상품의 기능 계층 구성도를 이용한 웹상에서 추출한 상품 상세 정보 처리

이근용, 박기선, 이용석
전북대학교 컴퓨터정보과학과
cpher2000@gmail.com, icarus@chonbuk.ac.kr, yslee@chonbuk.ac.kr

Processing Detailed Description of Product Extracted from Web Using The Functional Hierarchy of Goods

Keun-Yong Lee, Ki-Seon Park, Yong-Seok Lee
Dept. of Computer Information Science, Chonbuk National University

요 약

인터넷 쇼핑몰을 이용하는 구매자의 상품 구매를 돕는 방법으로 가격 비교 서비스가 가장 많이 이용되고 있다. 가격비교 서비스는 구매자가 구매할 상품을 이미 결정했다고 가정하고 동일 상품을 판매하는 사이트들의 가격과 서비스 정보를 비교하여 구매자의 구매 결정에 많은 도움을 주고 있다. 가격 비교 서비스는 구매자에게 인터넷 쇼핑몰 사이트에서 판매되는 수많은 상품 중 어떤 상품을 선택할지에 대한 고민을 해결해 주지는 못하고 있다. 구매자가 구매할 상품을 결정하지 못했을 때, 상품의 선택을 도울 수 있는 방법은 서로 다른 상품 모델에 대한 기능적 차이를 비교할 수 있도록 해주어야 한다. 상품에 대한 기능의 차이점은 상품의 상세 정보를 통해서 파악이 가능하다. 따라서 상품의 상세 정보를 구매자가 한눈에 파악할 수 있도록 제공하는 것은 상품을 선택하는데 매우 중요한 요소이다. 각 상품의 상세 정보는 구매자에게 해당 상품이 어떤 기능을 가지고 있는지 보기 쉽게 기술되어 있지만 다른 상품과는 기능을 설명하는 순서가 다르거나 사용한 용어 및 단위 표현에 같은 의미의 다른 표현 방식을 사용하기도 한다. 본 논문은 상품들의 기능적인 차이점을 파악하는 것을 도울 수 있도록 하기 위해서, 개별 상품에 대해서는 상품의 상세 정보가 보기 쉽게 기술되었다는 점을 이용하여 상품의 상세 정보로부터 상품의 정보를 추출한다. 추출된 정보는 상품을 구성하는 기능 계층 정보를 이용하여 각 상품들의 기능과 기능에 대한 설명을 일치시키는 방법을 제안한다.

1. 서론

인터넷 사용자가 2005년 현재 3300만 명을 넘어서고 인터넷 쇼핑몰에 대한 규모도 10년 만에 10조원 시장으로 성장하였다. 인터넷 쇼핑몰을 사용하는 사용자의 연령층 또한 10대에서 40대로 다양하며, 쇼핑몰 사용자의 규모 또한 1700만 명에 달하고 있다.[1]

인터넷 쇼핑몰은 상품을 판매하고자 하는 상품 판매자와 상품을 구매하고자 하는 구매자 간의 온라인을 통한 시장이라 볼 수 있다. 상품 판매자는 인터넷 쇼핑몰을 통하여 상품의 정보를 제공하여 상품이 선택받기를 원하고 있으며, 구매자는 같은 상품 중에 최저의 가격으로 상품을 구매하기를 원하고 있다. 구매자가 인터넷 쇼핑몰을 이용할 때 구매 패턴은 다음과 같다.

1단계. 구매를 원하는 상품을 결정
2단계. 가격 비교 사이트를 이용하여 가격 비교
3단계. 타 쇼핑몰과 비교[1)]
4단계. 구매 결정

특히, 최근 엄지족으로 대별되는 20대 초중반의 젊은 세대들은 소비생활에 있어서 가격과 성능을 비교해보고 구매하는 비교구매, 기존 구매자의 사용 후기, 상품평등을 참고하여 구매하는 리뷰구매, 휴대전화 및 인터넷을 통해 코드형태로 다운로드 받아 사용하는 온라인 쿠폰 등의 "지능형 쇼핑" 이 일반화되는 추세를 보이고 있다.[2]

기존의 가격 비교 사이트들은 사용자가 구매할 상품을 미리 결정하였다는 가정 하에[3] 구매 패턴에서의 2단계(가격 비교)와 3단계(타 쇼핑몰과의 비교)에서 사용자의 상품 구매를 돕고 있다.

인터넷 쇼핑몰에서 제시되는 상품들은 동종의 카테고리 내에 적게는 수십 점에서 많게는 수백 점의 상품정보를 제공한다. 인터넷 쇼핑몰에서 제공하는 수많은 상

1) 배송의 형태 또는 결재 방법과 같은 비교를 의미하며 상품의 기능을 의미하는 것은 아님

품 중에서 구매자가 구매하고자 하는 상품을 결정할 때 참고할 수 있는 정보는 인터넷 쇼핑몰의 추천정보, 판매량과 각 상품들의 상세 정보가 있다. 그러나 구매자 자신이 진짜 원하는 기능을 포함하는 상품이 무엇인지 구별할 수 있도록 도와주는 방법은 거의 없기 때문에 구매자는 구매하고자 하는 상품을 선택하기가 쉽지 않다.

인터넷 쇼핑몰에서 상품 선택을 어렵게 하는 이유는, 인터넷 쇼핑몰에서 구매자에게 제시해주는 상품들 사이의 차이점이 대부분 "모델명", "제조사", "가격", "나열식 상세 정보"가 주를 이루고 있기 때문이다. 더구나 상품의 상세한 정보는 한 번에 한 상품만의 상세 정보를 볼 수 있어서 상품들 사이의 차이점을 비교하기 어렵게 되어 있기 때문이다. 따라서 인터넷 쇼핑몰에서 제공하는 상품의 상세 정보를 이용하여 상품들 간의 기능적인 차이점을 비교할 수 있다면 구매자의 상품 선택을 도울 수 있을 것이다.

본 논문은 인터넷 쇼핑몰에서 제공되는 각 상품의 상세 정보가 해당 상품에 대해서는 구매자가 상품에 대한 기능을 파악하기 쉽게 구성되어 있다는 점을 이용하였다. 인터넷 쇼핑몰에서 상품의 기능에 대한 상세 정보는 HTML의 TABLE을 이용하거나 특정 분리자를 이용하여 상품의 기능과 해당 기능에 대한 설명을 제공하고 있다. 이러한 구조적 특성을 이용하여 기능을 표현하는 용어와 용어에 대한 설명을 추출한다. 상품의 계층 정보 사전을 이용하여 추출된 상품의 상세 정보를 다른 상품들과 비교 할 수 있도록 일관된 용어로 변환하는 방법을 제안한다.

2. 관련 연구

인터넷 쇼핑몰에서의 상품에 관련된 연구는 구매를 도와주는 에이전트 측면에서 연구가 활발히 진행되고 있지만 많은 연구들이 구매 패턴의 2단계인 가격 비교에 초점이 맞추어져 있었다.

근래에는 일부 인터넷 쇼핑몰에서 구매자가 선택한 상품의 상세 정보를 비교하여 서비스가 제공되고 있지만 아직은 미미한 상태이다. 그림 1과 그림 2는 서로 다른 두개의 모델을 선택하여 ENURI[2]와 DANAWA[3]에서 상품을 선택한 후 비교한 화면이다. 상세 정보 비교 서비스를 제공하는 쇼핑몰 중 ENURI는 선택한 상품들의 상세 정보를 단지 한 화면에서 볼 수 있도록 하고 있다. DANAWA의 경우는 좀 더 구체적인 비교를 하고는 있지만, 비교 내용 중에는 선택한 상품들 모두 기능을 가지고 있지 않은 정보와 함께 상품을 구성하는 기본 정보만을 제시하고 있다. DANAWA의 경우도 세부적인 상세 정보는 개별적으로 직접 확인하도록 하고 있다.

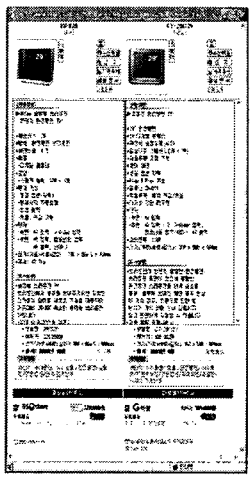

그림 1 ENURI의 상품 비교

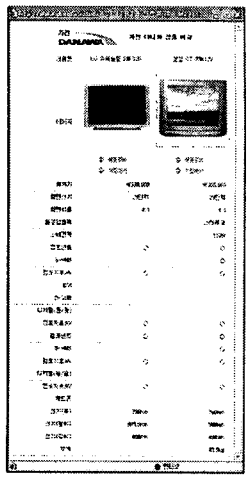

그림 2 DANAWA의 상품 비교

비교 구매 에이전트의 초기 모델은 앤더슨 컨설팅의 Bargain Finder로서, 구매자가 특정 상품을 찾고자 할 때 에이전트가 구매자를 대신하여 인터넷 상의 쇼핑몰에서 해당 상품을 비교하는 시스템[4]이 있다.

국내에서는 다수의 비교 구매 사이트가 서비스를 제공하고 있으며, [5]는 이런 비교 구매 기능에 대한 쇼핑 에이전트의 원형에 대해서 제안하였다.

[6]은 다양한 정보 출처에 존재하는 데이터 모델에 대해서 웹 정보를 추출하고 통합하기 위한 래퍼 시스템에 대해 제안하였다.

[7]은 비교 구매를 위해서 웹 온톨로지를 구축하는 방안에 대해서 연구하였다. [7]에서는 비교 구매를 위한 쇼핑몰 마다 다른 표현을 사용하는 가격과 서비스 관련 용어를 온톨로지를 이용하여 변환 처리하는 방법을 제안했다.

[3]은 의류와 같이 시각적인 비교가 필요한 부분에 대해서 비교하여 구매할 수 있도록 하는 이미지 검색을 이용한 비교 쇼핑 시스템을 제안하였다.

기존의 연구들은 웹에서의 정보를 추출하는 방법과 추출된 정보를 이용하여 동일한 상품에 대한 다른 인터넷 쇼핑몰 사이의 가격 비교에 중심이 맞추어져 있다. [3]은 상품들 사이의 비교가 이루어졌지만 상세 정보가 아닌 상품들의 이미지를 보고 선택할 수 있도록 하였다.

3. 상품 상세 정보의 추출 및 특징

2) ENURI: http://www.enuri.com
3) DANAWA: http://www.danawa.com,

3.1 상품 상세 정보의 추출

많은 인터넷 쇼핑몰들이 자사가 보유한 정보에 대해서 폐쇄적인 형태로 운영하고 있기 때문에 구조화된 일관된 정보를 추출하는 작업은 매우 어려운 실정이다. 쇼핑몰 사이트는 자사가 보유한 상품의 상세 정보를 복사하지 못하게 하는 장치를 한 경우도 많다. 인터넷 쇼핑몰에서 상세 정보의 추출은 각 쇼핑몰의 특징에 맞게 이루어져야 하며 확장성을 기대하기가 어렵다[8]. 인터넷 쇼핑몰로부터 상세 정보를 추출하기 위해서 래퍼(Wrapper)[6][8]를 구성하고 구성된 래퍼를 이용하여 정보를 추출 할 수 있도록 한다.

추출한 상품의 상세 정보는 InnerText라 불리는 웹브라우저에 보이는 텍스트와 추출한 정보의 구성 형식을 분석하기 위해서 HTML 태그가 부착된 형태인 OuterHtml 두 가지 모두를 추출하여 이용한다.

3.2 상품 상세 정보의 특징

표 1은 한 인터넷 쇼핑몰로부터 추출한 비슷한 가격대의 서로 다른 두 TV 상품에 대한 상세 정보이다. 두 상품의 상세 정보는 래퍼에 의해서 추출된 TV에 대한 상세 정보 부분으로서 어떤 특정 형식 없이 단순히 나열되어 있으며 기능 설명에 대한 순서나 표현 방법에 일관성이 없다. 예를 들어, 모델A와 모델B의 화면 크기를 나타내는 방법에 있어서도 모델A는 "화면크기: 29""와 "방식: 완전 평면 브라운관"으로 화면 크기와 방식을 분리하여 표현하고 있는 반면 모델B는 "29" 완전평면"과 같이 하나로 묶어서 표현하고 있다.

두 상품에 대한 상세 정보가 표 2와 같은 형식으로 일관된 표현 형식을 이용하여 표현될 수 있다면 두 모델 사이의 차이점을 쉽게 파악하여 상품 선택에 활용할 수 있을 것이다.

모델 A	모델 B
*화면크기: 29" *방식: 완전평면 브라운관 *화면비율: 4:3 *화질: -디지털 콤필터 *음향: -스피커 출력: 10W + 10W *편의 기능: -영문 캡션(자막) -현재시각 자동설정 -취침 예약 -꺼짐, 켜짐 기능 *단자: -측면: AV 입력, S-Video 입력 -후면: AV 입력, 컴포넌트 입력, AV 출력, 안테나 *크기(가로x세로x깊이): 796 x 600.5 x 406mm *무게: 42.7kg	*29" 완전평면 *1H 디지털 콤필터 *자연색 보정회로(AKB) *음성다중 스테레오(7W + 7W) *자동음량 조절 기능 *확대 화면 *영문 캡션 자막 *Plug & Play 기능 *멜로디 On/Off *취침예약, 예약 켜짐/꺼짐 *다기능 간단 리모콘 *단자: -측면: AV 입력 -후면: AV 입력 x 2, S-Video 입력, 컴포넌트 입력(480i), AV 출력 *소비전력: 110W *크기(가로x세로x깊이): 760 x 588 x 194mm

표 1 TV의 두 모델에 대한 제품 상세 정보

기능 \ 모델			Model A	Model B
화면	크기		29"	29"
	방식		완전평면 브라운관	완전평면
	비율		4:3	
	화질	엔진	디지털 콤필터	1H 디지털 콤필터
				자연색 보정회로(AKB)
음향	스피커	출력	10W+10W	7W+7W
		형식		음성다중 스테레오
	단자	측면	AV 입력 S-Video 입력	AV입력
		후면	AV 입력 컴포넌트 입력 AV 출력, 안테나	AV입력x2 S-Video 입력 컴포넌트 입력(480i) AV 출력
외관	크기	가로	796	
		세로	600.5	
		깊이	406	
	소비전력			110W
	무게		42.7kg	
편의 기능	캡션		영문 캡션(자막)	영문 캡션
	자동음량조절			O
	리모콘			다기능 간단 리모콘
	확대 화면			O
	현재시각		자동 설정	
	예약	취침	O	O
		꺼짐	O	O
		켜짐	O	O
	Plug & Play			O

표 2 두 상품의 용어와 설명 매핑 결과

서로 다른 모델에 대한 상품들의 가격 차이가 어떤 기능적인 차이에서 비롯한 것인지를 상품의 상세 정보를 통해서만 가능하다. 이러한 정보를 구매자가 한눈에 파악할 수 있도록 제공하는 것은 상품에 대한 구매를 결정하는데 매우 중요한 요소가 될 수 있다.

인터넷 쇼핑몰에서 제공하는 개별 상품의 상세 정보는 해당 상품에 대해서 쉽게 파악할 수 있도록 간략하게 제시되어 있다. 상세 정보를 표현하는 문장들이 갖는 특징들은 그림 3과 같이 크게 두 가지로 나눌 수 있다.

특징1: 하나의 기능 또는 비슷한 기능은 하나의 문장에 표현
특징2: 상세 정보 표현에 어느 정도 일정한 형식을 따름

그림 3 상세 정보 표현 문장의 특징

각 특징에 대해서 좀 더 살펴보면 다음과 같다. 특징1은 문장이 구성하는 내용은 일반적으로 하나의 기능을 설명하고 있고 문장의 끝은 일반적인 자연어 문장의 서술어 형태로 나타나기 보다는 HTML의
, <P> 또는

<TABLE>의 <TR>과 같은 태그로 나누어진다. 특징2는 구매자에게 상품이 갖는 기능을 보기 쉽게 보여주기 위해서 HTML의 TABLE 형식을 이용하거나 어떤 특별한 구분자을 이용하고 있다. 또한 TABLE을 이용하거나 구분자를 사용하는 경우에 앞부분은 기능에 대한 용어가 뒷부분은 기능의 설명이 따른다. 특징2는 그림 4와 같이 구분자의 위치에 따라 구분할 수 있다[4].

특징2-1 구분자가 InnerText에 나타남 특징2-2 구분자가 OuterHtml에 나타남

그림 4 구분자를 포함하는 문장

특징2-1은 {용어}와 {설명}이 웹브라우저에 나타나는 일반적인 텍스트 형태의 심벌로 나타난다. 표 1에서는 {용어}와 {설명}을 구분하기 위해서 InnerText에 구분자 ":"를 사용하고 있다. 특징 2-2는 HTML 태그 정보를 이용하지 않으면 얻을 수 없는 특징으로 태그를 이용하여 구분하는 경우는 대부분 <TABLE> 태그의 <TR>에서 앞 <TD> 태그에 {용어}가 뒤 <TD> 태그에 {설명}이 나타나는 경우이다. 그림 5는 특징2-1과 특징2-2의 형태를 나타내고 있다.

{용어} : {설명} <TABLE><TR><TD>{용어}</TD><TD>{설명}</TD>

그림 5 구분자를 포함하는 문장에서의 {용어}와 {설명}의 위치

특징2의 예외적인 사항은 InnerText나 OuterHtml에 구분자가 없는 경우가 있다. 그림 6은 구분자를 포함하지 않는 문장의 특징이다.

특징2-3 단순히 그러한 기능이 있음을 나타냄 특징2-4 익히 알려진 단위나 {설명}을 이용하여 {용어}를 생략

그림 6 구분자를 포함하지 않는 문장

특징2-3은 모델B의 "다기능 간단 리모콘" 과 같은 표현이나 "영문 캡션 자막" 과 같은 표현이다. 특징2-4는 "인치" 또는 "" 같은 단위를 사용하여 {용어}를 생략하는 경우이다. 예를 들어, 모델B에서 "29" 완전평면" 과 같은 표현은 "" 를 이용하여 "화면 크기"라는 {용어}를 생략한 경우이다.

4) 구분자의 앞부분에 놓이는 기능에 대한 용어를 {용어,T<상위용어>}로, 구분자의 뒷부분에서 용어를 설명하는 부분을 {설명,D<용어>}으로 사용한다.

4. 상품 상세 정보 분석

그림 7은 상품 상세 정보 분석을 위한 시스템 구성도이다. 시스템은 크게 상품 상세 정보 입력문의 전처리 부분과 상세 정보 분석 부분으로 나누어진다.

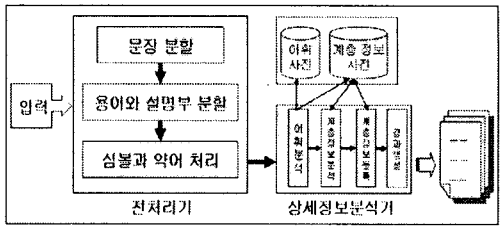

그림 7 시스템 구성도

4.1 계층 정보 사전

계층 정보 사전은 상품을 구성하는 기능의 계층성을 고려하여 구축하였다. 상품의 기능을 표현하는 의미로 {용어}를 사용하고 기능을 설명하는 의미로 {설명}을 사용하였다.

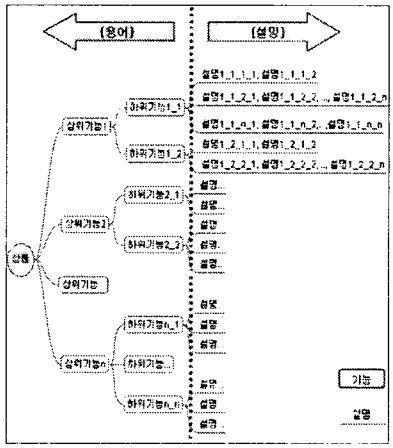

그림 8 상품의 기능 계층 구성도

그림 8은 상품의 기능을 계층 구성도를 이용하여 표현하는 방식을 나타낸 것으로 구성도의 왼쪽은 {용어}를 중심으로 상품의 기능을, 오른쪽은 왼쪽의 {용어}를 표현하는 {설명}을 나타내고 있다. 계층 구성도를 구성할 때 사용하는 엔트리는 상품의 상세 정보에서 추출한 명사를 대상으로 하였다.

그림 9는 본 논문에서 실험의 대상으로 삼은 TV에 대한 기능 계층 구성도의 일부로 {용어}부분에 대한 것이다.

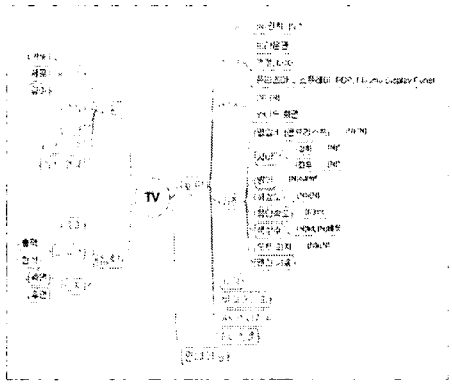

그림 9 TV에 대한 기능 계층 구성도

그림 9의 {설명}에 해당하는 부분에서 "{N}" 과 같은 표현은 숫자가 올수 있음을 의미하고, {용어} 부분에서 "명암비{콘트라스트}" 는 같은 의미의 다른 {용어}를 나타내고 있다. {용어}부의 "◦" 은 하위에 하위 기능이나 {설명}이 있음을 의미한다. {용어}의 "크기" 는 표현 방식에 따라서 {크기,T<화면>}으로 나타내고, {설명}의 "와이드 화면" 은 {와이드 화면,D<비율>}로 나타낸다.

TV에 대한 기능의 계층 정보 구조로 사용한 엔트리는 래퍼를 이용하여 추출한 인터넷 쇼핑몰 업체의 가전제품 중 TV에 대해서 추출한 20개 모델의 데이터를 분석하여 구성하였다.

4.2 입력 문장의 전처리

상품 상세 정보를 분석하기 위한 입력 문장으로는 상품 상세 정보 표현 문장의 특징을 이용할 수 있도록 OuterHtml 형태의 문장을 사용한다.

전처리 단계에서는 상세 정보 분석기가 입력 문장에서 정보를 추출하기 전에 문장의 구성 형태를 이용하여 다음의 정보와 InnerText 형태의 입력을 추출한다.

1. 문장의 경계와 문장의 구성 형식을 파악한다.
2. 구분자를 이용하여 {용어}의 위치와 {설명}의 위치 정보를 구한다.
3. 의미있는 심벌과 약어에 대한 처리하여 용어를 통일시킨다.

OuterHtml 형태의 입력 문장으로부터 문장의 경계를 구하기 위해서는 특징1을 이용한다. 문장의 경계를 구할 때 이용하는 태그 정보로는 줄 바꿈에 이용되는
과 <TABLE>의 Row를 구성하는 <TR> 태그를 이용하여 문장의 경계를 구한다.
이나 <TABLE>이 사용되지 않은 경우에는 OuterHtml에 <PRE> 태그가 사용되어 "₩n"으로 문장의 구분이 가능하다. 문장의 들여 쓰기 정보를 이용하여 문장의 구성 형식을 파악한다. 문장의 구성 형식은 한 문장이지만 이전 문장의 기능을 세부적으로 나눌 때 일반적으로 사용하는 형식이다.

상세 정보 분석 단계에서 특징2를 이용할 수 있도록 입력 문장에 구분자가 나타날 경우 {용어}의 위치와{설명}의 위치 정보를 추출한다. {용어}와 {설명}의 위치를 파악하기 위해서 InnerText에 나타나는 구분자와 OuterHtml에 나타나는 구분자를 이용한다. InnerText에 나타나는 구분자는 각 쇼핑몰마다 달라질 수 있기 때문에 전처리기에 미리 지정 해 준다. OuterHtml에 나타나는 구분자는 대부분 <TABLE> 태그의 <TD>를 이용하여 구할 수 있다.

전처리기에 의해서 처리되는 심벌과 약어에 대한 처리는 InnerText에 나타나는 심벌들로서 무의미하여 분석기가 처리하지 않도록 하거나 의미를 부여하여 분석기가 분석에 이용하도록 하여야 한다. 표 3은 상품의 상세 정보로부터 수집된 데이터에서 검출된 TV에 대한 의미 있는 심벌이다. 표 4는 추출된 약어의 일부이다.

심벌	의미	심벌	의미
"	인치	kg	무게
:	비율	°	각도
W	출력	ms	속도
Ω	저항	ch	채널

표 3 의미 있는 심벌 예

약어	의미	대표 약어
AV	Audio-Visual	AV
A/V	Audio-Visual	AV
HD	High-Defined	HD
VM	윤곽 보정 회로	VM
AKB	자연색 보정 회로	AKB

표 4 약어 예

전처리기가 담당하는 심벌의 처리는 단위를 표현하는 심벌에 대해서 상세 정보 분석 단계에서 처리 가능한 의미 있는 정보를 부여한다. 상품의 상세 정보를 표현할 때, 심벌과 그 심벌의 의미를 표현한 단어를 혼용해 사용하기 때문에 두 표현을 동일하게 만들어 줄 필요가 있다. 예를 들어 "18인치" 와 "18"" 나 "160도"와 "160°"의 표현들은 각각 같은 의미를 다른 방법으로 표현한 것이다.

상품 정보에서 사용하는 약어의 경우도 같은 약어를 표현 방식이 다르게 사용되는 경우가 있다. 예를 들어서 "A/V" 와 "AV" 는 "Audio-visual"의 약어로 상품의 상세 정보에 "A/V" , "AV" , "Audio-visual" 세 표

현이 나타나는 것이 가능하다. 전처리기는 약어를 하나의 대표 약어로 통일시켜 상세 정보 분석 단계에서 이용할 수 있도록 한다.

표 5는 입력 문장에 대한 전처리를 수행하여 형태소 분석 단계에 전달해주는 처리된 형태에 대한 예이다.

<table>
<tr><td colspan="3">1. 구분자를 갖는 입력 문</td></tr>
<tr><td>입력</td><td colspan="2">화면 크기 : 21인치</td></tr>
<tr><td rowspan="3">전처리 결과</td><td>{용어}부</td><td>화면 크기</td></tr>
<tr><td>분리자</td><td>:</td></tr>
<tr><td>{설명}부</td><td>21인치</td></tr>
<tr><td colspan="3">2. 구분자가 없는 입력 문</td></tr>
<tr><td>입력</td><td colspan="2">17" 완전평면</td></tr>
<tr><td>전처리 결과</td><td colspan="2">17인치 완전평면</td></tr>
</table>

표 5 입력 문에 대한 전처리 결과

4.3 상세 정보 분석

상세 정보 분석은 네단계로 구분되며 첫 번째 단계는 계층 정보 사전을 이용하여 어휘적인 분석을 한다. 두 번째 단계는 분석된 결과를 이용하여 상품의 계층 정보 사전에서 얻어진 정보를 이용하여 계층 정보를 분석한다. 세 번째 단계는 분석된 계층 정보중 미등록어에 대한 추정 결과를 계층 정보 구조에 반영한다. 마지막으로 다음 단계에서 사용할 수 있는 결과를 튜플 형태로 생성해 준다.

1. 어휘 분석

어휘 분석 단계는 입력 문을 구성하는 어절에 대한 계층 정보를 계층 정보 사전을 검색하여 부여한다. 계층 정보 사전에 등록된 어절은 해당 계층 정보를 부여하고 계층 정보 사전에 등록되지 않은 어절(미등록어)는 규칙에 의해서 계층 정보를 추정하여 부여한다. 어휘 분석단계에서는 계층 정보를 검색하기 전에 계층 정보를 검색할 단어인지를 알아보기 위해서 저수준의 형태소 분석을 수행한다. 형태소 분석 결과 용언과 같은 어절에 대해서는 계층 정보 검색에 이용하지 않도록 한다.

미등록어에 대한 계층 정보 추정하는 규칙은 입력 문에 {용어}와 {설명}을 구분하는 구분자의 유무에 따라 적용하는 규칙을 선택한다. 구분자가 있는 입력 문에서의 미등록어의 추정은 미등록어의 위치에 따라 추정의 범위를 제한한다. 미등록어가 {용어}부에 나타난 경우 {용어]로 추정을 시도하고, {설명}부에 나타난 경우는 {설명}으로 추정을 시도한다. 다음의 규칙1에서 규칙5까지는 구분자가 있는 문장에서의 미등록어 추정규칙이다.

규칙1 {용어}부를 이루는 어휘 가운데 일부가 검색에 성공하여 계층 정보를 가지고 있다면 미등록어의 계층 정보에 검색에 성공한 계층 정보를 부여한다.

규칙2 {용어}부를 이루는 모든 어휘가 미등록어라면 새로운 {용어}가 출현했다고 추정하고 {설명}부 어휘가 속하는 계층에 속하는 {용어}라 추정한다.

규칙3 {설명}부를 이루는 어휘 가운데 일부가 검색에 성공하여 계층 정보를 가지고 있다면 미등록어의 계층 정보에 검색에 성공한 계층 정보를 부여한다.

규칙4 {설명}부를 이루는 모든 어휘가 미등록어라면 새로운 {설명}이 출현했다고 추정하고 {용어}부 어휘가 속하는 계층에 속하는 {용어}라 추정한다.

규칙5 입력 문의 모든 어절이 규칙2와 규칙4에 해당하는 경우에는 구분자를 중심으로 {용어}부의 미등록어를 새로 출현한 {용어}로 추정하고 {설명}부의 미등록어를 새로 출현한 {용어}에 대한 설명으로 추정한다.

규칙1과 규칙2는 구분자를 중심으로 {용어}부에 출현한 미등록어를 추정하는 규칙이고, 규칙3과 규칙4는 {설명}부에 출현한 미등록어 추정규칙이다.

다음의 규칙 6과 규칙7은 구분자가 없는 입력 문장에 대한 추정 규칙이다.

규칙6 계층 정보 검색에 성공한 어휘 정보의 계층 정보를 따른다.

규칙7 입력 문의 모든 어휘가 미등록어이면 새로운 용어의 출현으로 추정한다.

규칙8은 미등록어가 출현한 문장이 위에서 처리된 문장의 세부 항목을 나타내는 문장일 때 미등록어 추정 규칙이다.

규칙8 {용어}부의 어휘가 미등록어이면 위 문장이 갖는 {용어}부의 계층정보를 이어 받는다.

그림 10과 그림 11은 규칙1에서 규칙7까지를 상품의 상세정보에 적용하는 예를 보이기 위한 계층 구성도의 일부와 입력의 일부이다.

```
* 화질:
  - 주사방식 : 순차 주사(프로그레시브 스캔)
```

그림 10 예1>을 위한 입력 문장

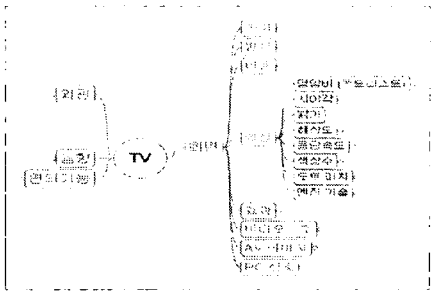

그림 11 TV 기능 중 화질을 나타내는 계층 구성 부분

"주사방식 : 순차 주사(프로그레시브 스캔)" 문장은 이전 문장 "화질" 의 기능을 설명하는 세부 기능중의 하나이고, 분리자 ":" 을 중심으로 {용어}부와 {설명}부를 구분하고 있음을 알 수 있다.

입력	- 주사방식 : 순차 주사(프로그레시브 스캔)
{용어}부	{주사방식,T<UnKnown>}
{설명}부	{순차 주사,D<UnKnown>}
	{프로그레시브 스캔,D<UnKnown>}
상위어	{화질,T<화면>}

표 6 그림10 에 대한 어휘 처리 결과

표 6은 그림10에 대한 어휘 처리 결과로서 T<UnKnown>은 {용어}부에서 D<UnKnown>은 {설명}부에서 미등록어임을 나타낸다. 상위어는 전처리에서 전달해주는 정보로 입력문의 구성 형식에 의해서 상위 문장의 하위 기능임으로 표현되었음을 나타낸다. {용어}부, {설명}부의 모든 어휘가 미등록어이므로 미등록어를 추정하기 위해서 규칙5와, 상위어가 존재하므로 규칙8을 적용할 수 있다. 규칙5를 적용하여 새로운 {용어}({주사방식,T<TV>})와 새로운 {설명} ({순차 주사,D<주사방식>}, {프로그레시브 스캔,D<주사방식>})이 출현했음을 알 수 있다. 규칙8을 적용하여 새로운 {용어}는 {화질,T<화면>}의 하위 기능임을 알 수 있다. 표 7은 규칙5를 적용한 후 미등록어 추정 결과를 표 8은 규칙 8을 적용한 후 미등록어 추정의 결과이다.

{용어}부	{주사방식,T<TV>}
{설명}부	{순차 주사,D<**주사방식**>}
	{프로그레시브 스캔,D<**주사방식**>}

표 7 규칙5를 적용한 결과

{용어}부	{주사방식,T<**화질**>}
{설명}부	{순차 주사,D<주사방식>}
	{프로그레시브 스캔,D<주사방식>}

표 8 규칙8을 적용한 결과

미등록어 추정 후, 다음 절의 계층 정보 분석결과 계층 정보가 모두 일관된다면 다음과 같은 계층 정보 구성이 수정된다. 그림12는 미등록어 추정 후 수정된 계층 구성도를 나타내고 있다.

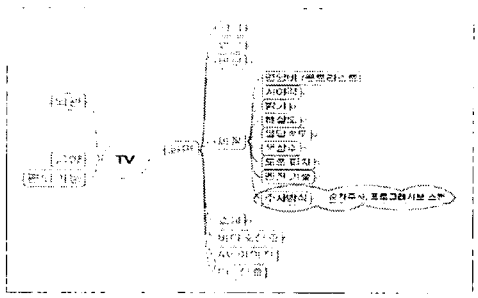

그림 12 새로 발견된 내용이 추가된 계층 구성도

2. 계층 정보 분석

계층 정보 분석 단계는 어휘 분석에서 얻어진 어휘의 계층 정보 사이의 불일치 정보를 수정한다. 계층 정보의 불일치 형태는 다음 4가지 패턴으로 나타날 수 있다.

패턴1: {용어}부의 계층 정보와 {설명}부의 계층 정보가 서로 다르다
패턴2: {용어}부 검색에 성공한 어휘들의 계층 정보가 서로 다르다
패턴3: {설명}부 검색에 성공한 어휘들의 계층 정보가 서로 다르다.
패턴4: 구분자가 없는 경우 계층 정보가 둘 이상 나타날 수 있다.

다음 규칙9에서 규칙12까지는 각 패턴에 따라서 불일치하는 계층 정보를 수정하는 규칙들이다.

규칙9 패턴1은 {용어}부의 계층 정보를 {설명}부의 계층 정보보다 우선하도록 하여, {설명}부의 계층 정보를 {용어}부의 계층 정보로 수정한다.
규칙10 패턴2는 {설명}부의 계층 정보들 사이에 불일치가 없고, {용어}부의 계층 정보에 {설명}부의 계층 정보를 포함하고 있다면 {설명}부의 계층 정보를 따르도록 한다.
규칙11 패턴3은 {용어}부의 계층 정보 내에서 불일치가 없다면 규칙8을 적용한다.
규칙12 패턴4는 계층 정보에 따라서 {용어}를 설정하여 세분화 시켜준다.

3. 계층 정보 추가 및 결과 생성

계층 정보 추가는 미등록어의 추정이 이루어진 결과를 이용하여 새로운 기능이나 기능에 대한 새로운 설명을 추가한다. 결과 생성은 상품의 상세 정보에 대한 입력을 계층 정보 분석의 결과를 이용하여 {용어, 설명}

의 형태로 결과를 생성하여 결과물 출력에 활용할 수 있도록 한다.

5. 실험 및 평가

본 논문은 인터넷 쇼핑몰에서 TV에 대한 상품 20개를 선정하여 분석할 결과를 토대로 상품을 구성하는 {용어} 사이의 계층 정보 구성도를 구하고 해당 {용어}에 사용된 {설명}에 대한 용어를 구축하여 실험을 하였다. 표 9는 계층 정보 사전을 초기 구축할 때 사용된 엔트리 수로 {용어}부의 경우 상위 기능과 같은 의미의 다른 용어를 포함한 수이다.

본 논문이 제안한 방법에 대해서 평가하기 위해서 선택한 동종의 상품들에 대해서 적절한 결과를 나타냈는지 분석하여 보았다. 비교하기 위한 상품의 선택은 계층 정보 사전을 구축할 때 사용한 상품 20여개의 상세 정보와 구축에 사용하지 않은 상품 20개를 대상으로 하였다.

평가 기준은 (1)분리({용어}와 {설명}이 잘 분리되었는가)와 (2)범주 선택({용어}와 {설명}이 적절한 범주에 소속되었는가)로 하였으며 정확률을 구하였다. 표 10은 평가 결과로서, A 항목은 계층 정보 구축에 이용한 상품에 대한 결과를, B 항목은 전체 상품 (960여개) 가운데 계층 정보 구축에 이용하지 않은 상품 중 20개를 선택하여 실험하였다.

분류	개수
{용어}	41
{설명}	268

표 9 계층 정보 구성 엔트리

평가기준	A	B
분리	100%	98.75%
범주선택	100%	95.12%

표 10 평가 결과

실험 B의 실험 결과 중 분리에서 발생한 오류는 {용어}와 {설명}을 분리하기 위해 사용한 분리자가 다른 문장의 형태로 {설명}부에 출현한 경우가 있었기 때문이다. 범주 선택에 대한 오류는 분리에서 파생된 결과와 어휘 분석을 하는 단계에서 주로 발생하였다. 어휘 분석 과정에서 간단한 형태소 분석을 사용하였기 때문에 복합명사에 의해서 미등록어가 아니지만 미등록어로 추정한데서 비롯되었다.

6. 결론

인터넷 사용이 일반화 되고 인터넷 쇼핑몰에서 상품을 구매하는 사용자가 늘고 있는 추세이다. 구매자의 구매 패턴을 파악하고 구매자의 편의를 증진시켜 줄 때 인터넷 쇼핑몰을 이용하는 구매자의 만족도는 더욱 증대 될 수 있다.

구매 패턴의 1단계인 상품의 선택을 도울 수 있는 방법으로 상품의 상세 정보를 비교할 수 있어야 한다. 각 상품들의 상세 정보는 표현의 방식과 기능을 설명하는 용어와 순서가 서로 다르기 때문에 상품들의 상세 정보를 비교하기는 쉽지 않다.

본 논문은 상품의 상세 정보들이 구매자가 쉽게 파악할 수 있게 구성되어 있다는 점을 이용하여 상품의 상세 정보를 표현하는 문장 구성의 특징을 제시하고 문장 구성의 특징을 이용하여 추출된 상품의 상세 정보로부터 상품의 계층 정보를 구축하였다. 상품 상세 정보를 표현하는 문장의 특징과 상품의 계층 정보를 이용한 상품의 상세 정보를 분석 비교 결과를 보여줄 수 있도록 계층 정보에 등록되지 않은 미등록어의 계층 정보 추정 방법과 계층 정보들 사이의 충돌에 대한 해결 방법을 제안했다.

본 논문은 가전제품 가운데 TV에 대해서 실험을 하였지만 각 상품에 맞는 계층 정보 구성을 통하여 다른 상품들에도 적용할 수 있을 것이다.

향후 연구 과제로는 상품의 상세 정보를 분석하기 위해서 필요한 상품들의 계층 정보의 구성을 본 논문이 제시한 상품 상세 정보 문장의 특징과 언어 분석 방법을 이용하여 자동화할 수 있는 방법의 연구와 함께 상품의 계층 정보를 온톨로지로 구축하여 이용할 수 있도록 해야 할 것이다. 또한 상품 모델사이의 기능적인 차이를 파악하여 각 기능이 가격에 미치는 영향을 분석할 수 있다면 구매자의 상품선택에 더 많은 정보를 제공할 수 있을 것이다.

참고 문헌

[1] 중앙일보 기사, 인터넷 쇼핑, 10년만에 10조원 시장으로...,2006년 6월 1일

[2] 오정연, 엄지족의 생활 백서, 한국 전산원 NCA ISSUE REPORT 제 8호, 2006.4.20

[3] 이기성, 유영훈, 조근식, 시멘틱 웹 기반의 이미지 검색을 이용한 비교 쇼핑 시스템, 한국정보과학회 학술발표논문집 2004년도 봄(B), 2004. 4, pp. 556 ~ 558

[4] 이경전, 전자 상거래 소프트웨어 에이전트, 정보처리 학회지 제 6권 제 1호, 1999.1, pp. 54-62

[5] 김현돈, 유창국, 최명근, 조성배, 비교 구매 기능을 갖는 쇼핑 에이전트의 원형 개발, 1997. 한국 정보과학회 가을 학술 발표 논문집 Vol. 24. No. 2 pp. 289-292

[6] 정재목, 김형주, 웹 정보의 추출 및 통합을 위한 래퍼 시스템, 정보과학회논문지:컴퓨팅의 실제 제 9권 5호, 2003.10, pp. 551-559

[7] 김수경, 안기홍, 시멘틱 웹 기반의 비교구매 에이전트를 위한 동적 웹 온톨로지에 대한 연구, 한국 지능정보시스템학회논문지 제 11권 제 2호, 2005.11, 31-45

[8] 최중민, 인터넷 정보 추출 에이전트, 정보과학회지 제18권 제5호, 2000. 5, pp. 48-53