(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2013-0103249
(43) 공개일자 2013년09월23일

(51) 국제특허분류(Int. Cl.)
G06F 17/27 (2006.01) *G06F 17/10* (2006.01)
(21) 출원번호 10-2012-0024733
(22) 출원일자 2012년03월09일
심사청구일자 2012년03월09일

(71) 출원인
가톨릭대학교 산학협력단
서울특별시 서초구 반포대로 222, 가톨릭대학교 성의교정내 (반포동)
(72) 발명자
강행봉
서울특별시 용산구 이촌1동 강촌아파트 101-1108
조상현
인천광역시 서구 연희동 우성아파트 102동 1702호
(74) 대리인
김건우

전체 청구항 수 : 총 8 항

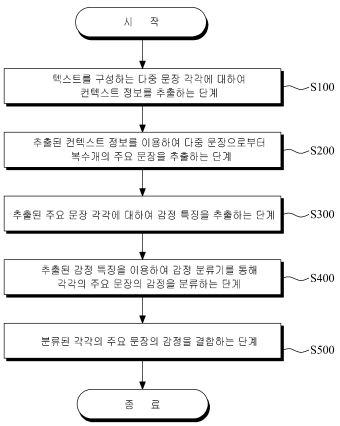
(54) 발명의 명칭 **컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법**

(57) 요 약

본 발명은 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법에 관한 것으로서, 보다 구체적으로는 (1) 텍스트를 구성하는 다중 문장 각각에 대하여 컨텍스트 정보를 추출하는 단계, (2) 추출된 컨텍스트 정보를 이용하여 다중 문장으로부터 복수 개의 주요 문장을 추출하는 단계, (3) 추출된 주요 문장 각각에 대하여 감정 특징을 추출하는 단계, 및 (4) 추출된 감정 특징을 이용하여 감정 분류기를 통해 각각의 주요 문장의 감정을 분류하는 단계를 포함하는 것을 그 구성상의 특징으로 한다.

본 발명에서 제안하고 있는 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법에 따르면, 컨텍스트 정보를 이용하여 텍스트를 구성하는 다중 문장으로부터 주요 문장을 추출하고, 추출된 주요 문장에 대하여 감정을 분류하고, 분류된 감정을 결합함으로써 온라인 상에서 수집할 수 있는 다중 문장으로부터 감정을 정확하게 분류하여, 마케팅 전략에 활용할 수 있다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호	1091205006109960010
부처명	문화체육관광부
연구사업명	문화기술연구소 육성 사업
연구과제명	비쥬얼커뮤니케이션을 위한 차세대 실감형 마이크로블로그 트위터 콘텐츠 제작 기술연구
주관기관	가톨릭대학교
연구기간	2010.05.15 ~ 2012.03.31

특허청구의 범위

청구항 1

(1) 텍스트를 구성하는 다중 문장 각각에 대하여 컨텍스트 정보를 추출하는 단계;

(2) 추출된 상기 컨텍스트 정보를 이용하여 다중 문장으로부터 복수 개의 주요 문장을 추출하는 단계;

(3) 추출된 상기 주요 문장 각각에 대하여 감정 특징을 추출하는 단계; 및

(4) 추출된 상기 감정 특징을 이용하여 감정 분류기를 통해 각각의 상기 주요 문장의 감정을 분류하는 단계를 포함하는 것을 특징으로 하는, 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법.

청구항 2

제1항에 있어서,

(5) 분류된 각각의 상기 주요 문장의 감정을 결합하는 단계를 더 포함하는 것을 특징으로 하는, 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법.

청구항 3

제1항에 있어서, 상기 단계 (1)은,

(1-1) 하기의 수학식을 이용하여 문장에 포함된 키워드 정보를 산출하는 단계;

$$\alpha(S)=\frac{1}{K}\sum_{i}\omega_i k_i(S)$$

(여기서, S는 입력문장이고, k_i는 입력문장 S에 포함되어 있는 i번째 키워드, $\omega_i \in R$는 키워드 가중치이며, K는 입력문장 S에 포함되어 있는 키워드 개수이다.)

(1-2) 하기의 수학식을 이용하여 상기 텍스트 내에서 문장의 위치에 대한 가중치를 산출하는 단계;

$$\beta(S)=\begin{cases} e^{-\frac{\mathrm{index}(S)}{T}} & \text{if } \mathrm{index}(S) \le \frac{T}{2} \\ e^{\frac{\mathrm{index}(S)}{T}} & \text{if } \mathrm{index}(S) > \frac{T}{2} \end{cases}$$

(여기서, $\mathrm{index}(S_i)$는 주요문장 S_i의 인덱스이고, T는 텍스트 안의 문장의 수이다.)

(1-3) 하기의 수학식을 이용하여 문장 간의 감정 변화도를 산출하는 단계; 및

$$\gamma(S)=\begin{cases} 1 & \text{문장의 감정이 변하지 않을 때} \\ e^{n(S_{ps})} & \text{문장의 감정이 변할 때} \end{cases}$$

(여기서, $n(S_{ps})$은 동일한 감정을 가지는 앞 문장의 수이다.)

(1-4) 산출된 상기 키워드 정보, 문장의 위치에 대한 가중치 및 문장 간의 변화도를 이용하여 하기의 수학식에

의해 문장의 컨텍스트 정보를 산출하는 단계

$$C(S) = \alpha(S) \cdot \beta(S) \cdot \gamma(S)$$

를 포함하는 것을 특징으로 하는, 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법.

청구항 4

제1항에 있어서, 상기 단계 (3)은,

(3-1) 형태소 분석기를 이용하여 문장에 포함된 단어를 품사별로 분류하는 단계; 및

(3-2) 감정 사전을 이용하여 상기 단어에 대한 감정 특징을 추출하는 단계를 포함하는 것을 특징으로 하는, 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법.

청구항 5

제4항에 있어서, 상기 단계 (3-2)에서,

상기 감정 사전은 어휘 사전 기반의 형식적 감정 사전에 도메인 기반 감정 사전을 추가하여 구축된 것을 특징으로 하는, 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법.

청구항 6

제4항에 있어서, 상기 감정 사전은,

명사, 동사, 형용사, 부사 및 이모티콘별 감정 특징을 포함하는 것을 특징으로 하는, 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법.

청구항 7

제6항에 있어서, 상기 이모티콘은,

불규칙적으로 자주 사용되는 이모티콘 중 가장 간단한 이모티콘 형태인 참조 이모티콘인 것을 특징으로 하는, 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법.

청구항 8

제7항에 있어서,

문장에 포함된 불규칙적인 이모티콘은 베이지안 프레임워크를 이용하여 상기 참조 이모티콘으로 변환하여 감정 특징을 추출하는 것을 특징으로 하는, 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법.

명 세 서

기 술 분 야

[0001] 본 발명은 감정 분류 방법에 관한 것으로서, 보다 구체적으로는 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법에 관한 것이다.

배 경 기 술

[0002] 스마트폰의 대중적인 보급으로 인해 트위터, 페이스북과 같은 소셜 네트워크 서비스(Social Network Service; SNS)가 보편화됨에 따라 다양한 주제에 대하여 수많은 의견들이 실시간으로 개진되고 있다. SNS는 기존의 인맥

을 강화하고 새로운 인맥을 형성하여 폭넓은 인적 네트워크를 형성할 수 있도록 해주는 서비스로서, 많은 사람은 이와 같은 서비스를 통해 서로에게 댓글을 달아주는 형태로 막대한 양의 텍스트 정보를 생성하고 있다.

[0003] 최근에는, 상품에 대한 리뷰(review), 영화 감상평, 음식 평가 등의 주요 이슈에 대하여 바이럴 마케팅(viral marketing), 즉 입소문을 통한 마케팅 전략이 많이 이용되고 있는바, 이와 같은 SNS 정보로부터 소비자들의 의견을 정확히 판단하는 것이 마케팅 전략에 매우 중요한 것으로 인식되고 있는 실정이다.

[0004] 따라서 SNS 사용자들에 의해 작성된 막대한 텍스트들로부터 의미 있는 정보를 찾기 위한 연구가 관심의 대상이 되고 있고, 특히, 문장에 담겨 있는 감정은 활용 범위가 매우 넓은 정보인바, 문장으로부터 감정을 분류 또는 인식하는 연구가 이루어지고 있지만(공개번호 제10-2002-0042248호 참조), 매우 미약한 실정이다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 기존에 제안된 방법들의 상기와 같은 문제점들을 해결하기 위해 제안된 것으로서, 컨텍스트 정보를 이용하여 텍스트를 구성하는 다중 문장으로부터 주요 문장을 추출하고, 추출된 주요 문장에 대하여 감정을 분류하고, 분류된 감정을 결합함으로써 온라인 상에서 수집할 수 있는 다중 문장으로부터 감정을 정확하게 분류하여, 마케팅 전략에 활용할 수 있는, 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법을 제공하는 것을 그 목적으로 한다.

과제의 해결 수단

[0006] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른, 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법은,

[0007] (1) 텍스트를 구성하는 다중 문장 각각에 대하여 컨텍스트 정보를 추출하는 단계;

[0008] (2) 추출된 상기 컨텍스트 정보를 이용하여 다중 문장으로부터 복수 개의 주요 문장을 추출하는 단계;

[0009] (3) 추출된 상기 주요 문장 각각에 대하여 감정 특징을 추출하는 단계; 및

[0010] (4) 추출된 상기 감정 특징을 이용하여 감정 분류기를 통해 각각의 상기 주요 문장의 감정을 분류하는 단계를 포함하는 것을 그 구성상의 특징으로 한다.

[0011] 바람직하게는,

[0012] (5) 분류된 각각의 상기 주요 문장의 감정을 결합하는 단계를 더 포함할 수 있다.

[0013] 바람직하게는, 상기 단계 (1)은,

[0014] (1-1) 하기의 수학식을 이용하여 문장에 포함된 키워드 정보를 산출하는 단계;

[0015] $$\alpha(S)=\frac{1}{K}\sum_{i}\omega_i k_i(S)$$

[0016] (여기서, S는 입력문장이고, k_i는 입력문장 S에 포함되어 있는 i번째 키워드, $\omega_i \in R$는 키워드 가중치이며, K는 입력문장 S에 포함되어 있는 키워드 개수이다.)

[0017] (1-2) 하기의 수학식을 이용하여 상기 텍스트 내에서 문장의 위치에 대한 가중치를 산출하는 단계;

[0018]
$$\beta(S)=\begin{cases} e^{-\frac{\text{index}(S)}{T}} & \text{if index}(S) \le \frac{T}{2} \\ e^{\frac{\text{index}(S)}{T}} & \text{if index}(S) > \frac{T}{2} \end{cases}$$

[0019] (여기서, index(S_i)는 주요문장 S_i의 인덱스이고, T는 텍스트 안의 문장의 수이다.)

[0020] (1-3) 하기의 수학식을 이용하여 문장 간의 감정 변화도를 산출하는 단계; 및

[0021]
$$\gamma(S)=\begin{cases} 1 & \text{문장의 감정이 변하지 않을 때} \\ e^{n(S_{ps})} & \text{문장의 감정이 변할 때} \end{cases}$$

[0022] (여기서, n(S_{ps})은 동일한 감정을 가지는 앞 문장의 수이다.)

[0023] (1-4) 산출된 상기 키워드 정보, 문장의 위치에 대한 가중치 및 문장 간의 변화도를 이용하여 하기의 수학식에 의해 문장의 컨텍스트 정보를 산출하는 단계

[0024]
$$C(S)=\alpha(S)\cdot\beta(S)\cdot\gamma(S)$$

[0025] 를 포함할 수 있다.

[0026] 바람직하게는, 상기 단계 (3)은,

[0027] (3-1) 형태소 분석기를 이용하여 문장에 포함된 단어를 품사별로 분류하는 단계; 및

[0028] (3-2) 감정 사전을 이용하여 상기 단어에 대한 감정 특징을 추출하는 단계를 포함할 수 있다.

[0029] 더욱 바람직하게는, 상기 단계 (3-2)에서,

[0030] 상기 감정 사전은 어휘 사전 기반의 형식적 감정 사전에 도메인 기반 감정 사전을 추가하여 구축될 수 있다.

[0031] 더욱 바람직하게는, 상기 감정 사전은,

[0032] 명사, 동사, 형용사, 부사 및 이모티콘별 감정 특징을 포함할 수 있다.

[0033] 더욱더 바람직하게는, 상기 이모티콘은,

[0034] 불규칙적으로 자주 사용되는 이모티콘 중 가장 간단한 이모티콘 형태인 참조 이모티콘일 수 있다.

[0035] 더욱더 바람직하게는,

[0036] 문장에 포함된 불규칙적인 이모티콘은 베이지안 프레임워크를 이용하여 상기 참조 이모티콘으로 변환하여 감정 특징을 추출할 수 있다.

발명의 효과

[0037] 본 발명에서 제안하고 있는 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법에 따르면, 컨텍스트 정보를 이용하여 텍스트를 구성하는 다중 문장으로부터 주요 문장을 추출하고, 추출된 주요 문장에 대하여 감정을 분류하고, 분류된 감정을 결합함으로써 온라인 상에서 수집할 수 있는 다중 문장으로부터 감정을 정확하게 분류하여, 마케팅 전략에 활용할 수 있다.

도면의 간단한 설명

[0038] 도 1은 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법의 순서도.

도 2는 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법의 단계 S100에 대한 세부 순서도.

도 3은 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법의 단계 S300에 대한 세부 순서도.

도 4는 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법에서, 베이지안 프레임워크를 이용하여 불규칙 이모티콘을 처리하는 세부 흐름을 도시한 도면.

도 5는 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법에서, 문장의 감정 분류 성능 실험 결과를 도시한 도면.

발명을 실시하기 위한 구체적인 내용

[0039] 이하, 첨부된 도면을 참조하여 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 본 발명을 용이하게 실시할 수 있도록 바람직한 실시예를 상세히 설명한다. 다만, 본 발명의 바람직한 실시예를 상세하게 설명함에 있어, 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략한다. 또한, 유사한 기능 및 작용을 하는 부분에 대해서는 도면 전체에 걸쳐 동일한 부호를 사용한다.

[0040] 덧붙여, 명세서 전체에서, 어떤 부분이 다른 부분과 '연결' 되어 있다고 할 때, 이는 '직접적으로 연결' 되어 있는 경우뿐만 아니라, 그 중간에 다른 소자를 사이에 두고 '간접적으로 연결' 되어 있는 경우도 포함한다. 또한, 어떤 구성요소를 '포함' 한다는 것은, 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있다는 것을 의미한다.

[0041] 도 1은 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법의 순서도이다. 도 1에 도시된 바와 같이, 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법은, 텍스트를 구성하는 다중 문장 각각에 대하여 컨텍스트 정보를 추출하는 단계(S100), 추출된 컨텍스트 정보를 이용하여 다중 문장으로부터 복수 개의 주요 문장을 추출하는 단계(S200), 추출된 주요 문장 각각에 대하여 감정 특징을 추출하는 단계(S300) 및 추출된 감정 특징을 이용하여 감정 분류기를 통해 각각의 주요 문장의 감정을 분류하는 단계(S400)를 포함하여 구성될 수 있고, 분류된 각각의 주요 문장의 감정을 결합하는 단계(S500)를 더 포함할 수 있다.

[0042] 단계 S100에서는, 텍스트를 구성하는 다중 문장 각각에 대하여 컨텍스트 정보를 추출하며, 단계 S100의 세부적인 흐름에 대하여는 도 2를 참조하여 상세히 설명하도록 한다.

[0043] 도 2는 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법의 단계 S100에 대한 세부 순서도이다. 도 2에 도시된 바와 같이, 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법의 단계 S100은, 문장에 포함된 키워드 정보를 산출하는 단계(S110), 텍스트 내에서 문장의 위치에 대한 가중치를 산출하는 단계(S120), 문장 간의 감정 변화도를 산출하는 단계(S130) 및 산출된 키워드 정보, 문장의 위치에 대한 가중치 및 문장 간의 변화도를 이용하여 문장의 컨텍스트 정보를 산출하는 단계(S140)를 포함할 수 있다.

[0044] 단계 S110에서는, 하기의 수학식 1을 이용하여 문장에 포함된 키워드 정보를 산출한다.

수학식 1

[0045] $$\alpha(S) = \frac{1}{K}\sum_{i} \omega_i k_i(S)$$

[0046] 여기서, S는 입력문장이고, k_i는 입력문장 S에 포함되어 있는 i번째 키워드, $\omega_i \in R$는 키워드 가중치이며, K는 입력문장 S에 포함되어 있는 키워드 개수를 나타낸다. 키워드는 도메인별로 미리 수집된 문장을 통해 해당 어휘의 빈도를 측정하여 이 빈도를 가중치(0~1)로 부여한다.

[0047] 단계 S120에서는, 텍스트 내에서 문장의 위치에 대한 가중치를 산출한다. 보다 구체적으로, 텍스트를 작성하는 사람은 일반적으로 자신의 의견을 텍스트의 첫 문장이나 끝 문장을 통해 표현하는 경우가 많기 때문에, 텍스트 내에서 문장의 위치는 텍스트의 전체 감정을 추정하는데 매우 중요한 정보인바, 하기의 수학식 2를 이용하여 텍스트 내에서 문장의 위치에 대한 가중치를 산출한다.

수학식 2

[0048] $$\beta(S) = \begin{cases} e^{-\frac{\text{index}(S)}{T}} & \text{if } \text{index}(S) \le \frac{T}{2} \\ e^{\frac{\text{index}(S)}{T}} & \text{if } \text{index}(S) > \frac{T}{2} \end{cases}$$

[0049] 여기서, $\text{index}(S_i)$는 주요문장 S_i의 인덱스이고, T는 텍스트 안의 문장의 수를 나타낸다.

[0050] 단계 S130에서는, 문장 간의 감정 변화도를 산출한다. 보다 구체적으로, 문장의 감정이 유지되다가 갑자기 감정의 변화가 생기는 경우 문장 전체의 감정이 변하는 경우가 많기 때문에, 문장 간의 감정 변화도 또한 전체 문장의 감정을 추정하는데 매우 중요한 정보인바, 하기의 수학식 3을 이용하여 문장 간의 감정 변화도를 산출한다.

수학식 3

[0051] $$\gamma(S) = \begin{cases} 1 & \text{문장의 감정이 변하지 않을 때} \\ e^{n(S_{ps})} & \text{문장의 감정이 변할 때} \end{cases}$$

[0052] 여기서, $n(S_{ps})$은 동일한 감정을 가지는 앞 문장의 수를 나타낸다.

[0053] 단계 S140에서는, 단계 S110 내지 단계 S130에 의해 산출된 키워드 정보, 문장의 위치에 대한 가중치 및 문장 간의 변화도를 이용하여 하기의 수학식 4에 의해 문장의 컨텍스트 정보를 산출한다.

수학식 4

[0054] $$C(S) = \alpha(S) \cdot \beta(S) \cdot \gamma(S)$$

[0055] 단계 S200에서는, 단계 S100에 의해 추출된 컨텍스트 정보를 이용하여 다중 문장으로부터 주요 문장을 추출한다. 즉, 단계 S200을 통해 텍스트 전체의 감정을 추정하는데 중요한 주요 문장을 추출하며, 추출되는 주요 문장은 복수 개로 추출될 수 있다.

[0056] 단계 S300에서는, 단계 S200에 의해 추출된 주요 문장 각각에 대하여 감정 특징을 추출하며, 단계 S300의 세부적인 흐름에 대하여는 도 3을 참조하여 상세히 설명하도록 한다.

[0057] 도 3은 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법의 단계 S300에 대한 세부 순서도이다. 도 3에 도시된 바와 같이, 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법의 단계 S300은, 형태소 분석기를 이용하여 문장에 포함된 단어를 품사별로 분류하는 단계(S310) 및 감정 사전을 이용하여 단어에 대한 감정 특징을 추출하는 단계(320)를 포함할 수 있다.

[0058] 단계 S310에서는, 형태소 분석기를 이용하여 문장에 포함된 단어를 품사별로 분류한다. 단어를 형태소 분석을 하게 되면, 다양한 활용을 하는 용언도 일치하는 어간으로부터 동일 단어 여부를 판단할 수 있고, 이러한 과정을 거쳐 단어를 품사별로 분류할 수 있다.

[0059] 단계 S320에서는, 감정 사전을 이용하여 단계 S310에 의해 분류된 단어에 대한 감정 특징을 추출한다. 보다 구체적으로, 감정 사전은 각 품사별로 감정에 따른 단어와 그 단어의 감정 세기를 포함할 수 있고, 단어에 대한 감정의 종류와 감정의 세기를 감정 특징으로 추출할 수 있다. 이때, 감정 사전은 명사, 동사, 형용사, 부사 및 이모티콘별 감정 특징을 포함할 수 있다.

[0060] 한편, 같은 어휘라 하더라도 특정 도메인에 따라 다른 감정을 나타내는 경우가 발생할 수 있다. 예컨대, “가볍다”라는 어휘는 “인물” 도메인에서는 부정적인 의미를 나타내는 반면, “통신” 도메인에서는 긍정적 의미를 나타낸다. 즉, 같은 어휘가 특정 도메인에 따라 감정이 달라질 수 있는바, 감정 사전은 어휘 사전에 기반한 기존의 형식적 감정 사전에 도메인 기반 감정 사전을 추가하여 구축하는 것이 바람직하고, 이를 통해 다중 문장에 대한 보다 정확한 감정 분류를 할 수 있다.

[0061] 더욱이, 단계 S320에서, 문장에 포함된 이모티콘은 문장의 감정을 분류하는데 매우 중요한 요소임에도 불구하고, 사용자의 취향이나 오타, 그리고 기타 여러 가지 요인으로 인해 같은 의미를 가짐에도 매우 불규칙하게 쓰여서 그 자체를 감정 특징으로 사용하는 것이 어려운 문제가 있다. 예컨대, “^_^”과 “^_____^”은 같은 의미이지만 개인에 따라 “_”의 개수를 다르게 사용할 수 있으며, 이러한 불규칙 이모티콘을 그대로 사용하는 것은 정확한 감정 분류를 어렵게 하는 요인 중 하나이다.

[0062] 이를 위해, 문장에 포함된 불규칙적인 이모티콘을 감정 사전에 포함된 참조 이모티콘으로 변환하여 이로부터 감정 특징을 추출하는 것이 바람직하다. 여기서 "참조 이모티콘"이란 감정 사전에 포함된 이모티콘으로서, 불규칙적으로 자주 사용되는 이모티콘 중 가장 간단한 이모티콘 형태를 말한다. 즉, 불규칙적으로 사용하는 이모티콘을 이러한 참조 이모티콘으로 변환함으로써 문장에 포함된 불규칙 이모티콘으로부터 정확한 감정 분류를 수행할 수 있다.

[0063] 보다 구체적으로, 문장에 포함된 불규칙한 이모티콘은 베이지안 프레임워크를 이용하여 참조 이모티콘으로 변환할 수 있으며, 도 4는 본 발명의 일실시예에 따른 컨텍스트 정보를 이용한 다중 문장으로부터의 감정 분류 방법에서, 베이지안 프레임워크를 이용하여 불규칙 이모티콘을 처리하는 세부 흐름을 도시한 도면이다. 도 4에 도시된 바와 같이, 이모티콘을 분해한 후 히스토그램을 이용한 정규화 과정을 거쳐 확률분포 모델을 구성하고, 불규칙 이모티콘과 참조 이모티콘 간의 유사도(likelihood)를 산출하여 최적의 참조 이모티콘을 추출함으로써 불규칙 이모티콘을 처리할 수 있다.

[0064] 단계 S400에서는, 단계 S300에 의해 추출된 감정 특징을 이용하여 감정 분류기를 통해 각각의 주요 문장의 감정을 분류한다. 즉, 단계 S300에 의해 추출된 감정 특징을 특징 벡터로 구성하여 감정 분류기를 통해 문장의 감정을 분류하며, 이때 문장 감정 분류를 위한 감정 분류기는 SVM(Support Vector Machine)을 이용할 수 있다.

[0065] 단계 S500에서는, 단계 S400에 의해 분류된 각각의 주요 문장의 감정을 결합하고, 이를 통해 다중 문장의 감정을 분류하여 최종적으로 텍스트의 전체 감정을 추정할 수 있다.

[0066] [실험예]

[0067] 문장의 감정 분류 성능 실험

[0068] 트위터, 페이스북, 미투데이와 같은 소셜 네트워크 서비스(SNS)에서 사용자가 작성한 글들을 일반, 제품리뷰, 여행, 음식 및 영화 도메인별로 무작위로 수집한 후, 수집된 텍스트를 각각 네 가지 방법을 사용하여 감정 분류를 수행하였다.

[0069] 즉, 기존의 형식적 사전만을 이용한 방법(case 1), 도메인 기반 감정 사전을 추가하여 구축한 감정 사전만을 이용한 방법(case 2), 컨텍스트 정보와 기존의 형식적 사전을 이용한 방법(case 3) 및 컨텍스트 정보와 도메인 기반 감정 사전을 추가하여 구축한 감정 사전을 이용한 방법(case 4)을 사용하여 감정 분류를 수행하였다. 수행된 각각의 방법에 따른 문장의 감정 분류 성능은 하기의 수학식 5 내지 수학식 7에 의한 정확률(precision, “p”) 및 재현율(recall, “r”)을 이용한 F_1-measure를 사용하여 평가하였고, 그 결과를 표 1 및 도 5에 나타내었다.

수학식 5

[0070]
$$\text{precision}(\rho) = \frac{\text{해당 감정으로 분류된 실제 해당 감정 텍스트 수}}{\text{해당 감정으로 분류된 텍스트 수}}$$

수학식 6

[0071]
$$\text{recall}(\gamma) = \frac{\text{해당 감정으로 분류된 실제 해당 감정 텍스트 수}}{\text{해당 감정 전체 텍스트 수}}$$

수학식 7

$$F_1 - \text{measure} = \frac{2\gamma\rho}{\gamma+\rho}$$

[0072]

표 1

[0073]

도메인	Case	감정	p	r	F1
일반	Case 1	긍정	0.5798	0.5644	0.5719
		부정	0.6377	0.4891	0.5536
		중립	0.6841	0.5991	0.6387
	Case 2	긍정	0.6213	0.5891	0.6047
		부정	0.6124	0.6401	0.6259
		중립	0.7135	0.6787	0.6956
제품리뷰	Case 1	긍정	0.6012	0.8181	0.6930
		부정	0.6663	0.2513	0.3649
		중립	0.5387	0.6578	0.5923
	Case 2	긍정	0.6648	0.7273	0.6946
		부정	0.6259	0.6211	0.6234
		중립	0.9121	0.6806	0.7795
	Case 3	긍정	0.8122	0.7301	0.7689
		부정	0.6381	0.6114	0.6244
		중립	0.7533	0.8101	0.7806
	Case 4	긍정	0.8129	0.7013	0.7529
		부정	0.6587	0.7759	0.7125
		중립	0.8264	0.8585	0.8421
여행	Case 1	긍정	0.7512	0.7598	0.7554
		부정	0.6602	0.3289	0.4390
		중립	0.4451	0.6654	0.5333
	Case 2	긍정	0.7146	0.8336	0.7695
		부정	0.6657	0.3328	0.4437
		중립	0.4281	0.5045	0.4631
	Case 3	긍정	0.7498	0.7592	0.7544
		부정	0.5722	0.6687	0.6166
		중립	0.8007	0.6618	0.7246
	Case 4	긍정	0.6599	0.8304	0.7353
		부정	0.5431	0.5007	0.5210
		중립	0.8704	0.5011	0.6360
음식	Case 1	긍정	0.7141	0.8401	0.7719
		부정	0.506	0.1916	0.2779
		중립	0.3754	0.4894	0.4248
	Case 2	긍정	0.7271	0.8891	0.7999
		부정	0.2035	0.2789	0.2353
		중립	0.902	0.2531	0.3952
	Case 3	긍정	0.8334	0.8136	0.8233
		부정	0.5014	0.9042	0.6450
		중립	0.8576	0.7234	0.7848
	Case 4	긍정	0.7891	0.8341	0.8109
		부정	0.5301	0.9012	0.6675
		중립	0.8249	0.6402	0.7209

영화	Case 1	긍정	0.6304	0.6681	0.6487
		부정	0.7813	0.4285	0.5534
		중립	0.2111	0.5131	0.2991
	Case 2	긍정	0.6936	0.4462	0.5430
		부정	0.7288	0.5898	0.6519
		중립	0.3312	0.7366	0.4569
	Case 3	긍정	0.5813	0.7777	0.6653
		부정	0.6054	0.4284	0.5017
		중립	0.8652	0.4809	0.6181
	Case 4	긍정	0.6148	0.8876	0.7264
		부정	0.9384	0.6278	0.7523
		중립	0.5812	0.6857	0.6291

[0074] 표 1 및 도 5에 나타낸 바와 같이, 감정 사전만을 이용한 경우에 비해 컨텍스트 정보와 감정 사전을 이용한 경우 감정 분류 성능 효과가 우수함을 확인하였다. 또한, 일반 감정 사전을 사용한 것에 비해 도메인 기반 감정 사전을 추가하여 구축한 감정 사전의 경우에 감정 분류 성능 효과가 보다 우수함을 확인하였다. 따라서 본 발명에 따른 방법은 텍스트를 구성하는 다중문장으로부터의 감정 분류 성능 효과가 우수함을 알 수 있다.

[0075] 이상 설명한 본 발명은 본 발명이 속한 기술분야에서 통상의 지식을 가진 자에 의하여 다양한 변형이나 응용이 가능하며, 본 발명에 따른 기술적 사상의 범위는 아래의 특허청구범위에 의하여 정해져야 할 것이다.

부호의 설명

[0076] S100: 텍스트를 구성하는 다중 문장 각각에 대하여 컨텍스트 정보를 추출하는 단계

S200: 추출된 컨텍스트 정보를 이용하여 다중 문장으로부터 복수 개의 주요 문장을 추출하는 단계

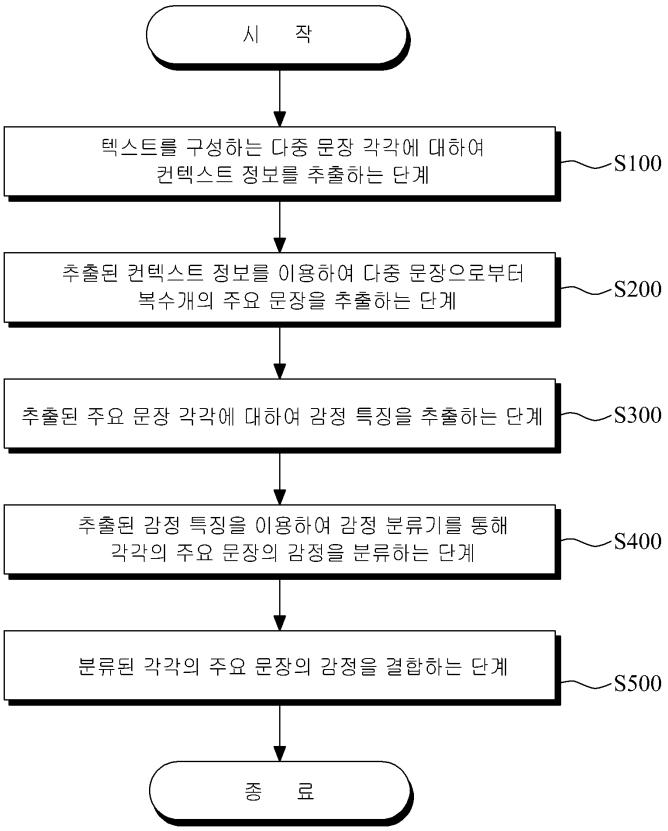
S300: 추출된 주요 문장 각각에 대하여 감정 특징을 추출하는 단계

S400: 추출된 감정 특징을 이용하여 감정 분류기를 통해 각각의 주요 문장의 감정을 분류하는 단계

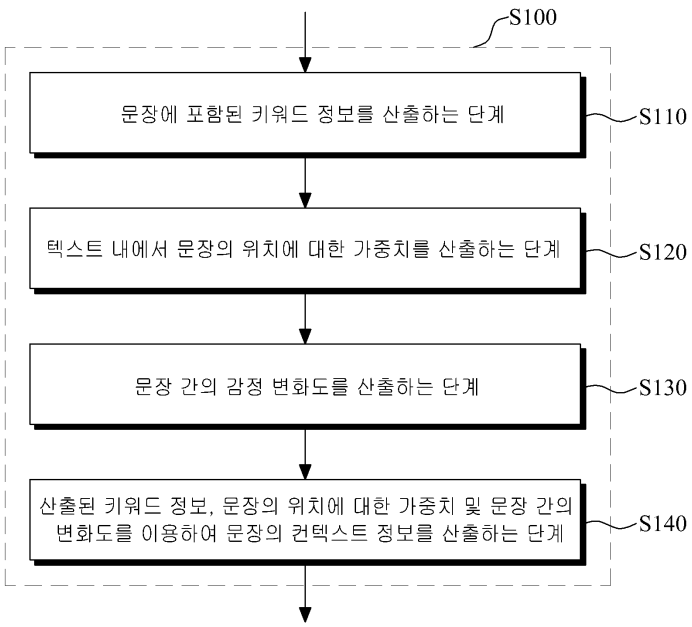
S500: 분류된 각각의 주요 문장의 감정을 결합하는 단계

도면

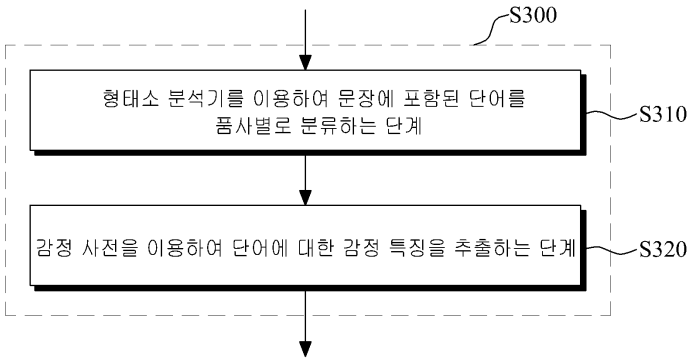
도면1


시 작
텍스트를 구성하는 다중 문장 각각에 대하여 컨텍스트 정보를 추출하는 단계
S100
추출된 컨텍스트 정보를 이용하여 다중 문장으로부터 복수개의 주요 문장을 추출하는 단계
S200
추출된 주요 문장 각각에 대하여 감정 특징을 추출하는 단계
S300
추출된 감정 특징을 이용하여 감정 분류기를 통해 각각의 주요 문장의 감정을 분류하는 단계
S400
분류된 각각의 주요 문장의 감정을 결합하는 단계
S500
종 료

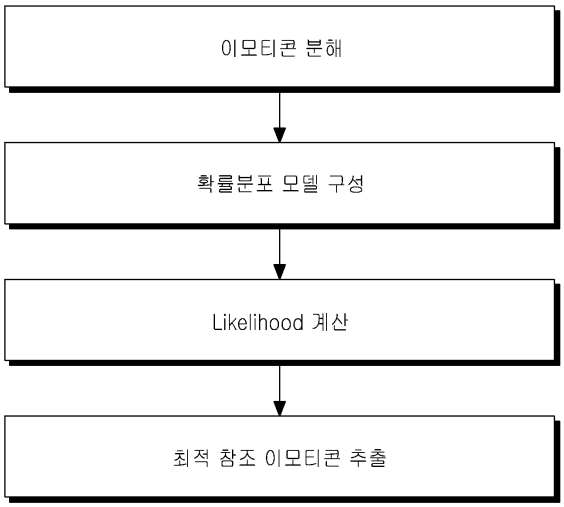
도면2


S100
문장에 포함된 키워드 정보를 산출하는 단계
S110
텍스트 내에서 문장의 위치에 대한 가중치를 산출하는 단계
S120
문장 간의 감정 변화도를 산출하는 단계
S130
산출된 키워드 정보, 문장의 위치에 대한 가중치 및 문장 간의 변화도를 이용하여 문장의 컨텍스트 정보를 산출하는 단계
S140

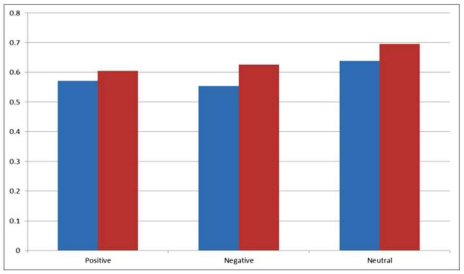
도면3


S300
형태소 분석기를 이용하여 문장에 포함된 단어를 품사별로 분류하는 단계
S310
감정 사전을 이용하여 단어에 대한 감정 특징을 추출하는 단계
S320

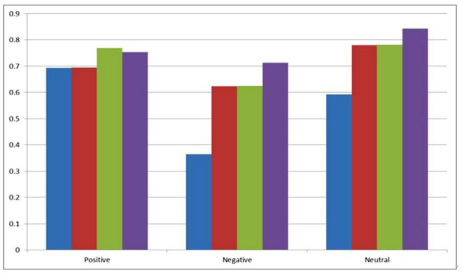
도면4


이모티콘 분해
확률분포 모델 구성
Likelihood 계산
최적 참조 이모티콘 추출

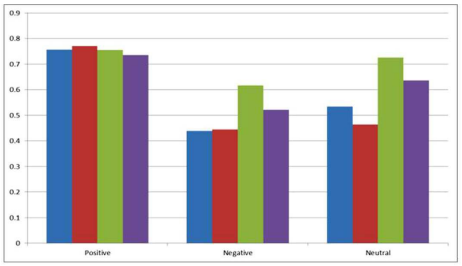
도면5



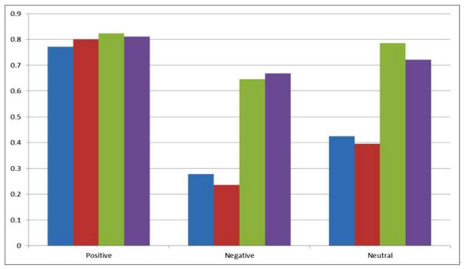
(a) 일반



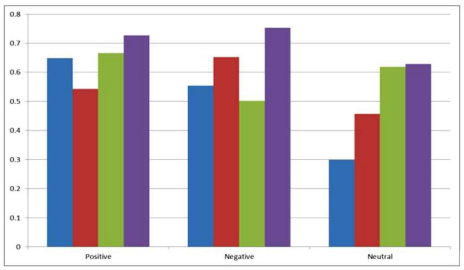
(b) 제품 리뷰



(c) 여행



(d) 음식



(e) 영화