진다. 따라서 영화정보와 검색순위가 영화리뷰 감성벡터와 함께 순위 재정렬 시스템으로 전달된다.

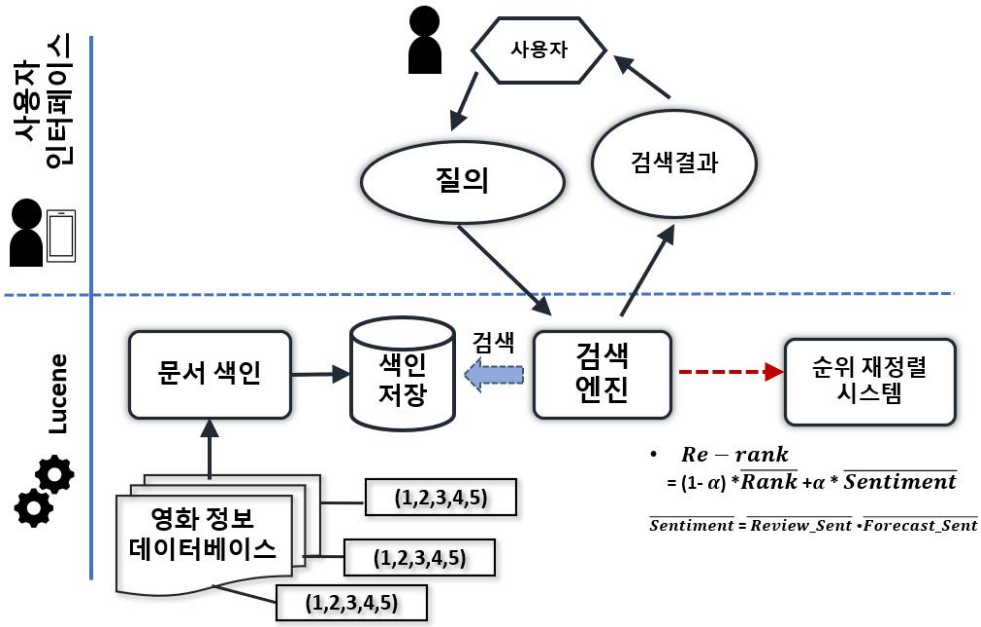


그림 3.6 루씬 라이브러리를 사용한 검색시스템 개략도

3.4.1. 순위 재정렬 시스템

다음 그림 3.7은 순위 재정렬 시스템에 대한 개략적인 모형이다. 실질적으로 순위 재정렬 시스템 단계에서 사용자의 감성이 반영된다.

루씬을 기반의 검색엔진 통해 사용자의 검색 질의 대한 검색요청으로 쿼리와 관련된 순으로 1위부터 20위까지 영화의 검색 순위$\overline{\boldsymbol{Rank}}$를 도출하였다. 이때 영화리뷰에 대한 감성벡터 또한 함께 검색되어 순위와 함께 순위 재정렬 시스템으로 전달되었다.

다음 절에서 설명할 영화검색 프로그램을 통해, 선택지의 날씨를 실험 참여자의 주변여건에 해당되는 감성으로 가정하여, 해당 날씨를 선택 시 감성을 순위 재정렬 시스템의 알고리즘에 반영하도록 설계하였다.

본 연구에서 다음 수식 3.6와 같이 새로운 *Movie* 함수를 고안하였다.

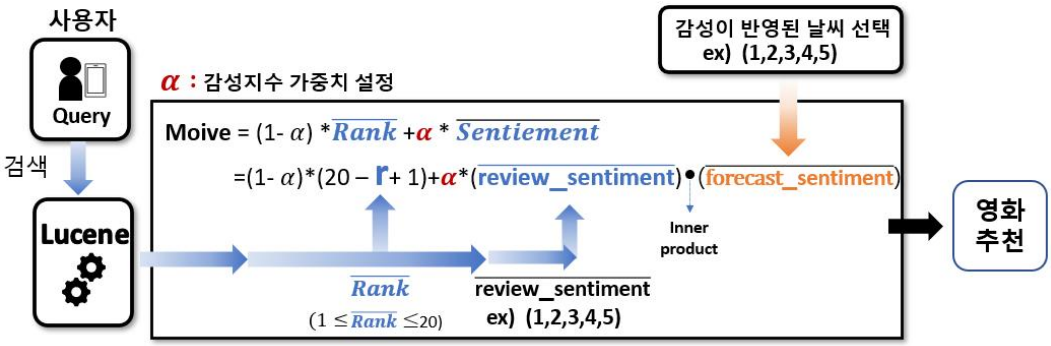


그림 3.7 순위 재정렬 시스템 개략도

- Movie 함수 정의

$$Movie = (1-\alpha) \times \overline{\boldsymbol{Rank}} + \boldsymbol{\alpha} \times (\overline{\boldsymbol{Sentiment}})$$

$$= (1-\alpha) \times (20-\mathrm{r}+1) + \boldsymbol{\alpha} \times (\overline{\boldsymbol{forecast_sent}}) \bullet ((\overline{\boldsymbol{review_sent}})) \quad 3.6$$

* 리뷰벡터($\overline{\boldsymbol{review_sent}}$)의 1, 2, 3 중에서 ($\overline{\boldsymbol{forecast_sent}}$)과 내적 계산을 하여 가장 큰 감성벡터의 ($\overline{\boldsymbol{Sentiment}}$) 값을 만족시키는 리뷰벡터만 반영하고 나머지는 2개의 벡터는 버린다.

α 값은 $\overline{\boldsymbol{Rank}}$ 벡터와 $\overline{\boldsymbol{Sentiment}}$벡터 사이에서 지수를 조절할 수 있다. 설정할 수 있는 감성 지수 $\boldsymbol{\alpha}$값의 지수 범위는 $0.1 < \boldsymbol{\alpha} \leq 1$이지만, 우리는 추천 영화에 대한 높은 만족도를 위해 최적의 $\boldsymbol{\alpha}$값을 찾는 실험을 하였다.
$\overline{\boldsymbol{Rank}}$ 벡터의 r값은 앞선 루씬 라이브러리 기반의 영화 검색 시스템에서 쿼리를 입력하여 추천된 검색순위 $\overline{\boldsymbol{Rank}}$값이다. 이때 $\overline{\boldsymbol{Rank}}$는 1위부터 20위까지의 범위이기 때문에 '0'의 값을 방지하기 위해서 1을 더했다.

감성을 나타내는 $\overline{\boldsymbol{Sentiment}}$ 감성벡터는 $\overline{\boldsymbol{forecast_sent}}$일기예보벡터와 $\overline{\boldsymbol{review_sent}}$벡터의 5차원 내적 값이다. 일기예보에 대한 감성벡터는 검색 사용자가 직업 선택하여 재정렬 시스템에 입력되었다. 영화리뷰에 대한 감성벡터는 앞서 설명한 20위 까지의 추천된 영화에서 순위와 함께 재정렬 시스템으로 전달되었다.

감성이 반영되는 효과를 보기 위해 일기예보와 영화리뷰의 내적 값이 가장 큰 값만 순위 재정렬 시스템에 사용되고 나머지 2개는 버리도록 설계하였다.

이렇게 *Movie* 함수를 통해 계산된 값을 바탕으로 영화검색에 감성을 반영하였다. 검색엔진에 의해 추천된 검색순위와 감성벡터와의 계산으로 검색순위를 재정렬하여 사용자의 감성을 반영한 영화를 추천한다.

3.5. 영화 검색 프로그램

실험을 위해 실험참가자에게 제공할 영화검색 프로그램을 설계하였다. 원활한 검색을 위하여 친숙한 영화검색 사이트의 디자인으로 외부 인터페이스를 제작하였다. 이러한 영화검색 프로그램의 구성과 순서는 본 연구의 실험목적에 부합하도록 설계하였고, 실험 참여자가 프로그램의 처음부터 끝까지 도달하여 실험을 마칠 수 있도록 설계하였다.

앞서 언급한 영화검색 인터페이스는 내부 프로그램과 연결되어 입출력을 통신해야 한다. 이러한 내부 프로그램은 앞서 설명한 검색엔진과 순위 재정렬 시스템이다. 그리고 감성분석을 위한 나이브 베이지안 분류기를 통해 생성된 감성벡터들은 내부 프로그램에서 사용할 데이터로서 저장해 두었다. 실험참가자가 전송한 입력 값을 기반으로 내부 프로그램이 알고리즘을 계산하도록 설계하였다. 계산된 결과 영화의 검색추천 결과를 외부의 유저 인터페이스로 전송되어 실험참가자에게 결과값을 보여주도록 설계하였다.

이러한 영화검색 프로그램은 자바 기반의 웹 프로그램을 개발을 지원해 주는 스프링부트(spring boot) 프레임워크를 사용함으로써 효과적으로 구축하였다[16]. 스프링부트 내부에는 톰캣(tomcat)서버가 포함되어 있고, 이를 통해 내부와 외부 프로그램의 통신을 가능하게 하였다. 그리고 실험 참여자가 입력하는 모든 값들은 검색시스템 평가를 위해 데이터베이스에 저장하도록 설계하였다.

이어지는 다음 절에는 영화검색 프로그램의 외부에 해당되는 영화검색 인터페이스에 대해 설명하겠다.

3.5.1. 영화검색 인터페이스

실험을 위해 실험 참여자에게 제공할 감성기반 영화검색 프로그램의 외부 인터페이스를 구축하였다(그림 3.8 ~ 3.13). 외부 프로그램을 통해 영화에 관련된 검색쿼리와 실험 참여자의 주변 여건에 해당되는 날씨를 내부 프로그램에 입력할 수 있도록 설계하였다. 따라서 내부 프로그램 단독으로 기능을 할 수 없다.

영화 검색을 유도하기 위해, 영화와 관련된 질문이 포함된 6개의 Task가 준비되어있다. 이러한 각각의 Task마다 서로 다른 감성 지수를 반영하는 검색엔진들로 구성되어 있다. 같은 Task에는 동일한 검색어가 서로 다른 감성 지수를 반영하는 검색엔진에 쿼리로 입력된다. 이를 통해, 서로 다른 검색엔진에서 추천되는 검색 결과는 1순위부터 20순위까지 사용자 인터페이스에 나타난다. 실험 참가자는 검색결과에 따른 만족도를 입력하여 영화검색 프로그램의 성능을 평가하도록 하였다.

다음 그림 3.8은 영화검색 프로그램의 인터페이스 시작 화면이다.

실험 참여자에게 실험에 대한 목적, 순서 등 정보를 제공해주고 실험시작 버튼을 누를 수 있도록 설계하였다. 이후 실험참여자가 원할 시 실험 중도에 종료할 수 있는 버튼을 만들었다.

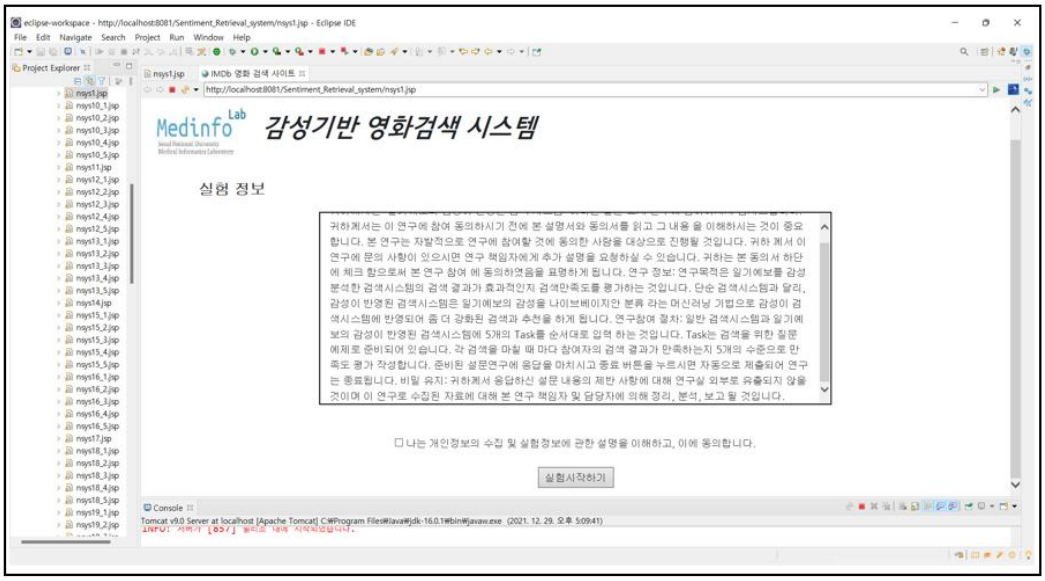


그림 3.8 감성기반 영화검색시스템의 시작화면

다음 그림 3.9는 영화검색 시스템의 메인 검색창 인터페이스를 나타낸다. Task는 1번부터 6번까지 구성 되어있고 각 Task 마다 영화검색을 유도하기 위한 영화에 관련된 질문을 준비하였다. 검색 쿼리는 모두 데이터베이스에 저장되며, 이러한 쿼리는 다음 차례의 비교 검색엔진으로 그대로 전송되어 동일한 검색어로 비교 검색엔진을 작동시킬 수 있도록 설계하였다.

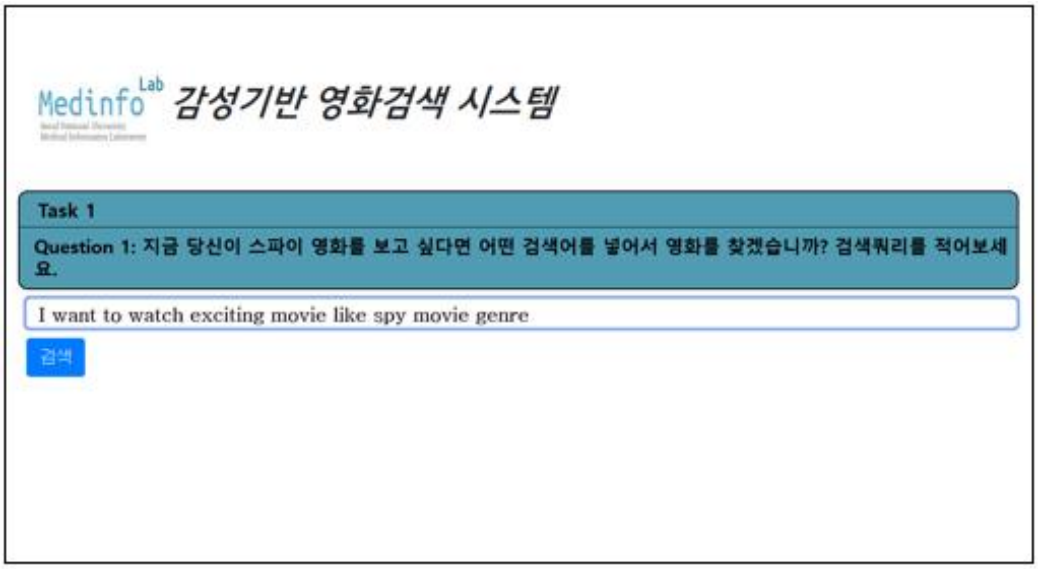


그림 3.9 영화검색 시스템의 메인 검색창

다음 그림은 메인 검색창에서 쿼리를 입력 후에 검색결과를 확인할 수 있는 검색결과화면이다. 1위부터 20위까지 검색에 대한 영화추천 리스트를 제공하도록 인터페이스를 제작하였다. 이 화면에서 영화제목, 상연연도, 러닝타임, 장르, 감독, 주연배우 평점에 대한 정보를 확인할 수 있도록 설계하였다(그림 3.10).

실험 참가자가 추천된 영화가 어떤 영화인지 모르는 경우를 추천된 검색결과에 대한 평가를 제대로 할 수 없다. 이를 대비하기 위해, 영화 포스터 이미지와 영화 1편당 서로 다른 영화리뷰 탭(tab)을 3개씩 제공하였다. 이러한 영화리뷰에는 줄거리도 포함 되어있어 영화에 대한 대략적인 내용을 파악할 수 있도록 하였다(그림 3.11)

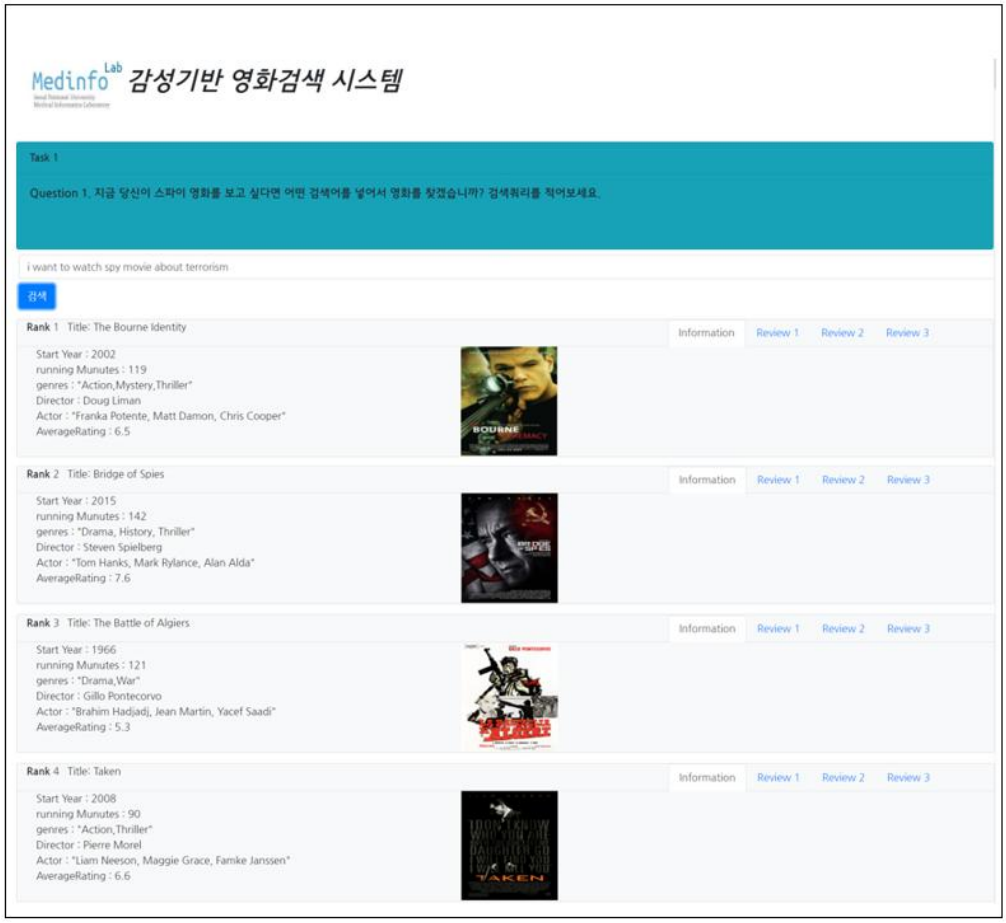


그림 3.10 영화추천 검색결과

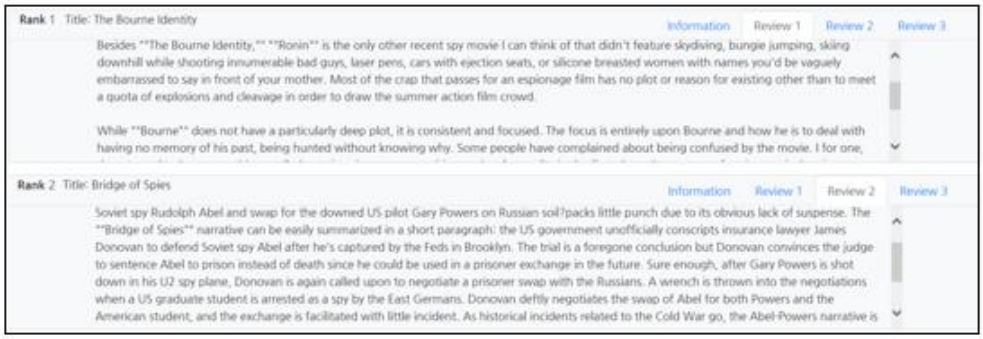


그림 3.11 영화 1편당 3개의 리뷰 탭

실험 참가자가 1순위부터 20순위까지의 추천 결과를 바탕으로 검색엔진의 성능을 평가하기 위해 검색결과에 대한 만족도를 평가할 수 있도록 그림 3.12과 같이 20순위의 검색결과 아래에 위치하도록 설계하였다.

검색 만족도 평가 기준은 5가지로 매우 만족(5점), 만족(4점), 보통(3점), 불만족(2점), 매우 불만족(1점) 순으로 라디오 버튼 형식으로 구현하여 실험참가자가 직접 검색 추천결과에 대한 만족도를 평가하도록 설계하였다. 이러한 실험참가자 만족도 평가의 입력 값은 모두 데이터 베이스에 저장되어 검색엔진의 성능평가에 대한 데이터로 사용하였다.

그림 3.12 영화 추천에 대한 실험참가자 만족도 평가

다음은 실험 참가자의 주변여건에 해당되는 날씨를 선택할 수 있는 프로그램의 화면이다. 선택지의 날씨를 실험 참여자의 주변여건에 해당되는 감성으로 가정하여, 해당 날씨를 선택 시 감성을 반영하도록 설계하였다. 그림 3.13에서 실험참여자가의 주변에 느껴지는 감성을 선택할 수 있도록 5개의 선택지를 라디오 버튼 형식으로 제공하였다. 라디오 버튼을 누르면 그림 3.14과 같이 선택한 감성의 기상예보에 대한 자료가 나타나도록 설계하였다. 이러한 자료는 동영상의 일기 예보방송과 해당 아나운서의 대본으로 구성되었고, 아나운서 대본은 나이브 베이지안 분류 모델을 활용한 감성분석을 통해 이어질 검색엔진에 감성이 반영된다.

따라서 이 과정을 통해, 실험참가자가 본인의 주변여건이나 현재 느끼는 감성을 선택지로 직접 선택함으로써, 실시간으로 검색프로그램에 감성이 반영하였다.

그림 3.13 감성을 반영한 일기에보 선택지

그림 3.14 감성을 반영하기위한 일기예보 시청각 자료

3.6. 실험 구성

본 연구의 영화검색 시스템은 6가지 task가 순서대로 진행되도록 설계하였다. 이러한 task는 실험 참가자의 영화에 관련된 검색쿼리를 유도하도록 영화에 관한 질문이 포함되어 있도록 구성하였고, 6개 task의 내용은 다음 표 3.5와 같다. 이를 통해, 실험 참여자들이 영화에 관련되지 않거나 너무 엉뚱한 쿼리를 입력하여 본 연구의 목적에 벗어나는 것을 방지하기 위함이다.

그리고 세상 모든 영화의 데이터를 수집하면 좋았겠지만, 현실적으로 제한이 있었다. 따라서 Task의 질문내용은 데이터베이스로 준비된 영화 범위 내에서 만들었다. Task 1과 Task 5과 달리, task 6은 자유로운 검색 형식으로 질문을 구성해 대조군을 설정하였다. 영화에 관련되지 않는 질문이나 데이터 베이스 범위에 넘어가는 검색을 한다면 검색성능이 떨어질 것으로 예상된다.

실험목적에 따른 실험 조건은 다음절에서 이어진다.

Task	Question
Task 1	지금 당신이 스파이 영화를 보고싶다면 어떤 검색어를 넣어서 영화를 찾겠습니까? 검색쿼리를 적어보세요.
Task 2	지금 당신이 전쟁과 관련된 영화를 보고싶다면 어떤 검색어를 넣어서 영화를 찾겠습니까? 검색쿼리를 적어보세요.
Task 3	귀신이 출몰하는 영화 중에 어떤 특징이 있는 영화를 추천할지 검색쿼리를 적어보세요.
Task 4	코미디 영화 중에 어떤 특징이 있는 영화를 추천할지 문장을 작성해 주세요.
Task 5	추리영화 중에 추천할 만한 영화의 특성을 문장으로 작성해 주세요.
Task 6	지금 당신이 시청 하고 싶은 영화를 자유롭게 문장 형식으로 적어보세요.

표 3.5 Task 구성과 내용

3.6.1. 실험 조건

본 연구에서는 감성지수를 변경하며 감성지수의 영향력을 확인하기 위한 실험 목적을 가지고 실험을 진행하였다. 그림 3.15은 이러한 목적을 위한 실험 구성을 나타낸다. 실험 참가자는 성인 남녀 50명이 참가였고, 서울의대 IRB 심의 면제 신청을 하여 승인을 받았다.

[서울의대 IRB승인번호: IRB No.2112-046-1281]

실험은 task 1부터 task 6까지 쿼리를 유도하는 영화관련 검색질문에 대한 검색을 진행하고 프로그램은 종료된다. 각 task에는 검색엔진이 3개씩 구성되어 있는데, 첫번째 검색엔진은 감성지수가 $\alpha = 0$ 으로 반영되지 않는 일반적인 검색엔진이다. 다음 두번째, 세번째 검색엔진은 감성이 반영된 검색엔진인데, 각각 감성지수가 $\alpha = 0.3$ 과 $\alpha = 0.7$ 로 반영된 검색엔진으로 구성하였다. 감성검색엔진은 실험 참가자의 주위여건에 해당하는 감성을 반영하는 날씨를 선택한 후에, 이전에 검색에 사용되었던 검색쿼리가 동일하게 적용되어 검색되어진다. 이를 통해 동일한 검색쿼리를 기반으로 세가지 검색엔진을 비교하여 검색성능을 평가할 수 있었다.

- 실험: 감성지수의 영향력에 대한 비교 실험

실험참여자: 50명

검색엔진 구성 ($0 \leq \alpha \leq 1$)

1. 일반 검색엔진 A Type 감성지수($\alpha = 0$)
2. 감성 검색엔진 B type 감성지수($\alpha = 0.3$)
3. 감성 검색엔진 C type 감성지수($\alpha = 0.7$)

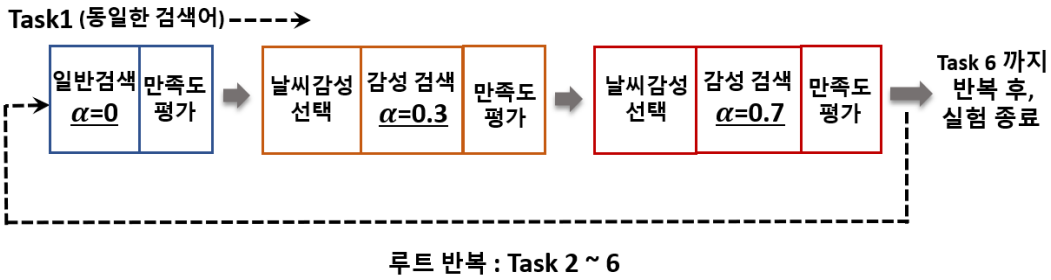


그림 3.15 전체 실험의 구성에 대한 도식도

4. 결과

4.1. 감성지수의 영향력에 대한 실험 결과

실험참가자 50명을 대상으로 감성이 반영되지 않는 일반검색($\alpha = 0$)과 감성지수가 각각 $\alpha = 0.3$과 $\alpha = 0.7$로 반영된 감성검색엔진의 성능을 비교하는 실험을 진행하였다. 일반 검색엔진은 감성이 반영되지 않아 검색순위가 재정렬 되지 않았다. 그에 반해, 감성이 반영된 두 종류의 검색엔진은 검색 재정렬 시스템에 의해 검색 순위가 재정렬 되었다.

앞서 설명한 3종류의 검색엔진을 기반으로 task 1부터 task 6까지 실험한 실험 참가자의 만족도를 표 4.1와 같이 조사하였다.

일반검색엔진 ($\alpha = 0$)의 실험 참가자 만족도 평가는 평균 3.53점을 기록하였다. 감성지수가 $\alpha = 0.3$과 $\alpha = 0.7$ 반영된 검색엔진의 만족도 평가는 각각 평균 3.63점과 평균 3.89점을 기록하였다. 그리고 이것을 3.15와 같이 감성 지수에 따른 검색 성능 만족도를 그래프로 나타내었다.

Task	α=0	α=0.3	α=0.7
Task 1	3.42	3.50	3.94
Task 2	3.60	3.68	3.90
Task 3	3.64	3.76	3.92
Task 4	3.56	3.58	3.88
Task 5	3.68	3.78	4.06
Task 6	3.30	3.46	3.64
평균	3.53	3.63	3.89

4.1 감성지수에 따른 검색만족도 표

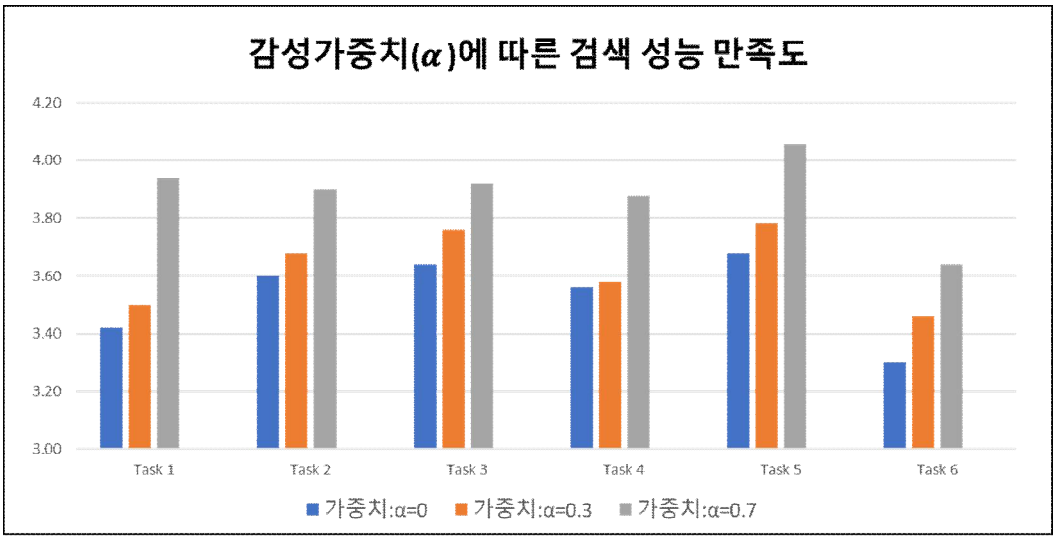


표 4.2 감성 지수에 따른 검색 성능 만족도 그래프

4.2. T-test 유의성 검정

일반 검색엔진과 감성이 반영된 검색엔진과에 대한 성능 평균점수가 통계적으로 유의한 차이가 있는 검증을 하기 위해 T-test 검정을 하였다.

첫번째로는 감성지수가 없는 일반검색엔진과 0.3으로 반영하는 감성검색엔진에 대한 T-test 검정을 하였다. T-test검정 결과 P value는 0.2744로 유의수준의 0.05보다 크므로 일반검색엔진과 감성지수가 0.3으로 반영된 검색엔진의 성능 차이는 통계적으로 유의하지 않았다.

두번째로 일반검색엔진과 감성지수가 0.7로 반영하는 검색엔진에 대한 T-test 검정을 하였다. T-test검정 결과 P value는 0.0014로 유의수준의 0.05보다 작기 때문에 일반검색엔진과 감성지수가 0.7으로 반영된 검색엔진의 성능 차이는 통계적으로 유의함을 보였다.

이를 통해 감성지수가 0.7로 반영하는 감성검색엔진이 통계적으로 유의성이 있는 검색 만족도가 나타냄을 알 수 있었다.

5. 고찰 및 결론

감성지수의 영향력을 확인하기 위해 감성 지수를 변경해가며 일반검색엔진과 감성지수가 $\alpha = 0.3$ 과 $\alpha = 0.7$인 감성검색엔진의 성능을 비교할 수 있도록 실험을 설계하였다. 본 연구의 실험 결과는 $\alpha = 0$ 일때, 평균 3.53점, $\alpha = 0.3$ 일 때, 평균 3.63점, $\alpha = 0.7$일 때, 평균 3.89점의 검색 만족도를 기록하였다. 이를 통해 서로 다른 감성지수를 반영한 검색엔진 중에 감성이 $\alpha = 0.7$로 반영된 검색엔진이 사용자의 검색 만족도가 가장 높음을 확인하였다.

또한 실험결과에 대한 통계적인 유의성을 확인하기 위해서 T-test검정을 하였다. 일반검색엔진과 서로 다른 감성 지수를 반영한 감성검색엔진과의 T-test 검정을 통해, 감성지수가 0.7로 반영하는 감성검색엔진이 통계적으로 유의한 검색 만족도를 나타냄을 확인하였다.

소비자의 패턴의 변화에 의해 검색과 추천의 플랫폼 수요는 계속 늘어날 전망이다. 그러나 기존의 검색과 추천시스템에서 벗어나 새로운 패러다임의 시스템이 필요하다. 이에 본 연구에서는 정보 검색 분야와 기계학습을 이용한 감성분석 분야를 접목시켜 새로운 검색시스템을 만들고자 하였다. 본 연구에서 개발한 감성을 반영한 검색엔진은 기존에 접근하지 못한 감성의 영역을 기계적인 모델을 사용하여 접근할 수 있을 것이다. 이를 통해 음악, 영화, 쇼핑과 같은 검색플랫폼 등에서 상업적으로 활용될 시 유용할 것으로 기대된다.

또한 본 연구를 통해 검색시스템을 이용하는 사람의 감정상태를 기계적으로 파악할 수 있어 의료분야에서 활용할 가능성이 있을 것이라 예상된다. 예를 들어 평소에 우울하거나 암울한 영화를 많이 보는 사람의 감성을 미리 분석한다면 자살과 같은 위험들을 미연에 방지할 수 있을 것으로 기대한다.

참고 문헌

[1] Fauci, Anthony S., H. Clifford Lane, and Robert R. Redfield. "Covid-19—navigating the uncharted." (2020): 1268-1269.

[2] Narae Park. “How to solve the problem of Netflix Syndrome”. *UX Collective*, 3 Aug 2019. https://uxdesign.cc/how-to-solve-the-problem-of-netflix-syndrome-6111cb9e7c94-6111cb9e7c94. Accessed 29 Nov 2021.

[3] Zhou, Guorui, et al. "Deep interest network for click-through rate prediction." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

[4] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." Computer networks and ISDN systems 30.1-7 (1998): 107-117.

[5] Renaud-Deputter, Simon, Tengke Xiong, and Shengrui Wang. "Combining collaborative filtering and clustering for implicit recommender system." 2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA). IEEE, 2013.

[6] Basilico, Justin, and Thomas Hofmann. "Unifying collaborative and content-based filtering." Proceedings of the twenty-first international conference on Machine learning. 2004.

[7] Feldman, Ronen. "Techniques and applications for sentiment analysis." Communications of the ACM 56.4 (2013): 82-89.

[8] Todd Atkins, Baton Rouge therapist, https://batonrougecounseling.net.

[9] Saritas, Mucahid Mustafa, and Ali Yasar. "Performance analysis of ANN and Naive Bayes classification algorithm for data classification." International Journal of Intelligent Systems and Applications in Engineering 7.2 (2019): 88-91.

[10] Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001..

[11] Baylis, Patrick, et al. "Weather impacts expressed sentiment." *PloS one* 13.4 (2018): e0195750.

[12] 기상청 날씨누리, https://www.weather.go.kr/weather/warning/wtouchqNew.jsp

[13] IMDb Datasets, https://www.imdb.com/interfaces/

[14] Daniel Jurafsky, James Martin, 『Speech and Language Processing』, Pearson Higher Education(2020), PP.55 -75

[15] McCandless, Michael, et al. Lucene in action. Vol. 2. Greenwich: Manning, 2010.

[16] Walls, Craig. Spring Boot in action. Simon and Schuster, 2015.

[15] McCandless, Michael, et al. Lucene in action. Vol. 2. Greenwich: Manning, 2010.

Abstract

A Sentiment based Retrieval System using Naïve Bayesian Classification

Hyunsic Kim
Interdisciplinary program in Bioengineering
The Graduate School
Seoul National University

This paper is a study on a model that recommends search results that are more suitable for individual sentiments by re-adjusting search rankings based on an algorithm that reflects sentiments. Since December 2019, the viewing time on OTT (Over-the-top media service) platforms has increased due to the pandemic of COVID-19, a global infectious disease, but a phenomenon called Netflix Syndrome has occurred. Therefore, this study developed a movie search system that reflects the user's sentiments through sentiment analysis to increase the search accuracy and satisfaction of users using the search system to solve the above problems.

Sentiment analysis is a natural language processing technology, and it is newly defined in this study as five sentiments: angry, depressed, energetic, nervous, and tender. By using this definition as a machine learning technique of Naive Bayesian classification, sentiment analysis was performed on weather forecast and movie review data.

The Naive Bayesian classification technique is a statistical technique that infers posterior probabilities based on additional information newly entered into the prior probabilities of an inference target under the assumption that all word features are independent. As a result of evaluating the model through machine learning and testing using this principle, it was possible to confirm the accuracy performance of 0.84 and 0.81 for movie review and weather forecast, respectively.

The sentiment prediction score generated through Naive Bayesian classification was transferred to the sentiment vector generation model, and the sentiment prediction score was ranked and converted into an sentiment vector. The converted weather forecast and sentiment vectors of movie reviews are calculated together with the search engine devised in this study to recommend search results reflecting sentiments. At this time, the optimal performance is confirmed by introducing the sentimental index (**α**), which adjusts the degree to which sentiments are reflected in the algorithm experiment was conducted.

In addition, in this study, a search engine that reflects sentiments was created by using the Lucene library specialized in information retrieval. In addition, a movie search program including a movie search interface was developed so that the subjects could directly participate in the search.

In this study, by changing the sentimental indexes, a comparative experiment was conducted on the influence of the search results of the search system according to the sentimental indexes. An experiment was conducted to compare the performance through the search satisfaction of the general search engine α=0 and the search engine in which sensibility

was reflected as α=0.3 and α=0.7 for 50 subjects. The experimental results of this study recorded an average search satisfaction score of 3.53 points when α=0.7, an average of 3.63 points when α=0.3, and 3.89 points when α=0.7. Through the T-test with the general search engine and the sentimental search engine that reflects different sentimental indexes, it was confirmed that the sentimental search engine that reflects the sentimental index as α=0.7 showed statistically significant search satisfaction.

The search engine that reflects the sentiment developed in this study will be able to access the area of sentiment that has not been approached before using a mechanical model. Through this, it is expected that it will be useful when used commercially in search platforms such as music, movies, and shopping.

In addition, it is expected that this study will have the potential to be used in the medical field because it is possible to mechanically grasp the sentimental state of the person using the search system.

Keywords : Natural Language Processing, Sentiment Analysis, Machine Learning, Naive Bayesian Classification, Information Retrieval

Student Number : 2019-29690