



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2011-0044112
(43) 공개일자 2011년04월28일

(51) Int. Cl.

G06F 17/40 (2006.01) G06F 17/30 (2006.01)

(21) 출원번호 10-2009-0100948

(22) 출원일자 2009년10월22일

심사청구일자 2009년10월22일

(71) 출원인

건국대학교 산학협력단

서울 광진구 화양동 1 건국대학교내

세종대학교산학협력단

서울 광진구 군자동 98 세종대학교

(72) 발명자

강한훈

경기도 안산시 상록구 성포동 587 주공아파트 30
7동 101호

유성준

서울특별시 강동구 명일1동 삼익그린아파트 506동
1503호

(74) 대리인

특허법인 엠에이피에스

전체 청구항 수 : 총 2 항

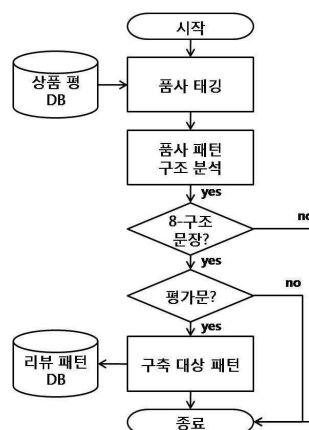
(54) 상품 속성별 리뷰의 마이닝을 위한 패턴 데이터베이스의 반자동적인 구축 방법

(57) 요약

본 발명은 상품평으로부터 상품의 속성별 긍정/부정 분류를 위해 상품평 패턴 사전을 이용하려고 했을 때, 상품평 패턴 사전을 반자동적으로 구축할 수 있는 방법에 관한 것이다.

상품평 패턴 사전은 상품의 속성별로 사용되는 어휘에 대한 긍정/부정 여부를 사전에 구축해 놓은 뒤에 이를 통해 새로운 문장에 대해서 긍정/부정으로 분류를 수행한다. 이러한 상품평 패턴 사전의 구축은 과거에 수작업으로 진행했지만 본 발명에서는 이를 반자동적인 방법으로 구축하는 방법에 관한 것으로 상품 평 문장에 대해서 품사 태깅을 수행한다. 이후에 태깅된 단어의 품사가 명사인 것 중에 상품의 속성 명으로 쓰이는 단어를 찾는다. 해당 단어를 기준으로 3어절의 문장을 추출한다. 추출된 문장이 8-구조 문장에 해당되는지 비교한 후에 8-구조 문장에 해당된다면 문장 내에 특정한 속성을 대상으로 평가가 이루어지고 있는지 찾는다. 만약 특정한 속성에 대해 평가가 이루어지고 있다면, 해당 문장을 리뷰 패턴 데이터베이스에 삽입하는 시스템이다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 10581

부처명 서울특별시

연구관리전문기관

연구사업명 서울시 산학연 협력사업 (2005년 기술기반구축사업)

연구과제명 u-라이프를 위한 디지털 콘텐츠 산업화 혁신 클러스터 구축

기여율

주관기관 건국대학교 산학협력단

연구기간 2005년 12월 01일 ~ 2010년 11월 30일

특허청구의 범위

청구항 1

상품의 속성별로 긍정 또는 부정을 분류함에 있어서, 리뷰 패턴 데이터베이스를 이용한다고 했을 때, 리뷰 패턴 데이터베이스를 구축하는 시간을 줄이기 위해 상품평에서 가장 많이 쓰이는 8-구조 문장 패턴의 분석 및 정의.

청구항 2

온라인 쇼핑몰에 게재된 상품평으로부터 품사 태깅(2)을 수행하고, 상품의 속성 명을 기준으로 좌측으로 3어절, 우측으로 3어절을 추출(3)한 후에 추출한 문장이 8-구조 문장(4)에 해당하는지 비교를 수행하고 8-구조 문장에 해당된다면 문장 내에서 평가문이 포함되어 있는지 분석(5)하고 평가문이 포함되어 있다고 했을 해당 패턴을 리뷰 패턴 데이터베이스에 추가(6)하는 시스템.

명세서

발명의 상세한 설명

기술 분야

[0001] 오피니언 마이닝

배경 기술

[0002] 본 발명은 오피니언 마이닝 연구에 해당한다. 기존에 이루어진 오피니언 마이닝 연구 중에는 첫째, 기계학습 기법을 적용한 전통적인 주제 기반 문서 분류 방법을 적용한 연구가 있다. 이 연구에서는 IMDb(Internet Movie Database)로부터 영화에 대한 700개의 positive review, 700개의 negative review 문서를 수집하여 Support Vector Machine, Naive Bayes의 분류 알고리즘을 사용하였다. 특징 단어는 unigrams, bigrams 등의 다양한 방법으로 선택했으며, 단어에 대한 가중치는 term frequency, presence 기법을 적용하였다. 실험 결과 특징 단어를 unigrams으로 추출하고, 단어 가중치는 presence 기법을 사용하였을 때, SVM 분류 알고리즘에서 82.9%로 가장 우수한 분류 성능을 보였다. 둘째, Score Function이라는 분류 알고리즘을 정의하고 이를 사용하였다. 실험에서는 7종류의 제품에 대한 리뷰 문서를 가지고 진행하였다. 리뷰 문서의 수는 대략 13000개 이상을 대상으로 하였다. 특징 단어에 대한 선택은 bigrams, trigrams 형태로 추출하여 실험한 결과 84.6%~88.3%의 성능 보였으며, naive Bayesian, SVM 그리고 다양한 종류의 분류 알고리즘을 통해서 실험한 결과를 보이고 있다. 셋째, 감정 단어 구(phrase)의 패턴을 기반으로 4종류의 서로 다른 주제에 대해서 분류를 수행한 연구도 있다. 이 연구에서는 170개의 positive review와 240개의 negative review를 Epinions[13]로부터 수집하였다. 수집한 문서로부터 특정 패턴에 맞는 품사를 추출하고 PMI(Pointwise Mutual Information)식을 변형하여 적용하였다. 변형된 PMI식에서는 altavista 검색엔진의 'near' 연산자를 사용하여 한 문서에 포함된 특정 패턴에 맞는 여러 개의 구(phrase)의 semantic orientation(SO)을 결정하도록 한다. 여러 개의 구(phrase)에 대한 SO의 합이 하나의 문서에 대한 semantic orientation이 된다. 실험 결과 'Automobiles' 주제와 'Banks' 주제에서 80%이상의 분류 정확도를 보였다. 넷째, supervised machine learning 기법으로 여행지에 대한 사용자의 리뷰의 긍정/부정 분류실험을 수행한 연구도 있다. 이 연구에서는 yahoo.com에서의 7개의 여행지에 대한 사용자의 리뷰를 수집했고, SVM, Naive Bayes, dynamic language model classifier의 분류 알고리즘을 사용하였다. SVM과 Naive Bayes에서 feature selection으로 사용한 가중치 알고리즘은 Information Gain을 사용하였다. 이 연구에서는 training data의 수에 따라 분류 결과를 보였고 긍정 데이터 셋과 부정 데이터 셋이 각각 300개 이상일 때 SVM에서 가장 높은 분류 성능을 보였다. 마지막으로 한국어 문서 중에서 상품리뷰, 영화리뷰, 신문기사를 데이터 셋으로 수집하여 이에 대해서 전통적인 문서 분류 기법으로 감정(긍정, 부정) 분류 실험결과를 보여준 연구도 있다. 감정 자질은 영어권의 워드넷을 확장하여 적용하였다. 분류 알고리즘으로는 SVM을 사용하였고, feature selection 방법은 TF-IDF, TF-ISF를 사용하였다.

발명의 내용

해결 하고자하는 과제

[0003] 본 발명은 오피니언 마이닝 연구의 한 분야로써, 기존에 제안된 연구의 대부분은 문장 전체를 대상으로 긍정 또는 부정인지를 분류하였다. 실제로 쇼핑몰 등에 게재되는 상품평에서는 상품의 속성별로 사용자의 감정이 나타나는데, 기존에 제안된 연구로는 상품의 속성별로 긍정 또는 부정을 분류하기 어렵다. 따라서 상품의 속성별 사용되는 어휘의 긍정/부정 여부를 패턴 사전에 정의 해놓고, 상품의 속성별로 새로운 상품평 문장에 대해서 비교하여 그 결과를 이용하는 것이 분류 정확성을 높일 수 있다. 그러나 패턴 사전을 구축하기 위해서 수작업 분석을 해야 하며 많은 시간이 소비될 수 있다는 문제를 가지고 있다. 본 발명에서는 리뷰 패턴 사전을 구축함에 있어서 수작업 분석 시간을 줄여보고자, 상품평에서 가장 많이 쓰이는 패턴을 분석하였고, 이를 찾았다. 본 발명에서는 8-구조 문장이라는 상품평으로서 가장 많이 쓰이는 패턴을 정의하고 이를 활용하여 리뷰 패턴 데이터베이스를 구축하는 시스템에 대해서 소개하였다.

과제 해결수단

[0004] 온라인 쇼핑몰에 게재된 실제 상품평으로 부터 그 내용을 분석하고 가장 많이 쓰이는 상품평의 활용 패턴을 분석하고 이를 구축한다. 본 발명에서는 이를 8-구조 문장이라고 정의하였고, 이러한 8-구조 문장을 활용하여 구축한 리뷰 패턴 데이터베이스를 활용하여 상품의 속성별 긍정/부정 분류 시스템을 소개한다.

효 과

[0005] 상품평에서 속성별로 긍정/부정을 분류하는 데 있어서 리뷰 패턴 데이터베이스를 이용한다고 했을 때, 기존에는 수작업을 통해서 분석하여 많은 시간을 요구로 했다. 그러나 본 발명에서는 분석 시간을 줄여보고자 상품평으로부터 가장 많이 활용되는 패턴을 분석하여 이를 찾아내었고, 이를 활용함으로써 리뷰 패턴 데이터 베이스의 구축 시간을 단축시킬 수 있다.

발명의 실시를 위한 구체적인 내용

[0006] 이하 첨부된 도면에 의해 상세히 설명하면 다음과 같다.

[0007] 도 1은 본 발명의 주요 구성도로써 온라인 쇼핑몰에서 수집된 상품평으로부터 품사 태깅을 통해 품사를 추출하고, 추출된 문장에 대한 품사 패턴의 구조를 분석한다. 만약 추출된 문장이 8-구조 문장을 만족한다면, 해당 문장에서 평가문장이 있는를 판단한다. 해당 문장 내에 평가 문장이 포함되어 있다면 이를 리뷰 패턴 데이터베이스에 삽입한다.

[0008] 도 2는 수작업을 통해 리뷰 패턴 데이터베이스를 구축하는 흐름도이며, 기존 방법에서 ‘제품 속성에 따른 문장 분석’ 과정은 단순히 수작업을 통해서 분석하는 과정이지만 본 발명에서의 ‘제품 속성에 따른 문장 분석’은 도1에 해당하는 과정이다. 이 과정 이후에 이루어지는 과정은 평가 문장에 사용된 어휘가 긍정의 의미를 갖는지 부정의 의미를 갖는지 분석하는 과정이다. 또한 이 과정에서는 강도 부사 등이 쓰였을 경우 강한 긍정 또는 강한 부정인지도 분석하여 이를 리뷰 패턴 데이터베이스에 삽입하는 과정이다.

[0009] 도 3은 실제로 구축된 리뷰 패턴 데이터베이스의 예를 보여준다. 리뷰 패턴 데이터베이스에 구성되는 항목으로는 속성명, 품사가 태깅된 상품평의 패턴, 그리고 해당 패턴의 등급화 정보이다. 이러한 패턴 사전을 통해 새로운 문장과 비교하여 그 결과를 긍정 또는 부정인지로 보여주게 된다.

[0010] 도 4는 8-구조 문장의 품사 패턴과 예문 그리고 품사과 태깅된 문장의 형태를 보여준다. 이러한 8-구조 문장의 정의는 가장 많이 활용되는 상품평의 패턴을 정의하는 것으로써 리뷰 패턴 데이터베이스를 구축함에 있어서 시간적으로 많이 단축시켜 줄 수 있음을 의미한다.

도면의 간단한 설명

[0011] 도 1은 본 발명이 적용되는 리뷰 패턴을 반자동적으로 구축하는 시스템 구성도.

[0012] 도 2는 일반적인 수작업 방법으로 리뷰 패턴을 구성하는 흐름도.

[0013] 도 3은 리뷰 패턴 데이터 베이스의 데이터 구조도.

[0014] 도 4는 8-구조 문장의 패턴 예시도.

[0015] <도면의 주요부분에 대한 부호의 설명>

[0016] 1: 상품의 속성별로 사용자의 선호도가 문장 형태로 이루어진 상품평

[0017] 2: 문장의 어절 또는 단어를 인식하여 해당 단어의 품사를 결정함

[0018] 3: 상품의 속성별로 상품평의 패턴 형태 분석 및 추출함.

[0019] 4: 추출된 상품평 패턴이 8-구조 문장에 해당되는지 분석함.

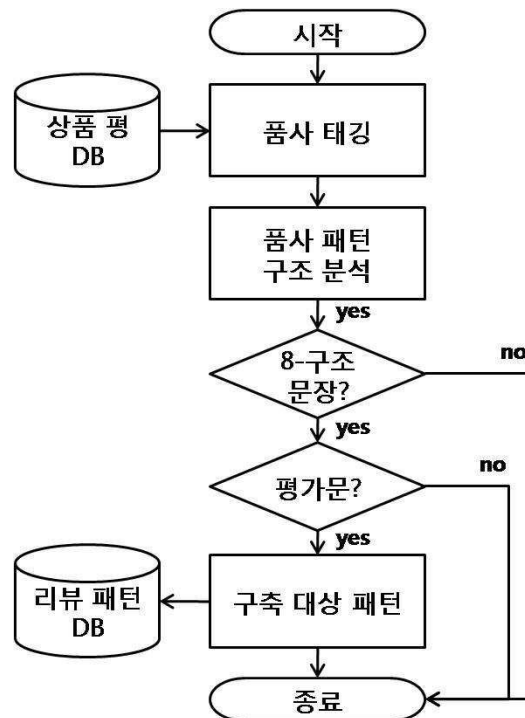
[0020] 5: 상품평 패턴 중에는 상품의 속성명과 긍정 또는 부정을 나타내는 술어 문장이 포함되어 있는데, 술어 문장이 평가문인지 비교함.

[0021] 6: 4와 5의 과정을 거친 후에 해당 문장을 리뷰 패턴으로 보고, 이를 패턴 사전에 삽입함.

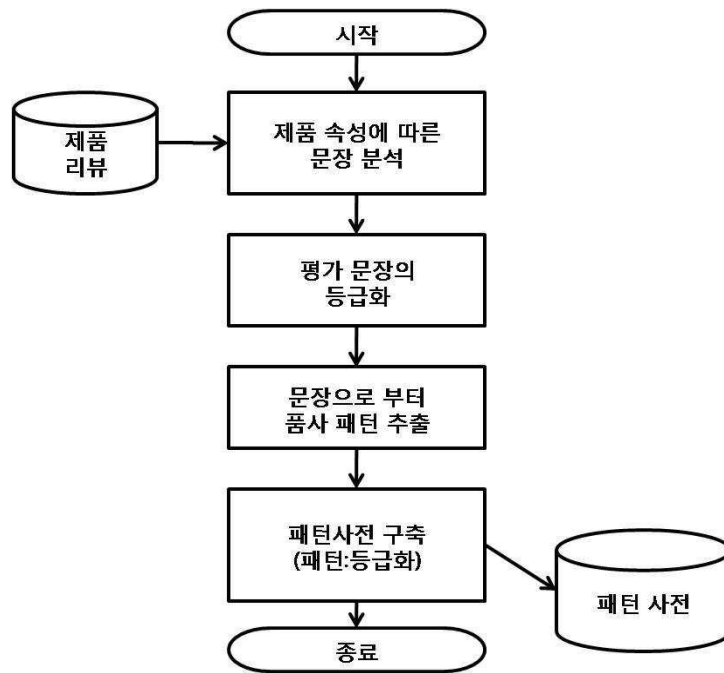
[0022] 7: 6의 과정을 통해 생성된 리뷰 패턴이 저장되는 있는 데이터베이스

도면

도면1



도면2



도면3

속성명	품사 패턴	등급
배송	배송(N)_느리(V)	부정
배송	배송(N)_빠르(V)	긍정
배송	배송(N)_아주(Z)_늦(V)	강한 부정
화면	화면(N)_선명(V)	긍정
화질	선명(N)_화질(N)	긍정
...		

도면4

	품사 패턴	예문	품사 태깅
구조 1	N[명사]+V[술어]	배송이 빠르다	배송(N)+빠르(V)
구조 2-1	N[명사]+Z[부사]+V[술어]	화면이 정말 크다	화면(N)+정말(Z)+크(V)
구조 2-2	N[명사]+Z[부사]+N[명사]	가격에 너무 만족한다	가격(N)+너무(Z)+만족(N)
구조 3	N[명사]+N[명사]+N[명사]	가격에 완전 만족한다	가격(N)+완전(N)+만족(N)
구조 4-1	V[술어]+N[명사]	싼 가격	싸(V)+가격(N)
구조 4-2	N[명사]+N[명사]	선명한 화질	선명(N)+화질(N)
구조 5-1	Z[부사]+V[술어]+N[명사]	너무 빠른 배송	너무(Z)+빠르(V)+배송(N)
구조 5-2	Z[부사]+N[명사]+N[명사]	정말 저렴한 가격	정말(Z)+저렴(N)+가격(N)