감성과학, Vol. 12, No. 4, pp.433-442, December 2009

텍스트 문서 기반의 감성 인식 시스템*

An Emotion Scanning System on Text Documents

김명규**† · 김정호** · 차명훈** · 채수환***

MyungKyu Kim**† · JungHo Kim** · MyungHoon Cha** · Soo-Hoan Chae***

한국항공대학교 컴퓨터공학과**

Department of Computer Engineering, Korea Aerospace University**

한국항공대학교 항공전자정보통신공학부***

School of Electronics, Telecommunication and Computer Engineering, Korea Aerospace University***

Abstract

People are tending to buy products through the Internet rather than purchasing them from the store. Some of the consumers give their feedback on line such as reviews, replies, comments, and blogs after they purchased the products. People are also likely to get some information through the Internet. Therefore, companies and public institutes have been facing this situation where they need to collect and analyze reviews or public opinions for them because many consumers are interested in other's opinions when they are about to make a purchase. However, most of the people's reviews on web site are too numerous, short and redundant. Under these circumstances, the emotion scanning system of text documents on the web is rising to the surface. Extracting writer's opinions or subjective ideas from text exists labeled words like GI(General Inquirer) and LKB(Lexical Knowledge base of near synonym difference) in English, however Korean language is not provided yet. In this paper, we labeled positive, negative, and neutral attribute at 4 POS(part of speech) which are noun, adjective, verb, and adverb in Korean dictionary. We extract construction patterns of emotional words and relationships among words in sentences from a large training set, and learned them. Based on this knowledge, comments and reviews regarding products are classified into two classes polarities with positive and negative using SO-PMI, which found the optimal condition from a combination of 4 POS. Lastly, in the design of the system, a flexible user interface is designed to add or edit the emotional words, the construction patterns related to emotions, and relationships among the words.

Keywords : Measuring user's sensibility, emotion words, Text classification

요 약

요즈음 인터넷을 통해 물건을 구매하는 경향이 증가하고 있다. 또한 물건을 구매한 소비자는 리뷰, 댓글, 비평 또는 블로그 등의 형식으로 온라인에 그들의 사용 후기를 작성한다. 또한 작성된 사용 후기부터 많은 구매자들은 물건을 구매하기 전에 자신이 구입하고자 하는 물건에 대한 정보를 얻는다. 따라서 회사나 공공기관은 대중이 다른 사람의 의견에 관심을 기울인다는 점 때문에 대중의 의견을 수집하고 분석할 필요성에 직면하였다. 그러나 온라인상에 댓글이 너무 많고, 중복적이면서 짧은 경향이 있다. 이러한 환경 속에서 텍스

* 이 논문은 2009년도 고양시 산학관 공동 기술개발 사업으로 고양시 및 (주)RSN 지원을 받아 연구되었음.

† 교신저자 : 김명규 (한국항공대학교 컴퓨터공학과)

E-mail : kimmk@kau.ac.kr

TEL : 02-300-0146

트 문서의 감성을 인식하는 시스템의 필요성이 대두되었다. 텍스트로부터 작성자의 의견이나 주관적인 생각을 추출할 수 있게 영어에서는 단어에 속성이 주어진 GI와 LKB가 있으나 한글은 아직 속성이 주어진 사전이 존재하지 않는다. 이 논문에서는 한글 품사 중 4개의 품사(명사, 동사, 형용사, 부사)에 속성을 주었다. 그리고 학습 군을 만들어서 감성 단어의 패턴을 구성하고, 문장에서 단어 사이의 공기관계를 구성하여 학습 시켰다. 이 학습을 바탕으로, SO-PMI을 이용하여 문서를 긍정과 부정 2가지 극성을 분류하고, 4개의 품사(명사, 동사, 형용사, 부사)를 각각 조합하여 최상의 조건을 구하였다. 마지막으로 사용자 인터페이스를 통해 새로운 감성 표현, 구성형식, 단어 연관성을 반자동적으로 삽입하고 교정할 수 있는 시스템을 설계하였다.

주제어 : 사용자 감성측정, 감성어, 문서 분류

1. 서론

인터넷이 대중화됨에 따라 언론 기관 또는 개인이 온라인(on-line) 상에서 정치, 경제 등의 사안, 기업 및 제품의 이미지, 공인에 대한 자신의 의견을 게시판, 블로그, 포럼, 댓글 형식으로 나타낸다. 기업 및 공공기관의 입장에서는 언론 매체 및 대중이 기업, 제품 또는 공공 정책을 어떻게 생각하는지에 대한 자료를 수집하고 분석하는 서비스가 필요하게 되었다. 이러한 사용자의 반응 상태를 찾아내는 연구를 감정분석(sentiment analysis), 감정분류(sentiment classification), 의견추출(opinion extraction, or opinion mining)이라고 한다.

이 분야에서는 사용자의 생각을 인식하기 위해 문서 내부나 댓글 등에서 감성적 표현(emotional expression) 단어를 찾아내는 단어 중심적 시스템(keyword based system)에 관심을 둔다.

이러한 시스템 연구로는 영어권에서 오래전부터 말뭉치의 관계를 정립한 WordNet이나 WordNet에서 긍정과 부정의 속성을 주어서 극성(polarity) 용어를 나타낸 SentiWordNet[1)2)], GI(General Inquirer), LKB(Lexical Knowledge base of near synonym difference)와 같은 프로젝트 등을 들 수 있다.

국내 경우는 '21세기 세종 기획'과 같이 국립국어원에서 대표적으로 한글 정보화 작업을 주도하고 있으나, 별도의 감성 표현 말뭉치 연구는 아직 미흡한 편이다. 또한 영어 같은 경우에는 단어를 한 벡터로 사용하여 감성 분류가 가능하지만 한글의 경우에는 한글의 어순이 자유롭고, 언어의 중의적 표현 문제와, 단어와 단어 사이 관계로 느낌을 표현하고, 한 어절 안에 어미변화로 극성을 다르게 대표적으로기 때문에 몇 개의 논문[3)4)5)]에서 감성 어휘를 나타내고 있으나, 문서의 극성을 처리 시에는 아직 부족하다.

본 논문에서는 이런 문제를 유연하게 대처할 수 있는 (그림 1)과 같은 감성인식시스템을 제안한다. 제안한 감성 인식시스템은 문서로부터 검색 대상이 되는 개체명(named entity)을 인식하여 의견 표현 대상이 되는 객체(이하 개체라고 한다)에 대한 감성을 학습에 의하여 분류한다. 이 시스템의 주요 목적은 문서 및 문서에 대한 리뷰(review)로부터 개체 대상, 감성 언어, 극어(polarity), 강조어, 동사를 문장에서 추출하여 크게는 긍정, 부정 요소로 분류하는 것으로 세부적으로는 긍정 부정 요소를 10개로 소분류 하여 문서상에 나타난 의견을 평가하는 데 있다(<표 3> 참조).

시스템 구성 요소로 태깅 엔진을 포함한 문서 추출기, 문맥의 공기 정보(co-occurrence information)를 지닌 엔진, 문맥에서 감성적 규칙을 추출하는 엔진, 사전 리스트와 극성과 감정 속성을 표시한 감성 기본 사전을 바탕으로, 학습을 통해서 추출한 문장의 패턴 정보를 가진 엔진, 그리고 감성 극성을 판단하는 엔진으로 구성하였다.

1) Esuli A., Sebastiani F. (2006). SentiWordNet: A Publicily Avaliable Lexical Resource for Opinion Mining, In Proceeding of LREC-06, 5th Conference of Language Resource and Evaluation, Genova, IT, pp. 417-422.

2) Hatzivassiloglou V., Mackeown K. (1997). Predicting the Semantic Orientation of Adjectives, In Proc. of 35th ACL/8th EACL.

3) 박인조, 민경환 (2005). 한국어 감정단어의 목록 작성과 차원 탐색, 한국심리학회지, 사회 및 성격, 제 19권(1호), pp. 109-129.

4) 한덕웅, 강혜자 (2000). 한국어 정서 용어들의 적설정과 경험 빈도", 한국심리학회지, 일반, 제 19권(2호), pp.3-99.

5) 권용주, 신동훈, 한혜용, 반윤복 (2004). 과학적 관찰과 규칙성 발견 활동에서 나타나는 감성 단어 유형과 과학 지식 생성력과의 관계, 한국과학교육학회지, 제 24권(6호), pp.1106-1117.

여기에서는 대상이 되는 개체의 개체명을 사람과 기업 및 기업의 제품으로 한정하였다.

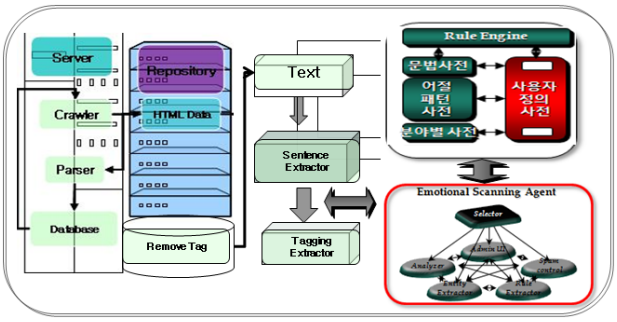

그림 1. 감성 인식 시스템 개요

2. 관련연구

감성이란 정서적 반응, 표현적 행동의 조합이 빚어내는 기분이나 감정 또는 강한 느낌에 해당하는 상태를 일컫는 말이다.[6] 감성 분류는 문서를 분류할 때, 문서를 객관적인(objective) 글과 주관적인(subjective) 글로 나누고 문서나 댓글내의 텍스트(text)로부터 긍정 또는 부정적인 주관성(subjectivity)을 인식하는 연구이다. 감성 상태의 측정 가능성으로 보면 어휘를 사용하여 정서를 측정하는 방법으로 감성의 차원이나 범주를 추론하는 작업이 감성 분석을 위한 선행 작업이다.

2.1. 감성 모델

2.1.1. 국내 모델 사례

국내의 감성 어휘는 형용사적 의미를 바탕으로 진행되었다. 형용사적 의미의 분류 체계는 연구하는 사람마다 기준이 다르다. 1980년 이전에는 형용사에서 감성이란 분류를 존재하지 않았으나, 1980년 중반 이후부터 감정이나 심리에 대한 사례가 보인다. 한국어 감성 단어에서 특히 감정에 관련된 단어를 연구한 박인조(2005)는 434개의 한국어 감정단어 목록을 만들었다. 또한 한국어 정서용어들의 적절성과 경험빈도를 연구한 한덕웅(2000)의 감성단어 연구도 있다. 과학수업에 감성을 적용한 사례 연구의 내용에서 제시된 학습자의 감성 단어는 한혜영, 권용주 등에 의해 제안되었고, 김은영은 현대 국어 감성 동사에 대한 범위와 의미 특성에 대한 연구로 기쁨(喜), 노여움(怒), 슬픔(哀), 두려움(懼), 좋아함(愛), 싫어함(惡), 바람(慾)과 같이 칠정(七情)을 중심으로 감성 언어를 분류한 연구가 있다.[7][8][9]

2.1.2. 국외 모델 사례

국외의 인지 평가 모델의 대표적 사례는 Ortony, Collins, Clore 세 사람이 제안한 OCC 모델이다.[10] OCC 모델은 심리와 관련을 가지며, OCC 감정 모델은 사람의 모든 감정을 나타내는 대신 감정 유형(emotion type) 이라는 감정 군집 추론에 대한 연구를 전체 내용으로 한다.

예를 들면, 고뇌라는 감정 유형은 불쾌한 사건이 원인이 되어 발생되는 모든 감정을 설명한다. 즉 고뇌는 불쾌하게 되는 사건의 원인과 정도의 차이에 따라 슬픔, 복받치는 감정, 사랑의 아픔 등과 같은 다양한 감정을 포함한다. OCC 모델에서는 감정 유형을 취급하며 유사한 감정의 군집화를 실제적으로 다룬다.

OCC 감정 모델은 특정 감정에 대한 세 가지 유형의 주관적(subjective) 평가를 제안하였다. 주관적 평가의 첫째는, 에이전트의 목표에 관련된 사건에 대한 만족에 관한 평가, 둘째는 행위에 관한 표준의 집합으로부터 에이전트 혹은 다른 에이전트의 행위의 승인에 관한 평가, 셋째는 그 에이전트의 태도에 관련된 대상을 좋아하는지 여부에 관한 평가이다. 또한, OCC 모델은 서로 다른 감정의 조합에 의해 생성되는 감정들을 제안하였다.

OCC 모델에서 정의한 대표적인 감정들은 즐거움, 고뇌, 희망, 두려움, 자부심, 수치심, 감탄, 치욕, 분노, 감사, 만족, 그리고 후회 등 22개의 도메인으로 감정의 수가 한정되어 있어서 구현이 용이하다.

6) Schlosberg, H. (1952). “The descriptions of facial expressions in terms of two dimensions.” Journal of Experimental Psychology,. 44, pp. 229-237.

7) 한혜영 (2005). 귀납적 과학지식의 생성과정에서 나타나는 감성분석을 위한 측정도구 개발, 한국교원대학교 대학원 석사학위논문

8) 권용주, 신동훈, 한혜용, 반윤복 (2004). 과학적 관찰과 규칙성 발견 활동에서 나타나는 감성 단어 유형과 과학 지식 생성력과의 관계, 한국과학교육학회지, 제 24권(6호), pp.1106-1117.

9) 김은영 (2005). 현대 국어 감정동사의 범위와 의미 특성에 대한 연구. 한국어 의미학, 16, 99-124.

10) Ortony, A., Clore, G.L., and Collins, A. (1998). The Cognitive Structure of Emotions., Cambridge University Press.

2.2. 의견 추출(Opinion Mining)

OM(Opinion Mining)은 IR(Information Retrieval)에서 개체명(named entity)에 대한 사용자의 의견을 긍정, 부정으로 추출하는 것이다. OM은 IR의 문서 요약(document summarize)뿐만 아니라 사용자의 생각을 반영하기 위한 것으로, OCC 모델의 세 가지 유형의 주관적 평가와 비슷하게 관련 연구나 작업을 세 가지로 나타낼 수 있다.

첫째는 주어진 텍스트가 실질적인 현상을 나타내는 객관적(objective)이거나 주관적(subjective)인 문제에 대한 긍정이나 부정의 극어(polarity)를 나타내는 것이다.[11] 두 번째는 주어진 주관적인 텍스트를 표현함에 있어서 주관적인 문제에 대해 긍정이나 부정으로 나타내는 것이다.[12] 세 번째는 긍정과 부정으로 나타낼 때 강한 극성을 지니고 있는지 약한 극성을 나타내는지 극성의 강도를 결정하는 것이다.

OM은 극성(긍정, 부정)을 분류하는 방법으로 PMI (Pointwise Mutual Information) method,[13] machine learning method, NLP(Natural Languages Processing) combined method[14]과 통계학적인 Rocchio algorithm,[15] SVM(Support Vector Machine)[16] 등이 있다.

2.3. SO-PMI

특정 두 단어의 PMI(Pointwise Mutual Information) 란, 단어와 단어 간의 연관 또는 관련 정도를 통계적인 방법을 통해 얻은 수치이다. $word_i$과 $word_j$의 PMI 값은 다음 (식 1)으로부터 구할 수 있다.

$$PMI(word_i, word_j) = \log_2[\frac{p(word_i\ AND\ word_j)}{p(word_i)p(word_j)}] \quad \text{(식 1)}$$

$p(word_i AND\ word_j)$: $word_1$, $word_2$가 한 문장에서 동시에 나올 확률

$p(word_i)$: $word_i$가 한 문장에서 나올 확률

이를 감성 인식 혹은 분류에 적용할 경우 (식 2), (식 3)과 같이 PMI를 적용할 수 있다. 여기서 $pword_j$와 $nword_j$는 각각 사전에 정의한 긍정, 부정 극성을 가지는 단어이다.

$$PMI(word_i, pword_j) = \log_2[\frac{p(word_i\ AND\ pword_j)}{p(word_i)\,p(pword_j)}] \text{(식 2)}$$

$$PMI(word_i, nword_j) = \log_2[\frac{p(word_i\ AND\ nword_j)}{p(word_i)\,p(nword_j)}] \text{(식 3)}$$

(식 2)와 (식 3)은 단어 $word_i$와 긍정, 부정 단어 간 각각의 연관 정도를 얻을 수 있다. 특정 단어뿐만 아니라 어절, 문장 간의 PMI값을 구할 수 있다.

이를 바탕으로 문장 혹은 문서의 극성(Sentiment Orientation)을 (식 4)를 통해 구할 수 있다.

$$SO(phrase) = PMI(phrase, pword_i) - PMI(phrase, nword_j) \quad \text{(식 4)}$$

if SO > 0, then phrase is positive polarity
else if SO < 0, then phrase is negative polarity
else, neutral

특정 단어, 문장, 또는 문서 등이 긍정 단어와 부정 단어와의 연관 정도를 각자 구한 뒤 그 두 연관 정도의 차를 이용하여 해당 단어, 문장, 또는 문서가 가지는 전체적인 극성이 무엇인지 수치로 나타내어 분류한다.

3. 감성 인식 시스템 구성

본 감성 인식 시스템의 주된 목적은 한글 문서에 대한 검색 대상 개체와 감성 용어 관계를 추출하는 것이다. 이를 위하여 시스템은 크게 문서 추출, 문장 분류, 태깅, 감성 분류 및 미분류 처리 시스템, 스팸처

11) Bo Pan, Lillian Lee (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, In proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, pages 271-278, Barcelona, ES.

12) A. Esuli, F. Sebastiani (2005). Determining the semantic orientation of terms through gloss classification., CIKM 2005, 17-624.

13) Peter D. Turney (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings 40th Annual Meeting of the ACL, pp. 417-424.

14) Christopher D. Manning, Hinrich Schutze (1999). Foundations of Statistical Natural Language Processing, MIT Press.

15) Hull, D.A, (1994). Improving text retrieval for the routing problem using latent semantic indexing. In Proceedings of SIGIR 94: 282-289.

16) Corinna Cortes and V. Vapnik (1995). Support-Vector Networks, Machine Learning.

리, 사용자 정의부분으로 나누어진다.

3.1. 문서 인식 및 태깅 시스템

자연 언어를 태깅함에 있어서 자연 언어는 문장을 구성하는 어절이 같은 단어라도 여러 가지 품사를 가질 수 있다는 모호성이 있기 때문에 주위의 문맥을 고려해서 단어의 품사를 결정하는 작업이 이루어져야 한다.

문서에 대한 태깅 작업을 하기 전에, 웹에서 수집한 문서에 대하여 각종 태그를 제거하여야 한다. 해당 문서에서 태그를 제거하는 방법으로 현존하는 태그 패턴을 사전에 저장하여 웹으로부터 수집한 문서에서 텍스트를 추출하여 문서번호(D_i)를 주었다.

단어가 한 개체의 특성을 나타내는 것임에 비해 문장은 '개체가 어떠하다'는 것을 나타내 주므로, 문서 분류기는 태깅 작업을 하기 전에 문서로 부터 전처리(preprocessing)한 텍스트를 문장 단위(S_i)로 분류한다.

분류된 문장은 국립국어원에서 발행한 대규모의 말뭉치를 기반으로 만든 어절 및 형태소 혼합 기반 태깅 시스템을 통해 문장의 품사 열을 구한다. 어절기반 태깅 시스템은, 각 어절에 대해 직접 가장 적합한 품사를 구하는 시스템이다. 가령, "나는 하늘을 나는 새를 보았다."라는 문장에서 나는(대명사+주격조사), 하늘을(명사+목적격조사), 나는(동사+관형형전성어미), 새를(명사+목적격조사), 보았다(동사+선어말어미+종결어미) 식으로 분석하는 것을 말한다.

HMM(Hidden Markov Model) 관점에서 보면, 각 어절들은 출력 상태(output state)이며, 이에 대응하는 품사들은 은닉 상태(hidden state)이다.

하나의 문장은 여러 어절들의 집합이다. 그래서 문장 S_i에 대하여 다음과 같이 표현된다.

문장 S에 대하여,

$$S_i = w_1, w_2, ..., w_k \quad \text{(식 5)}$$

여기에서 w_i, $i=1,k$는 문장을 구성하는 어절을 의미함. 그리고 문장을 이루는 각 어절에 대한 적절한 품사는

$$w_i = c_1, c_2, ..., c_l \quad \text{(식 6)}$$

여기에서 c_i, $i=1,l$는 어절을 구성하는 품사를 의미함. (식 7)은 한 품사에서 다른 품사가 나올 확률이고 (식 8)는 한 품사에서 한 어절이 나올 확률을 나타낸다.

$$P(c_i|c_{1,i-1}, w_{1,i-1}) \simeq P(c_i|c_{i-1}) \quad \text{(식 7)}$$

$$P(w_i|c_{1,i}, w_{1,i-1}) \approx P(w_i|c_i) \quad \text{(식 8)}$$

Markov 가정에 의하여, (식 7)은 현재 품사의 발생 이전 품사에만 의존하며, (식 8)는 현재 어절에서 발생된 현재 품사에 의해서 결정된다는 것을 알 수 있다.

결국, (식 9)처럼 한 품사에서 다른 품사로의 전이 확률과 한 품사에서 한 어절이 나올 확률의 곱을 최대로 하는 어절 및 품사가 태깅 결과가 된다.[17]

$$T(S) \simeq \arg\max_{w,c} \prod_{i=1}^{n} P(c_i|c_{i-1}) P(w_i|c_i) \quad \text{(식 9)}$$

여기서 S는 문장, T(S)는 문장 S에 대한 태깅 결과를 말한다. 그러나 어절기반 태깅 시스템은 자료부족 문제가 발생하기 쉽다. 자료부족 문제란, 모든 다양한 어절을 데이터로 저장할 수 없어 아직 분석되지 않은 어절에 대해 품사후보를 얻지 못하는 것을 말한다.

이를 보완하기 위해 분석되지 않은 어절에 대해서는 형태소 기반 시스템을 도입하여 분석하였다.

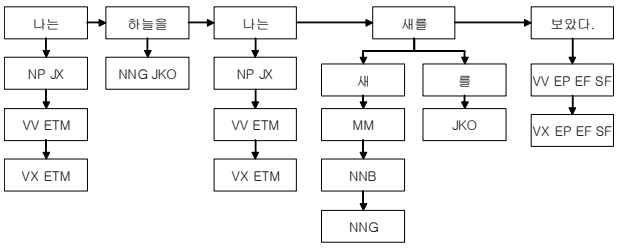

그림 2. 어절-형태소 혼합 기반 HMM

(그림 2)는 어절 기반 태깅 시스템의 모델이다. 각 어절에 대해 품사 후보를 두고, 문맥에 맞게 적절한 품사를 결정한다. 그러나 '새를'이라는 어절에 대해 품사후보 정보를 갖고 있지 않은 경우 자료부족문제가 발생하면, 어절 '새를'을 형태소 단위로 분리하여, 각 형태소에 대한 품사후보를 구한 후, 주변 어절의 품사후보들과 관계를 고려하여 적절한 품사를 결정하도록 한다.

형태소 단위로 데이터를 가지고 있으면, 어절 단위보다 비교적 적은 데이터를 보유해도 되기 때문에 데이터 관리가 용이하고, 자료부족문제도 보완할 수 있다.

17) 김진동, 임희석, 임해창 (1997). Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델, 정보과학회논문지(B), pp. 1502~1512.

3.2. 용어 인식 시스템

3.2.1. 개체명 인식

대량의 문서집합으로부터 사용자가 요구한 특정 문서나 특정 사실을 사용자에게 제공해 주기 위해서 문서나 사실 정보, 질의 속의 핵심어를 주요 검색 대상으로 한다. 그러나 핵심 대상이 되는 개체명들은 형태소의 품사로 고유명사이거나 미등록어인 경우가 많으며, 항상 새롭게 만들어지고, 때로는 같은 단어라도 사용되는 문장에 따라 상이한 의미를 가지고 오기 때문에 다른 일반 품사들처럼 사전으로 구축하여 사용하기 어렵다.[18)19)]

이런 이유로 해당 문서의 범주를 명확히 분류해 줄 필요가 있으며, 개체명들의 의미 범주를 결정하는 작업을 개체명 인식이라 한다. 본 시스템에서는 개체명을 추출하는 과정을 형태소 분석을 통하여 개체명 사전 검색을 하는 단계와 미등록어 추정 단계로 구분할 수 있다.

시스템 내에서 범주화의 도메인을 <표 1>과 같이 사람과 사물로 하여, 회사는 공기업(27), 외국 기업(695), 상장기업(541), 코스닥 상장기업(760) 그리고 그룹 내 계열사(1734)의 기업명과 조직명으로 한정하여 사전을 구성하였다. 인명에 관한 범주는 인명의 고유명사 사전과 직책명을 사전 리스트에 등록하였다. 또한 개체명에 대한 추가 삭제를 (그림 3)에 보인 바와 같이 별도의 인터페이스를 두어서 사용자가 수정을 할 수 있도록 하였다.

표 1. 개체명의 범위(개체명 종류(속성 번호))

개체명	사람	고유명사	인명(0), 직책명(1)
		대명사	인칭 대명사(2)
	사물	고유명사	지명(3), 국명(4), 기관명(5), 회사명(6), 제품명(7), 조직명(8)
		대명사	지시대명사(9)

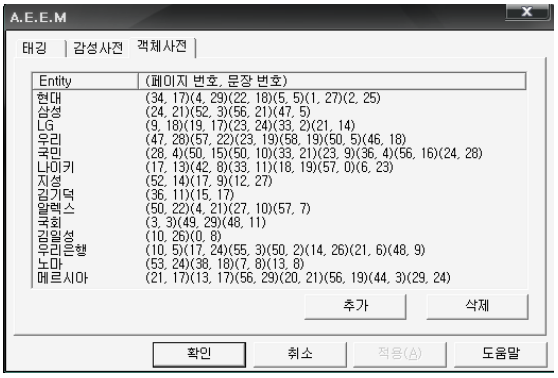

그림 3. 개체명 사전

3.2.2. 감성 용어 구성

한국어 감성 단어에서 특히 감정에 관련된 단어를 연구한 박인조(2005)는 434개의 한국어 감정단어 목록을 만들었다. 또한 한국어 정서용어들의 적절성과 경험빈도를 연구한 한덕웅(2000)의 감성단어 연구도 있다.

과학수업에 감성을 적용한 사례 연구의 내용에서 제시된 학습자의 감성 단어는 한혜영(2005), 권용주(2004) 등에 의해 선행 연구가 되었다. 본 시스템에서는 각 문헌에서 수집된 기본적인 한국어 감성 단어 700여 개와 국립 국어원에서 제공한 사전으로부터 극성(긍정, 부정)을 지니는 명사 2,199개, 동사 753개, 형용사 813개의 단어에 <표 2>와 <표 3>에서와 같이 2개의 관점으로 용어의 속성값을 주었다. <표 3>는 OCC 모델에서의 22개의 감정 도메인 중 대표성을 지닌 10개의 감정 도메인(domain of emotion)으로 기쁨(joy), 안심(relief), 만족(satisfaction), 재미(enjoy), 긍지(pride)를 긍정적인 도메인(positive domain)으로 하였고 분노(anger), 공포(fear), 혐오(dislike), 불만(dissatisfaction), 슬픔(sad) 등을 부정적 도메인(negative domain)으로 구분하였다.

감성용어는 단어의 속성상 공기 관계를 나타내는 것이 있고 사람, 사물, 사건에 대하여만 관계를 나타내는 속성을 지닌 용어들이 있다. 이를 위하여 감성 단어의 속성(TP: Term Property)을 <표 2>와 같이 4가지 기준으로 주어서 (그림 4)와 같이 나타내었다. <표 3>을 이용한 <예제 1>의 예제들과 같이 단순한 단어만으로 문장의 감정을 표현을 나타낼 수 있다. 그러나 이러한 문장은 전체 텍스트에서 극히 일부분이다. 또한 감성을 나타낼지라도 일반적인 기사문이 아닌 문장인 경우의 <예제 1>의 (1)~(3)번 예제와 같이 문법에 어긋난 경우가 발생하여 비문을 정제하는 작업이

18) 이경희, 이주호, 최명석, 김길창 (2000). 한국어 문서에서 개체명 인식에 관한 연구, 한글 및 한국어 정보처리 학술대회, pp.292-299.

19) 선충녕 (2002). 신경망과 규칙을 이용한 한국어 개체명 인식 시스템의 구현, 서강대학교 대학원 컴퓨터학과, 석사학위논문.

필요하다. 또한 <예제 1>의 (5)번 예제와 같이 감성을 표현하는 단어가 있을지라도 감성 단어가 있는 위치, 단어와 단어와의 관계에 따라 그 문장의 감성 표현 정도가 다를 수 있다. 따라서 이를 정립해 주는 공기 정보시스템이 필요하다.

예제 1. 일반적인 감성의 예

<신문 기사>

(a) 정확히 말하자면 스스로 출국을 거부했다.
=> 개체명 스스로(인칭) 부정
(b) 두꺼운 입술이라는 뜻의 '사치모'라는 별명을 얻은 그는 경쾌한 트럼펫 연주로 대중의 인기를 모으며 50여 편의 영화에 출연했을 뿐만 아니라, 라디오 쇼를 진행하고 국제무대 공연을 통해 최고의 스타로 발돋움했다.
=> 개체명 그(인칭) 긍정
(c) 그리고 관객은 감격했다.
=> 개체명 관객(인칭 대명사) 긍정
댓글
(1) 간편하면서도 아주 굿굿
=> 개체명 긍정
(2) 국회의원 행세 하는 꼬라지가 가소롭다
=> 개체명 국회의원(직책명) 부정
(3) mp3에 직결로도 충분히 소리내주구요..
=> 개체명 mp3(제품명)
아무튼 강추합니다.
=> 개체명 긍정
(4) ***들이 많이 거주하는 대림동, 구로동 ,방면을 운행중인 택시들이 술취한 ***인 경우 승차거부를 하고 피한다 고 합니다.
=> 개체명 *** 부정
(5) 배송된 상품이 나쁘지 않네요.
=> 개체명 상품 중립

표 2. 감성 용어 속성

TP	내용
1	감성 단어만으로 표현을 나타내는 용어
2	감성 용어와 동사의 관계를 지닌 용어
4	복합 단어에 의하여 감성을 나타내는 용어
8	사람, 사물, 사건에 대한 공기 관계로 특성을 지닌 용어

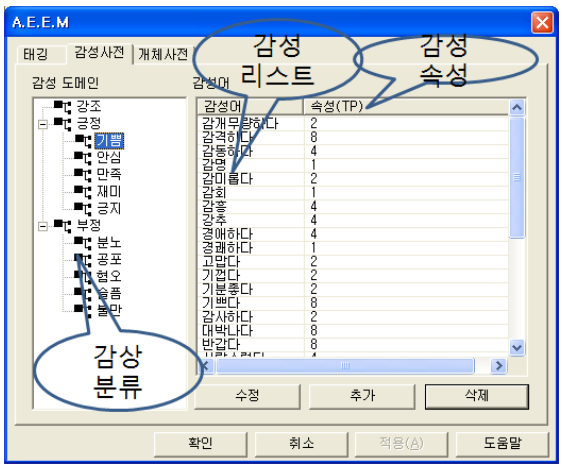

그림 4. 감성 사전 UI

표 3. 한국어 감성 도메인

긍정	기쁨 (a)	감개무량, 감격, 감명, 감흥, 강추, 경애, 경쾌 ..
	안심 (b)	고객감동, 공감하다, 괜찮다. 교감하다. 느끼다. 맘 놓다, 미소, 믿다, 안도하다, 안락하다. 안심하다….
	만족 (c)	가격만족, 가격저렴, 값싸다, 경탄하다, 굿, 귀엽다, 깔끔하다, 깨끗하다, 친절하다…..
	재미 (d)	매료되다, 매혹적이다, 박진감, 반하다, 살맛나다, 설레다, 신나다, 열광하다, 열렬하다, 열애…
	긍지 (e)	가슴뿌듯하다, 감복하다, 감사하다, 경외하다, 감탄하다, 긍지, 떳떳하다, 보람차다, 뿌듯하다….
부정	분노 (f)	가증스럽다, 개탄하다, 격노하다, 노기, 격분하다, 고깝다, 괘씸하다, 골나다, 굴욕, 어이없다
	공포 (g)	강압, 경악하다, 곤혹스럽다, 공포, 기겁하다, 끔찍하다, 무섭다, 두렵다, 무시무시하다, 섬뜩하다
	혐오 (h)	개쓰레기, 거부감, 경멸하다, 가소롭다, , 기만하다, 농락하다, 기고만장하다, 꺼리다, 당혹하다, 만행, …
	슬픔 (i)	가련하다, 가엾다, 고뇌하다, 고독하다, 고적하다, 고롭다, 낙담하다, 그립다, 낙망하다, 낙심하다, 망연자실하다..
	불만 (k)	가입거부, 가짜, 감언이설, 거부, 거절당하다, 결점, 결함, 계약위반, 고발하다, 고소하다, 과실, 권태롭다.

3.3. 감성 제어 시스템

3.3.1. 문장 공기 시스템

제안한 시스템은 대량의 문서로부터 개체에 대한 감성이입의 여부를 문장 단위로 판단하여 개체에 해당하는 감정적 요소를 나타낸다. 문장의 감성적 기준은 감성 용어 시스템에서 설명했듯이 긍정 부정 대분류 내에서 OCC 모델 내의 22개 감성 중 10개를 감성 도메인으로 긍정 부정을 분류하였다.

감성적 분류를 할 때 감성적 용어만으로 감성을 나타내는 것도 있지만, 감성적 용어와 동사들 간에 공기를 이루어 극성을 나타내는 경우가 많다. 또한 특정 용어는 용어에 대응하는 개체와 공기 관계를 이룬다.

본 시스템에서는 (식 9)의 결과 값이 출력되었을 때 6가지 단계로 문장 추출기로부터 개체, 극어, 감성어, 동사의 용어를 추출하는 데 순서는 다음과 같다.

1. 감성 단어를 포함한 단일 문장
2. 감성 용어와 동사의 관계
3. 공기 정보를 이용한 사람 및 개체에 대한 감성

표현 추출
4. 감성 용어와 극어와의 위치 관계
5. 전위 관계를 가진 복합문 처리
6. 중립적인 표현-평서문, 미분류

단계 1에서는 감성 용어 속성(TP)의 값이 1로 분류된 강한 부정 및 긍정 의미를 가진 단어를 포함하는 문장을 분류한다(예제 1). 그러나 대다수의 용어들은 긍정 및 부정에 대한 양면성을 가지고 있거나 <표 3>에서 나타냈듯이 개체의 관계에 따라 감성 의견이 표출된다. 또한 단어와 단어 사이의 연결 관계이외에도 단어의 위치에 따라 문장의 성격을 달리 나타낼 수 있다. 따라서 <단계 2>, <단계 3>, <단계 4>를 구한 경우, 부정 의미를 가진 용어가 있더라도, 부정적 요소가 없는 문장으로 나타날 수 있다.

<예제 2>에서 보듯이, 부정 극어 <표 4> 중 '아무도'는 대표적인 부정 극어로서 부정 극어를 허용하는데 있어서, 부정어의 위치가 문장의 마지막 부분에 위치해야 부정문으로 성립된다.[20]

이를 처리하기 위해서 (식 9)를 통해 추출된 극어와 관계되는 주위의 단어군을 찾아 관계가 성립되는 경우 공기 관계로 정리하였다. 본 시스템에서는 동사와 부정문의 대표적인 부정 극어의 단어군을 <표 4>[21]와 같이 단어형 극어와 구형 극어로 형성하여 부정적 요소를 지닌 단어 연관성을 공기 관계로 (그림 5)와 같이 나타내었다.

위와 같이 하나의 문서를 여러 개의 문장으로 분리한 뒤 반자동으로 문장 내의 개체와 이 개체와 관계를 지닌 명사, 형용사나 부사, 동사 품사 용어의 쌍을 이루어 단어구의 성격을 분류하고, SO-PMI(Semantic Orientation Pointwise Mutual Information)를 계산을 한 다음, 이를 단어 연관성 리스트에 추가하였다.

표 4. 대표적인 부정 극어

단어형 극어 :전혀, 영, 통, 도무지, 여간, 이/그다지, 미처, -밖에, 별반, -커녕, 차마, 절대, 도무지, 도통, 아무런, 당최, 아무도, 웬만해서(는), 여간해서(는), 아무도, 결코, 결단코, 도저히, 좀처럼, 좀체, 과히, 별로, ..
구형 극어 : 숨 한 번도, 동전 한 푼(도), 땡전 한 푼(도), 한 치도, 한 참도, 한 순간도, 한 푼도…, 하나(도), 눈 하나(도), 손 하나(도), 머리카락 하나(도)…, 눈곱만큼도, 손톱만큼도, 티끌만큼도, 털끝만큼도, 조금도…,

20) 임홍빈 (1987). 국어 부정문의 통사와 의미, 국어 생활 10호.
21) 구종남 (1992). 국어 부정문 연구, 전북대 박사학위 논문.

예제 2 '아무도' 부정 극어 사용

아무도 그 사실을 부인하지 않았다. (a) 아무도 감히 생각하지 못한 고르바초프적 신사고가 하나의 현실로 다가오고 있다. (b)

3.3.2. UI(User Interface)

본 시스템은 자동적으로 문서에 대한 감성 지수를 측정하는 시스템이다. 그러나 학습 활동을 통하여 개체 추출과 문장에 대한 규칙을 추출(rule extraction)을 하지만 학습량이 부족할 경우 오류가 발생할 수 있다. 오류를 수정하기 위하여 UI에서는 자동 감성 분류 기능뿐만 아니라 각종 사전에 대한 용어 추가 및 개체 추가, 오류에 대한 문장 규칙 수정 그리고 미분석 문장 및 용어에 대한 처리를 수행한다. 또한 대부분 문장 미분류 오류는 띄어쓰기 오류로서 기존의 오타에 대하여 수정한 것을 기록해 두어서 미분석 용으로 사용하였다.

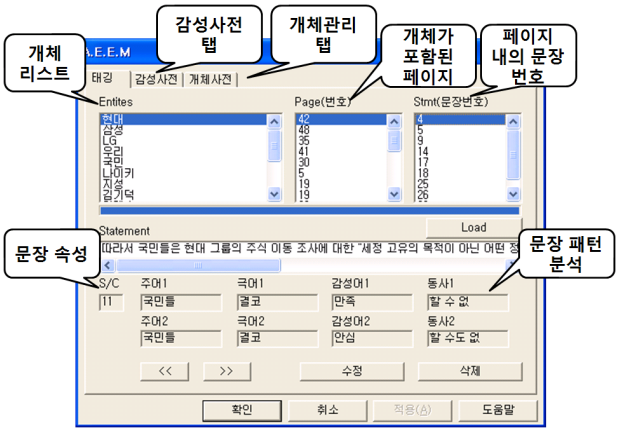

그림 5. 감성 사전 UI

3.3.3. 스팸 처리

문서와 관계없는 광고나 의견 내용과 관계없이 자기의 의견의 가중치를 올리기 위해서 한 사용자가 여러 번의 내용을 반복적으로 올리는 등 평가 시스템을 저해 시키는 요소가 있다. 이 문제를 해결하기 위하여 같은 아이디인 경우 아이디에 대한 중복적인 내용을 제외 및 URL이 들어간 댓글은 기본적으로 제외하였다.

3.3.4 데이터 구성 및 실험

문서의 극성(polarity)을 결정하는 주요한 품사는 형용사(adjective)이다. 형용사는 한국어에서 정서적 표현을 나타내는 용언이다. 하지만 특정 용언이 특정 주체를 서술함으로써 새로운 의미를 가질 수 있으므로 주체와 이 주체를 서술하는 서술어 관계(주체, 서술어)로 문장의 극성을 분류하는 것이 더 명확하다. 그러므로 각 품사별로 결합하여 상품평의 감성을 판별하였다. 실험에 사용된 데이터는 현재 온라인 상에서 운영중인 대표 쇼핑몰 사이트들의 상품 댓글, 대표 가격비교 사이트에서의 사용자 후기 및 소비자가 만드는 신문, 고객 센터에서 2,000개의 문서 및 댓글을 사용하였다. 이들 개체 분야로 헤드폰 500건, 노트북 500건, MP3 플레이어 500건, 디지털 카메라 500건 4개의 개체에 대해서 각 125개의 문서로 나누어 시스템의 성능을 평가하였다. 시스템의 성능은 정확률(precision), 재현율(recall) 그리고 정확성(accuracy)로 표현하였다. 정확률은 감성 범주(긍정, 부정)별로 분류 결과가 얼마나 정확한지에 대한 것이며, 재현율은 감성 범주 별 분류 할 문서들 중 실제로 정확히 분류된 정도를 의미한다. 마지막으로 정확성은 모든 감성 범주에 대해 얼마만큼 정확히 분류했는지를 말한다. 이를 각각 수식으로 정의하면 다음과 같다.

$$\text{Precision}(S_i) = \frac{C(S_i)}{R(S_i)} \quad \text{(식 10)}$$

$$\text{Recall}(S_i) = \frac{C(S_i)}{N(S_i)} \quad \text{(식 11)}$$

$$\text{Accuracy} = \frac{C(S_P) + C(S_N)}{N(S_P + S_N)} \quad \text{(식 12)}$$

S_i : 감성 범주(i={P(Positive), N(Negative)})
$R(S)$: 감성 범주 S라 분류 된 문서의 수
$C(S)$: 감성 범주 S에 대해 정확히 분류된 문서 수
$N(S)$; 분류에 사용된 감성 범주 S인 문서의 수

분류 실험할 때 사용된 500개의 댓글에서 기존 사전에 없는 새로운 어휘가 발견되면 감성 사전에 추가하여 학습시키는 방법으로 하여서 <그림 6~10>로 각 긍정, 부정에 대한 정확률과 재현율 그리고 전체 분류의 정확성을 구하였다. 그림이 보이는 바와 같이 학습(training)을 시킬 때 마다 성능이 높아지는 것을 알 수 있으며, 테스트(test) 되어지는 임의의 문서의 수에 관계없이 평균 70~80%의 정확성을 보인다.

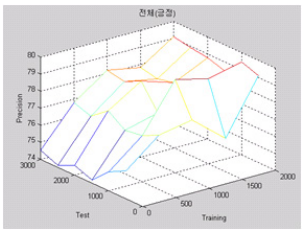

그림 6. Positive Precision

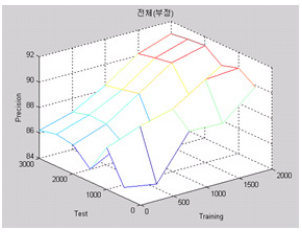

그림 7. Negative Precision

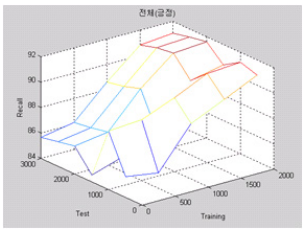

그림 8. Positive Recall

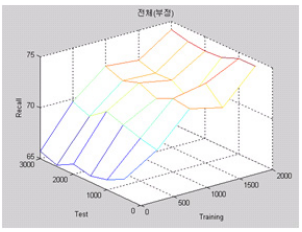

그림 9. Negative Recall

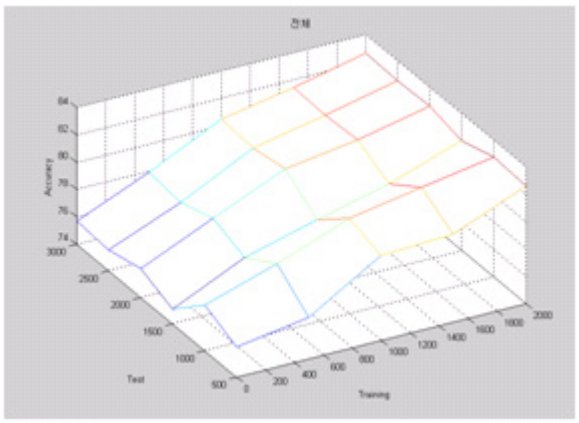

그림 10. Accuracy

4. 결론

인터넷을 통해 일반 대중이나 소비자의 의견을 수집하고 빠르게 분석해야 할 필요성이 점점 증가하고 있다. 본 논문에서는 온라인 텍스트 문서상에서 나타난 (주어진) 개체명에 대해 이 개체를 표현한 감성어 및 감성어에 관련된 문장의 구성, 단어 간의 관련성 등을 고려한 감성인식시스템을 제안하였고, 유연성 있는 사용자 인터페이스를 만들었다. 또한 감성인식 시스템을 바탕으로 SO-PMI 값을 이용하여 극성을 주었다. 품사의 영향도에서는 명사, 동사+형용사 등 한 개의 품사만 사용할 때는 정확도, 재현율, 정확성에 대해 낮은 값을 발생하나, 단어들 간의 관계를 나타내는 관계 용언과 결합하면 70% 이상의 결과 값을 얻는 것을 알 수 있다. 그러나 관계 용언이라는 것은 사람

이 반자동적으로 학습을 시켜 주는 것으로, 학습 방향에 따라서 좀 더 좋은 결과 값을 추출할 수 있다는 것을 알 수가 있다. 그러나 현재 구축된 시스템은 다양한 형식에 의해 표현된 감성을 제대로 추론하지 못하는 부분이 있고, 같은 감성 표현이라도 상황에 따라 달라질 수 있는데 이 모든 경우를 고려하지 못했다. 또한 감성의 중립적인 표현에 대한 정의를 아직 내리지 못했다. 중립적인 글이란 '긍정이다, 부정이다'라고 정확히 정의 내리지 못한 글로써, 감성은 가지고 있으면서, 약한 극성을 가지는 것이다. 이를 위해서, 감성에 영향을 주는 단어 정보들 간의 관계를 파악하여 감성 간의 구성 관계를 확장하는 연구로서 두 극성을 가지는 단어에 대한 연구와 극성에 대한 오류를 야기하는 단어의 패턴에 대한 연구를 할 필요성이 있다.

참고문헌

- 구종남 (1992). 국어 부정문 연구, 전북대 박사학위논문.
- 권용주, 신동훈, 한혜용, 반윤복 (2004). 과학적 관찰과 규칙성 발견 활동에서 나타나는 감성 단어 유형과 과학 지식 생성력과의 관계, 한국과학교육학회지, 제 24권(6호). pp.1106-1117.
- 김은영 (2005). 현대 국어 감정동사의 범위와 의미 특성에 대한 연구. 한국어 의미학, 16, 99-124.
- 김진동, 임희석, 임해창 (1997). Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델, 정보과학회논문지(B). pp. 1502~1512.
- 박인조, 민경환 (2005). 한국어 감정단어의 목록 작성과 차원 탐색, 한국심리학회지, 사회 및 성격, 제 19권(1호). pp. 109-129.
- 선충녕 (2002). 신경망과 규칙을 이용한 한국어 개체명 인식 시스템의 구현, 서강대학교 대학원 컴퓨터학과, 석사학위논문.
- 이경희, 이주호, 최명석, 김길창 (2000). 한국어 문서에서 개체명 인식에 관한 연구, 한글 및 한국어 정보처리 학술대회, pp.292-299.
- 임홍빈 (1987). 국어 부정문의 통사와 의미, 국어생활 10호.
- 한덕웅, 강혜자 (2000). 한국어 정서 용어들의 적설정과 경험 빈도", 한국심리학회지, 일반, 제 19권(2호). pp 3-99.
- 한혜영 (2005). 귀납적 과학지식의 생성과정에서 나타나는 감성분석을 위한 측정도구 개발, 한국교원대학교 대학원 석사학위논문.
- A. Esuli, F. Sebastiani (2005). Determining the semantic orientation of terms through gloss classification., CIKM 2005, 17-624.
- Bo Pan, Lillian Lee (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, In proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, pages 271-278, Barcelona, ES.
- Christopher D. Manning, Hinrich Schutze (1999). Foundations of Statistical Natural Language Processing, MIT Press.
- Corinna Cortes and V. Vapnik (1995). Support-Vector Networks, Machine Learning.
- Esuli A., Sebastiani F. (2006). SentiWordNet: A Publicily Avaliable Lexical Resource for Opinion Mining, In Proceeding of LREC-06, 5th Conference of Language Resource and Evaluation, Genova, IT, pp. 417-422.
- Hatzivassiloglou V., Mackeown K. (1997). Predicting the Semantic Orientation of Adjectives, In Proc. of 35th ACL/8th EACL.
- Hull, D.A. (1994). Improving text retrieval for the routing problem using latent semantic indexing. In Proceedings of SIGIR 94: 282-289.
- Ortony, A., Clore, G.L., and Collins, A. (1998). The Cognitive Structure of Emotions., Cambridge University Press.
- Peter D. Turney (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings 40th Annual Meeting of the ACL, pp. 417-424.
- Schlosberg, H. (1952). The descriptions of facial expressions in terms of two dimensions. Journal of Experimental Psychology,. 44, pp. 229-237.

원고접수 : 09.09.30
수정접수 : 09.10.17
게재확정 : 09.10.20