Journal of The Korea Society of Computer and Information
Vol. 17 No. 9, September 2012
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci/2012.17.9.017

효율적인 상품등록을 위한 워드넷 기반의 오픈마켓 카테고리 검색 시스템

홍 명 덕*, 김 장 우*, 조 근 식*

A WordNet-based Open Market Category Search System for Efficient Goods Registration

Myung-Duk Hong*, Jang-Woo Kim*, Geun-Sik Jo*

요 약

여러 오픈마켓에서 판매자가 동일한 상품을 등록할 시에 각 오픈마켓마다 다른 기준으로 제공되는 카테고리로 인하여 카테고리 선정에 어려움이 발생한다. 본 논문에서는 판매자가 오픈마켓에서 상품 등록 시 다른 오픈마켓에서 기 판매하고 있는 상품의 카테고리와 의미적으로 가장 연관성이 높은 카테고리를 추천하는 방법을 제안한다. 이 때 입력받은 카테고리를 의미 분석하는 방법으로 형태소 분석, Wiki 낱말사전, WordNet, Google 번역 서비스를 사용하여 추출된 색인어로 카테고리를 검색한 후, 의미적 연관성 측정을 통하여 가장 의미가 비슷한 카테고리를 추천하는 방법이다. 실험 결과로 색인어 기반의 검색방법 보다 제안하는 의미분석 검색방법이 정확한 검색결과를 보여주어 시스템의 신뢰도를 향상시켰으며, 카테고리를 선택하는데 드는 시간비용을 절감해주는 것을 보인다.

▸ Keywords : 전자상거래, 오픈마켓, 카테고리 검색 시스템, 워드넷

Abstract

Open Market is one of the key factors to accelerate the profit. Usually retailers sell items in several Open Market. One of the challenges for retailers is to assign categories of items with different classification systems. In this research, we propose an item category recommendation method to support appropriate products category registration. Our recommendations are based on semantic relation between existing and any other Open Market categorization. In order to analyze correlations of categories, we use Morpheme analysis, Korean Wiki Dictionary, WordNet and

• 제1저자 : 홍명덕 • 교신저자 : 홍명덕
• 투고일 : 2012. 04. 26, 심사일 : 2012. 06. 05, 게재확정일 : 2012. 07. 25.
* 인하대학교 컴퓨터정보공학과(Dept. of Computer and Information Engineering, Inha University)

Google Translation API. Our proposed method recommends a category, which is most similar to a guide word by measuring semantic similarity. The experimental results show that, our system improves the system accuracy in term of search category, and retailers can easily select the appropriate categories from our proposed method.

▸ Keywords : e-Commerce, Open market, Category search system, WordNet

I. 서 론

현대 산업사회에서 정보통신 기술이 발전함에 따라 인터넷이 비즈니스 수단으로써 이용가치가 높아지고 있다. 이로 인해 온라인으로 상품이나 서비스를 거래하는 새로운 경제 패러다임이 형성되었고, 시간적 공간적 제약을 뛰어넘는 e-비즈니스 또는 전자상거래(e-Commerce)가 새롭게 등장하였다. 전자상거래 시장의 규모는 2008년 약 630조원, 2009년 약 672조원, 2010년 824조원으로 지속적으로 매년 성장하고 있으며, 이 중 인터넷쇼핑 거래액은 2008년 약 19조원, 2009년 약 20조원, 2010년 약 25조원으로 2006년 이후 가장 높은 증가율을 보이고 있다[1].

인터넷쇼핑몰 시장의 일부인 오픈마켓(Open market)은 저비용으로 소비자가 판매자가 될 수 있는 공간을 제공하는 C2C(Customer To Customer)시장으로, 2011년 3/4분기 인터넷쇼핑몰 시장은 7조 2,770억원으로 전년 동분기 대비 16.3% 증가한 시장규모를 형성하였으며 또한 진입장벽이 낮다는 장점 때문에 국내 전자상거래 시장에서 차지하는 비율 또한 매년 성장하고 있는 추세이다[1]. 현재 옥션, G마켓, 11번가, 인터파크, 농협의 NH마켓, 네이버의 체크아웃이 한국의 대표적인 오픈마켓이다.

오픈마켓 시장이 매년 성장하고 판매자에게 무료 또는 저비용으로 판매장소를 제공하기 때문에 판매자들은 대부분 두 곳 이상의 오픈마켓에 반복적인 판매과정을 수행한다. 이러한 반복적인 판매 과정을 보다 쉽고 편리하게 제공하기 위해 써드파티(Third Party) 업체들이 제공하는 서비스에도 판매중인 상품을 다른 오픈마켓으로 등록해주는 기능이 있다. 써드파티란 오픈마켓에서 상품 및 배송관리를 통합적으로 지원해주어 판매자들이 판매 작업을 좀 더 편리하게 사용할 수 있는 환경을 제공해주는 기업을 말한다.

오픈마켓에 상품을 등록하거나 다른 오픈마켓에서 판매중인 상품을 복사등록하는 과정 중 중요한 단계로 카테고리 선정이 있다[2]. 이는 상품 등록 시 각 오픈마켓에서 제공되는 카테고리가 상이하기 때문이다. 이로 인해 발생되는 비용을 최소화하기 위해 써드 파티 업체에서는 카테고리 자동매칭 서비스를 제공하고 있다. 2010년 상반기에는 옥션과 G마켓에서 서로 간에 상품을 복사등록하는 서비스가 추가 되었으며, 이 과정에서 카테고리를 자동으로 매칭해주는 "카테고리매칭"을 제공하고 있다. 카테고리 매칭하는 방법은 일반적으로 오픈마켓에서 공개하는 카테고리를 웹스크랩하여 수집하거나, 오픈마켓 CM(Category Manager)으로부터 카테고리를 제공 받은 후 수동으로 매칭작업을 하는 것으로 조사되었다.

대량의 상품을 판매하는 판매자의 경우 상품을 새로 등록하거나 다른 오픈마켓에서 판매하고 있는 상품을 복사등록할 목적으로 오픈마켓 상품등록페이지에서 상품별로 카테고리를 일일이 설정하고 상품을 등록하는 일은 쉽지 않다.

본 연구에서는 오픈마켓을 위한 카테고리 검색 시스템을 제안하여 서로 상이한 카테고리를 가지는 오픈마켓의 판매자들이 상품등록 수행 시 빠르고 효과적으로 카테고리를 검색할 수 있도록 제안한다.

II. 관련 연구

사용자가 원하는 카테고리를 정확하게 검색할 수 있도록 하기 위해서 다양한 연구가 진행되어 왔다[2,3,4,5,12].

상품등록 전 적절한 카테고리를 찾기 위하여 키워드로 검색 시 오픈마켓 사이트의 검색어가 단일 키워드라면 그 의미를 파악하기 어려워 키워드를 포함하는 모든 상품을 나타내게 되는데, 2개 이상의 조합 키워드로 검색 시 의미가 단일 키워드로 검색할 때 보다 좀 더 명확해진다. 예를 들어 [표 1]과 같이 단일 키워드 "받침대"와 2개 이상의 조합 키워드 "노트북 받침대"로 검색할 경우 2개 이상의 조합 키워드가 비교적 정확한 검색 결과를 얻을 수 있다.

[표 1]과 같이 검색 의도를 명확하게 파악할 수 있도록 하기 위한 개인화된 정보 제공에 대한 연구가 있었다[3]. 이 방법은 사용자의 검색기록에 의존한 방식이기 때문에 다른 의도로 검색 시 정확한 결과를 얻을 수 없다는 단점이 존재한다.

사용자의 검색 의도를 파악하기 위한 연구로 OntoSeek이 제안되었다[4]. OntoSeek는 단일 공급자 카테고리에 대한 검색도구이며, 다양한 카테고리를 표준화하기 위한 방법으로 이를 온톨로지로 생성하였다. 이 방법은 미리 구축된 온톨로지를 사용함으로써 보다 정확하다.

표 1. 상품 검색 결과 비교
Table 1. Comparison of goods search results

구분	단일 키워드 "받침대"			2개 이상의 조합 키워드 "노트북 받침대"		
	정확한 상품수 (건)	검색된 상품수 (건)	정확도 (%)	정확한 상품수 (건)	검색된 상품수 (건)	정확도 (%)
옥션	3,307	29,926	11.05	5,716	7,327	78.01
G마켓	4,729	30,754	15.38	3,739	4,789	78.07
11번가	2,083	34,254	6.08	3,328	4,684	71.05
인터파크	2,393	15,381	15.56	1,608	2,047	78.55

온톨로지 매핑에 대한 방법론을 쇼핑몰 환경에 적용함으로써 서로 이질적인 상품 카테고리로 구성된 두 쇼핑몰 간의 상품에 대한 매핑 알고리즘에 대한 연구가 있었다[5]. 이 방법에서는 Amazon.com, Buy.com의 해외 대형 온라인 쇼핑몰의 상품 카테고리를 위한 온톨로지를 제안하는 매핑 알고리즘으로 연결하여 해결하였다.

앞서 온톨로지를 이용하는 제안된 방법들은 카테고리 매칭에 효과적이나 해외 오픈마켓과는 다르게 국내 오픈마켓은 카테고리가 온톨로지화 되어 있지 않고, 표준화된 카테고리의 기준이 없으며, 수시로 변하는 오픈마켓 카테고리에 대응하기에는 상대적으로 시간 및 유지보수 측면에서 비효율적이다.

본 연구에서는 오픈마켓 카테고리를 통합 및 표준화 하지 않고 의미분석 방법을 통하여 카테고리를 검색하는 시스템을 제안한다.

III. 오픈마켓 카테고리

3. 1. 오픈마켓 카테고리 분석

오픈마켓 판매자는 상품을 등록하고 주문 및 배송작업을 오픈마켓에서 제공하는 판매관리화면에서 처리하며, 국내 오픈마켓 중 하나인 11번가의 상품등록화면에서는 [그림 1]과 같이 카테고리를 3~4단계를 거쳐 선택할 수 있는 인터페이스가 제공되며 다른 오픈마켓에서도 [그림 1]과 유사한 인터페이스가 제공되고 있다. 본 논문에서는 상위 카테고리부터 대카테고리, 중카테고리, 소카테고리, 세부카테고리로 명명하였다.

오픈마켓 카테고리는 의미는 같지만 다른 용어의 카테고리로 구성되어있기 때문에 온톨로지를 사용하여 표준 온톨로지를 구축해 놓으면 검색 시 유용하나, 카테고리가 변경 및 수정 될 때 마다 온톨로지를 재구축하는 것은 비효율적이다.

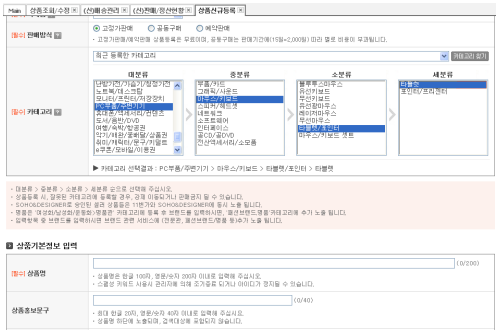


그림 1. 상품등록 시 카테고리 선택 인터페이스
Fig. 1. A interface that selects a category when registering a commodity

본 논문에서는 온톨로지를 구축하기 보단 의미분석 검색방법을 통해 카테고리 매칭을 하고자 한다. 이를 위해 카테고리 특성을 분석한 결과는 다음과 같다.

1. 공백, 특수문자[/,(,),[,],-,+]로 구분된다.
2. 대부분 명사/복합명사로 이루어져 있다.
3. 카테고리 의미는 같으나 명칭이 다르다.
4. 상품 속성에 대한 단위표시가 다르다.
5. 상품 속성을 상세하게 표시되거나 간략하게 표시된다.
6. 같은 상품이 한글 또는 영어로 표시된다.
7. 다분야에서 공통으로 사용되는 색인어로 조회 시 서로 다른 분류로 조회된다.
8. 의미는 같으나 상위카테고리가 다른 경우가 존재한다.
9. 상품의 모든 속성을 표시할 수 없으므로, "일반" 또는 "기타" 카테고리를 두어 카테고리 선택의 폭을 확장하였다.

위 특성들 중 몇몇 특성의 경우 카테고리 검색의 정확도에 큰 영향을 미친다. 특성 3의 경우 [표 2]와 같이 판매자가 카테고리를 계층적으로 검색하는 과정에서 원하는 카테고리를

찾을 수 없거나 잘못 찾는 경우가 빈번히 발생한다.

표 2. 의미는 같으나 상위 카테고리가 다른 경우
Table 2. Different classification criteria for same item

오픈마켓 구분	카테고리
옥션	모니터/프린터/부품 → 모니터 → 17형LCD -> 기타브랜드
G마켓	모니터/프린터/부품 → 모니터 → 기타브랜드
11번가	모니터/프린터/저장장치 → LCD모니터 → 43cm(17형) → 일반
인터파크	컴퓨터/노트북/프린터 → 모니터 → 기타브랜드 → 43cm(19인치)이하

특성 4의 경우 〔표 2〕를 보면 "17인치"를 옥션과 11번가에서는 "형"으로 표시하고 있으며, 인터파크는 "인치"로 표시하고 있다. 이처럼 의미는 같으나 다른 단위로 표시될 수 있으나 본 논문에서는 단위로 구분하더라도 "형(形)"이 단위인지 "형(兄)"인지 구분하기 어려우므로 형태소 분석기준에 따라 "17형", "17인치", "17" 형태로 검색을 하였다.

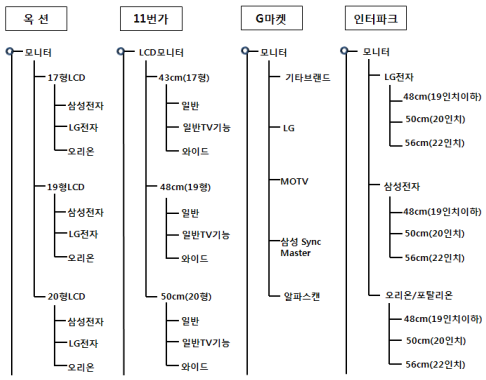


그림 2. 오픈마켓 카테고리 관계도
Fig. 2. Categories for Open Markets

특성 5의 경우 〔그림 2〕와 같이 옥션의 〔모니터/프린터/부품 → 모니터 → 17형LCD → 삼성전자〕 카테고리는 인터파크의 〔컴퓨터/노트북/프린터 → 모니터 → 삼성전자 → 48cm(19인치이하)〕 카테고리로 매칭 될 수 있으나 11번가와 G마켓 카테고리에는 맞는 카테고리가 없으므로 사용자가 판단하여 매칭 해야 한다.

특성 8의 경우 〔표 2〕와 같이 카테고리가 각각의 회사의 성격에 맞게 만들어졌기 때문에 같은 상품이라도 옥션과 G마켓에서는 "모니터/프린터/부품", 인터파크는 "컴퓨터/노트북/프린터", 11번가는 "모니터/프린터/저장장치"로 되어있다. 이처럼 다른 계층구조에서 동일성을 파악하고 이를 선택할 수 있는 방법이 요구된다. 이와 같은 문제를 해결하기 위해서 본 논문에서는 색인어로 조회된 모든 카테고리의 대카테고리와 원본카테고리의 의미를 분석하여 WordNet을 이용하여 상위 계층끼리 비교하는 방법을 제안한다.

WordNet은 프린스턴 대학의 조지 A. 밀러 심리학 교수가 지도하는 인지 과학 연구소에 의해 1985년에 만들어진 영어의 의미어휘목록이다〔6〕. 물론 한글에 관해서도 아시아 워드넷 프로젝트〔7〕, 한국어 어휘망 코렉스〔8〕, 코어넷〔9〕과 같이 공개적으로 연구가 많이 진행되고 있으나, 본 연구에서 사용하기 적합한 한국어의미어휘 목록을 제공하는 곳이 없어, 영어의미어휘 목록인 WordNet을 사용하고자 한다.

WordNet을 통한 검색어에 대해 5가지 관계가 존재하는데, 이는 Synonym(동의어, 유의어), Hypernym(상위어), Hyponym(하위어), Coordinate term(등위어), Meronyms(부분어)이다. 본 논문에서는 Hypernym을 사용하여 상위어휘를 검색하고자 한다. 〔그림 3〕과 같이 Hypernym은 단어의 상위어휘를 표시하며, 마지막 상위어휘인 개체(Entity)까지 계층적으로 표시하게 되나, 모든 단어의 너무 높은 단계끼리 비교하게 되면 모든 카테고리가 동일하다는 잘못된 결론이 도출되므로 정확한 검색을 위해 본 연구에서는 상위 3단계까지만 비교한다.

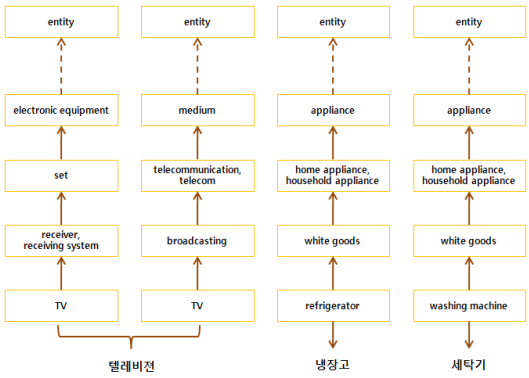


그림 3. WordNet에서 상위어휘를 조사한 결과
Fig. 3. Examples of upper level vocabularies in WordNet

특성 9의 경우, 상품의 속성은 다양하고, 고정적이지 않기 때문에 "일반", "기타" 카테고리를 두어 중요한 속성을 제외한 나머지 속성을 등록되도록 하였다. 그러나 이 용어는 모든 상품에 대해 적용되기 때문에 이 용어로 검색 시 범위가 불용한

카테고리까지 조회되는 문제가 있다. 따라서 본 연구에서는 "일반", "기타"와 같은 용어는 색인어 추출 시 제외하도록 하였다.

3. 2. 제안하는 형태소 분석 방법

3.1절에서 제시한 특성들을 가지는 오픈마켓의 카테고리 검색 시스템을 위한 의미 분석 방법으로 형태소 분석을 통하여 카테고리명의 구문을 분석하는 방법을 제안한다.

형태소 분석이란 자연 언어 분석의 첫 단계로서 단어(한국어의 경우 어절)를 구성하는 각각의 형태소들을 인식하고 불규칙 활용이나 축약, 탈락 현상이 일어난 경우 원형을 복원하는 과정을 말한다[10].

형태소 분석을 통해 카테고리명으로부터 명사를 추출하는 방법을 사용하였다. 보다 정확한 구문 분석을 위해서는 카테고리의 내용을 파악하고 미리 연관된 관계어를 정리해야 하나 시간 및 비용 측면에서 비효율적이며, 한국어 특성상 범위가 넓어지게 되고, 오히려 정확도가 떨어질 수 있게 된다. 본 연구에서는 구문 분석 범위를 명사 추출 및 상위어 비교분석으로 한정하도록 한다. 명사 추출 시 모든 사용자들이 참여하는 사전인 Wiki 낱말사전[11]에 등록된 명사를 사용하였으며, 의미를 비교하기 위해 앞서 기술한 WordNet의 상위어 유형 분석기능을 사용하고, 한영번역을 위해 Google에서 OpenAPI로 제공되는 번역 서비스를 사용하였다.

본 논문에서 제안하는 형태소 분석 방법은 다음 단계로 진행된다.

단계 1. 카테고리 특성 1에 따라 카테고리를 공백, 특수문자로 구분한다.

단계 2. 명사 추출 시 Wiki 낱말사전에 등록된 명사를 기반으로 한다.

단계 3. 영문자+숫자 형식인 경우 영문자 또는 숫자가 3음절 이상인 경우에만 분리작업을 한다.

단계 4. 영문자+한글, 숫자+한글, 영문자+숫자+한글 형식인 경우 영문자, 숫자, 영문자+숫자가 2음절 이상이라면 한글과 분리작업을 한다. 한글의 특성상 1음절 명사인 경우 검색 시 너무 많은 상황이 발생하기 때문에 2음절 이상이고 접속조사가 붙지 않은 복합명사인 경우에만 명사로 분리한다.

위 단계들을 예로 들면 단계 1의 경우 "모니터/프린터/부품"은 "모니터", "프린터", "부품"으로 분리되며, "48cm(19형 이하)"는 1차로 "48cm", "19형이하"로 구분된다.

단계 3의 경우 "2PM"이나 "MP3"는 분리하지 않는다. 하지만 "512MB"는 "512", "MB"로 분리한다.

단계 4의 경우 "토끼거북이"는 "토끼", "거북이"로 분리되나 "토끼와거북이"는 분리되지 않는다. 또한 "생활주방"의 카테고리인 경우 "생활", "활주", "주방" 3개의 명사가 추출되는데 여기서 "활주"는 카테고리 성격과 맞지 않는 명사이므로 제거하는 과정이 추가로 필요하겠다. 이 문제를 해결하기 위해 명사가 추출되면 해당 명사는 색인어에서 제외가 되고 남은 용어로 명사추출 작업이 진행되도록 하고자 한다.

IV. 오픈마켓 카테고리 검색 시스템

4.1. 검색 시스템 구조

본 논문에서는 판매자로부터 입력받거나 선택된 카테고리를 "원본카테고리"로 정의하였다. 원본카테고리는 형태소 분석을 통해 색인어가 추출되고 색인어로 각 오픈마켓에서 검색하게 된다. 그러나 오픈마켓 카테고리 특성으로 인해 원본카테고리와 다른 의미의 카테고리까지 조회된다. 이 문제를 해결하기 위하여 Google의 번역 서비스와 WordNet의 상위어 조회 기능을 사용하여 조회된 카테고리 목록 중 원본카테고리와 같은 상위어휘를 가지지 않는 카테고리를 제거하였으며, 의미적 연관성을 측정하여 판매자에게 원본카테고리와 가장 비슷한 카테고리를 추천하였다. 사용자는 [그림 4]와 같이 오픈마켓에 상품 등록 시 시스템에서 추천하는 카테고리 정보를 이용함으로써 상품등록 시간이 감소되고 대량의 상품을 효율적으로 관리할 수 있게 된다.

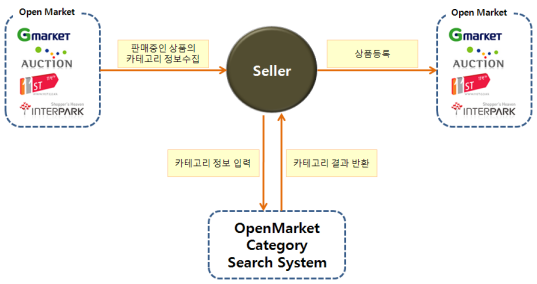


그림 4. 판매자 작업 흐름도
Fig. 4. Seller Work Process

오픈마켓 카테고리 검색 시스템은 [그림 5]와 같이 구성되어 있다. 판매자가 특정 오픈마켓 카테고리를 선택하면 시스

템은 카테고리를 분석하여 색인어를 추출한다. 오픈마켓 카테고리 특성상 동일한 구조와 용어를 사용하지 않으므로 단계별로 카테고리를 검색한다. 예를 들면 색인어를 포함하는 대카테고리, 대·중카테고리, 대·중·소카테고리, 대·중·소·세부카테고리식으로 조회한다.

원본카테고리를 분석하고 조회하는 과정은 다음과 같다.

과정 1. 판매자는 원본 오픈마켓 카테고리를 선택한다.
과정 2. 시스템은 원본카테고리를 형태소 분석하여 색인어를 추출한다. 추출 시 Wiki 낱말사전에 등록된 명사를 토대로 추출한다. 추출된 명사는 Google 번역 서비스를 이용하여 (한↔영)으로 번역된다. 또한 WordNet을 사용하여 추출된 색인어의 상위어휘를 조회한다.
과정 3. 오픈마켓 카테고리 테이블에서 추출된 색인어를 포함하는 카테고리를 조회한다.
과정 4. 조회된 카테고리와 입력받은 카테고리의 상위어휘를 비교하여 의미가 다른 카테고리를 조회된 카테고리 목록에서 제거한다.
과정 5. 원본카테고리와 조회된 카테고리간의 의미적 연관성을 측정한다.
과정 6. 판매자에게 조회된 카테고리를 연관성이 높은 카테고리를 기준으로 표시한다.

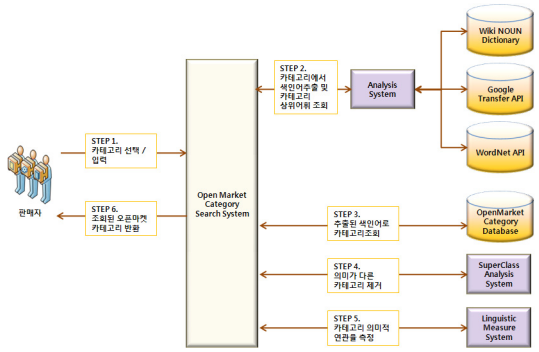


그림 5. 오픈마켓 카테고리 검색 시스템 구성도
Fig. 5. System architecture

4.2. 카테고리 분석 모듈 구성

본 절에서는 형태소 분석, Wiki 낱말사전, WordNet을 이용하여 원본카테고리의 색인어를 확장한 후 오픈마켓 카테고리에서 색인어를 포함하는 카테고리를 조회한다. 그 후 연관성을 측정하여 판매자에게 연관성이 높은 순으로 추천한다.

표 3. 오픈마켓 카테고리 수 (2010년 10월 기준)
Table 3. The number of categories

오픈마켓 구분	카테고리
옥션	10,859
G마켓	6,683
11번가	6,363
인터파크	10,703

〔표 3〕에서와 같이 오픈마켓 카테고리는 6,000 ~ 11,000여개로 조사되었다. 오픈마켓 간의 카테고리 수가 차이가 남에 따라 같은 상품에 대한 카테고리 표현방식도 다르게 된다. 예를 들면 G마켓의 카테고리가 옥션이나 인터파크에서는 의미가 좀 더 세분화된 다수의 카테고리로 표현되기도 한다.

이런 다양성 때문에 추출된 색인어로 조회된 카테고리는 다음과 같이 구분될 수 있다.

- 원본카테고리와 이름, 구조 모두 같은 경우
- 카테고리명이 일부만 일치하는 경우
  · 형태소 분석을 통해 추출된 색인어를 포함하거나 같은 경우
  · 색인어를 포함하거나 같은 카테고리는 없으나 WordNet 상위어휘가 일치하는 경우

따라서 조회된 오픈마켓 카테고리에 대해 우선순위를 정하는 작업이 필요하며, 이는 의미적 연관성 척도〔12〕를 변형하여 측정하였다.

4.2.1. 의미적 연관성 측정

조회된 오픈마켓 카테고리의 우선순위를 정하기 위하여 언어적 및 구조적 측정 방법을 사용한다. 의미적 연관성 척도는 언어적 측정값과 구조적 측정값을 합한 비율로 정의하며, 이 값이 2에 가까울수록 비교 대상인 두 카테고리가 의미적으로 연관성이 높다고 볼 수 있다.

언어적 측정(Linguistic Measure)은 색인어로 조회된 오픈마켓 카테고리에서의 용어와 입력받은 카테고리에서의

용어간의 연관성을 측정한 것이다.

언어적 가중치를 구분하기 위해 가중치 항목을 3단계로 나누어 의미적으로 멀어질수록 가중치값을 낮게 부여하는 연구가 있었다[13]. 본 논문에서도 이 기준을 바탕으로 [표 4]와 같이 가중치값을 1에서 0까지 부여하였다.

표 4. 언어적 가중치
Table 4. A numerical standard for linguistic weight

번호	관계	가중치
1	같음	1
2	구분자로 분리된 용어와 같음 (ex. 컴퓨터/모니터/프린터 -> "컴퓨터","모니터","프린터")	0.8
3	구분자로 분리된 용어가 포함되는 경우	0.6
4	WordNet에서 조회된 상위어와 같은 경우	0.3
5	다름	0
6	비교대상인 카테고리가 구조적으로 차이가 날 경우 측정된 가중치값이 〈W〉 0)이라면 가중치값에서 0.1을 뺀다.	x - 0.1 단, x = 가중치값 (1, 0.8, 0.6, 0.3, 0)

본 논문에서는 매칭행렬을 사용하여 언어적 상관관계를 측정하였으며, 측정된 가중치 중 가장 높은 값을 해당 카테고리명에 대한 가중치로 설정하고, 이는 수식(1)과 같이 정의된다.

$$\text{언어적 측정값} = \frac{\text{설정된 가중치 중 최대값의 합}}{\text{원본카테고리의 언어적 측정값}} \quad (1)$$

수식(1)이 적용된 [표 5]의 매칭행렬에서 원본카테고리로 사용된 11번가 카테고리와 옥션 카테고리의 가중치를 보면 "노트북/데스크탑"에 측정된 가중치값들 {1, 0.8, 0} 중에 가장 큰 수는 1이므로 "노트북/데스크탑"↔"노트북/데스크탑"에 대한 가중치는 1로 설정이 되며, "노트북"↔"노트북"은 1로, "IBM/레노버"↔"ASUS"는 0으로, "코어i3"↔"코어i시리즈"는 0.6으로 설정이 된다. 마찬가지로 원본카테고리에 대한 언어적 측정값을 측정하면, 정방형으로 매칭하게 되므로 가중치값이 모두 1로 설정되어 총 4가 된다. 따라서 [표 5]의 매칭행렬에 대한 언어적 측정값은 2.6 / 4 = 0.65로 계산된다.

표 5. 매칭행렬에서의 측정 예제
Table 5. An example of Measurements from a matching matrix

옥션 / 11번가	노트북/데스크탑	노트북	ASUS	코어i 시리즈
노트북/데스크탑	1	0.8	0	0
노트북	0.8	1	0	0
IBM/레노버	0	0	0	0
코어i3	0	0	0	0.6

만약에 11번가 세부카테고리인 "코어i3"카테고리가 없다고 가정하면 11번가와 옥션카테고리가 구조적으로 다른 문제가 발생하게 된다. 이 경우 [표 4]의 6번 항목인 언어적 가중치값 설정 방법에 따라 옥션의 중·소카테고리인 "ASUS", "코어i시리즈"카테고리의 언어적 연관도 측정값 들 중 가장 높은 값에 0.1을 빼서 언어적 측정값을 계산한다. 마찬가지로 만약에 11번가 카테고리가 중카테고리까지만 있고 옥션카테고리는 세부카테고리까지 있다면 옥션의 중·소·세부카테고리에서 측정된 언어적 연관도 측정값 들 중 가장 높은 값에서 0.1을 빼서 언어적 측정값을 계산한다.

의미적 연관성 척도에서 구조적 측정(Structural Measure)은 입력받은 카테고리와 조회된 오픈마켓 카테고리 간에 구조적으로 의미가 같은지를 측정하는 것이다. 이는 언어적 측정값이 높게 나와도 같은 단계의 카테고리 간 비교를 통해 더 정확한 연관성을 추정하기 위함이다. 수식(2)는 구조적 가중치이다[13]. 수식(2)는 최하위 카테고리부터 값이 증가되도록 하여 최상위 카테고리로 갈수록 가중치값이 감소되도록 하였다. 그러나 본 논문에서는 오픈마켓 카테고리 특성상 최상위 카테고리에 따라 전체 카테고리 성격이 달라지므로 하위단계로 내려갈수록 가중치값이 감소되도록 설정하였다.

$$\text{구조적 가중치} = \begin{cases} 10 & \text{if } d = 1 \\ 10(1 - 1/n)^{d-1} & \text{if } d > 1 \end{cases} \quad (2)$$

수식(2)에서 n값은 카테고리 총 단계수를 의미하며, d값은 1부터 시작하며 카테고리 단계를 나타낸다. 대->중->소-> 세부카테고리 순으로 가중치가 부여된다. 따라서 대카테고리의 구조적 가중치는 10이 되며 d값이 n까지 증가하는 동안 구조적 가중치는 점점 감소된다. 이는 오픈마켓 카테고리 특성상 중·소·세부카테고리가 바지에 대한 의미를 가지더라도 대카테고리가 "남성의류", "여성의류"로 나뉘는 것처럼 최상

위 카테고리가 원본카테고리와 밀접하게 연관되어야 하기 때문이다. 만약 같은 단계의 카테고리 간에 언어적으로 연관된 측정값이 있다면 가중치값에 1을 곱하고, 없다면 0을 곱해 구조적 가중치값을 부여하지 않는다.

〔표 4〕에서의 구조적 가중치를 계산하면 다음과 같다.

원본카테고리는 "노트북/데스크탑→노트북→IBM/레노버→코어i3"의 총 4단계로 이루어져 있으며, 언어적 측정 방법과 동일하게 정방향으로 매칭하게 되므로 구조적 가중치는 W1 = 10 + 7.5 + 5.625 + 4.219 = 27.344 로 계산된다. 조회된 카테고리의 구조적 가중치를 측정하면 소카테고리 간에는 언어적 측정값이 없으므로 가중치값이 0이 부여되어 W2 = 10 + 7.5 + 0 + 4.219 = 21.719로 계산되고, 수식(3)에 따라 〔표 4〕의 구조적 측정값은 SM2 = 21.719 / 27.344 = 0.794로 계산된다.

$$\text{구조적 측정값} = \frac{\text{조회된 카테고리의 구조적 가중치의 합}}{\text{원본 카테고리의 구조적 측정값}} \quad (3)$$

정리하면 〔표 4〕의 언어적 측정값은 0.65값을 얻었고 구조적 측정값은 0.794를 얻었다. 따라서 두 측정값의 합으로 정해지는 의미적 연관성 측정값은 1.444를 얻었다. 두 측정값의 합이 더 높을수록 원본카테고리와 비슷하다고 판단한다.

V. 실험 및 평가

본 연구의 실험 환경은 Intel Pentium 4, MS Windows XP, RAM 2GB이며, 구현을 위해 Microsoft 사의 C# 언어와 데이터 처리를 위한 DBMS로는 Microsoft 사의 Access를 사용하였다. 실험 데이터는 옥션, 11번가, G마켓, 인터파크의 카테고리를 수집하여 저장하였다. 여기서 오픈마켓 카테고리 정보는 오픈마켓 CM(Category Manager)으로부터 제공받거나 직접 수집한다는 가정을 한다.

〔그림 6〕은 오픈마켓 카테고리 검색 시스템의 실험 결과를 보여준다. 프로그램은 카테고리 선택(상단), 조회된 카테고리 목록(중단), 판매자가 원하는 카테고리를 찾아볼 수 있는 오픈마켓 카테고리목록(하단)에 대한 부분으로 구성되어 있다. 실험 데이터를 통해 색인어기반 검색방법과 의미분석 검색방법의 검색 성능을 비교하기 위해 평균 검색 시간과 각 오픈마켓 카테고리를 검색 원본으로 하여 다른 오픈마켓의 카테고리와의 매칭 여부를 실험하였다.

그림 6. 오픈마켓 카테고리 검색 시스템
Fig. 6. Open Market category search system

실험의 수행 시간은 색인어기반 검색방법은 카테고리별로 평균 1초 내로 검색결과를 확인할 수 있었고, 의미분석 검색방법의 경우 평균 5초 내로 검색결과를 확인할 수 있었다. 이는 단순히 카테고리를 조회하는 색인어기반 검색방법과는 다르게 OpenAPI를 이용하는 의미분석 검색방법이 상대적으로 더 많은 시간을 소요한다. 하지만 색인어기반의 검색은 낮은 정확성으로 인해 반복적인 카테고리 검색 비용이 추가로 발생하게 된다. 따라서 판매자가 최종적으로 카테고리를 결정 시 의미분석 검색방법이 보다 빠른 것을 보였다.

표 6. 검색결과 비교
Table 6. Comparison of search results

검색대상 / 검색원본		옥션	G마켓	11번가	인터파크	평균
옥션 (10,859개)	색인어 기반		73%	80%	77%	77%
	의미분석		88%	91%	87%	89%
G마켓 (6,683개)	색인어 기반	65%		53%	52%	57%
	의미분석	81%		82%	70%	78%
11번가 (10,703개)	색인어 기반	75%	56%		72%	68%
	의미분석	90%	84%		86%	87%
인터파크 (6,363개)	색인어 기반	73%	58%	61%		64%
	의미분석	86%	80%	79%		82%

〔표 6〕은 오픈마켓 카테고리에 대한 검색 결과를 비교한 표이다. 검색 기준은 색인어기반 검색방법의 경우 원본카테고리 와 검색된 카테고리 중에 매칭이 가능한 카테고리가 있는 경우에 검색된 것으로 측정하였고, 의미분석 검색방법의 경우

원본카테고리를 통해 검색된 카테고리들 중에서 언어적 측정값과 구조적 측정값의 합을 내림차순으로 정렬하여 가장 큰 값을 가지는 카테고리를 선정하여 그 값이 1.4보다 큰 경우 검색된 것으로 측정하였으며, 카테고리가 조회되지 않거나, 없는 경우에는 검색실패로 처리했다.

〔표 6〕은 본 논문에서 제안한 의미분석 검색방법이 색인어 기반 검색방법보다 최소 12%부터 최대 21% 사이로 평균 17.5% 더 향상된 검색결과를 보여주고 있다.

표 7. 그룹별 검색결과 비교
Table 6. Comparison of search results

검색대상 / 검색원본		그룹1	그룹2	평균
그룹1 (옥션, 11번가)	색인어 기반	77.5%	69.5%	73.5%
	의미분석	90.5%	86.3%	88.4%
그룹2 (G마켓, 인터파크)	색인어 기반	63.0%	55.0%	59.0%
	의미분석	82.0%	75.0%	78.5%
같은 그룹 (그룹1→그룹1, 그룹2→그룹2)	색인어 기반	66.3%		
	의미분석	66.3%		
다른 그룹 (그룹1→그룹2, 그룹2→그룹1)	색인어 기반	84.1%		
	의미분석	82.8%		

〔표 7〕은 오픈마켓을 전체카테고리 수를 상대적인 기준으로 보면 그룹1은 약 10,000개로 구성된 옥션과 11번가, 그룹2는 약 6,000개의 G마켓과 인터파크로 구분할 수 있다. 두 그룹 간에 카테고리 매칭 정확도를 보면 많은 카테고리를 가진 그룹1→그룹1의 색인어기반 검색방법과 의미기반 검색방법 모두가 가장 높은 정확도를 보여주었다. 반면 상대적으로 적은 카테고리를 가진 그룹2→그룹2의 색인어기반 검색방법과 의미기반 검색방법 모두 가장 낮은 정확도를 보여주었다. 이는 카테고리의 수가 적은 경우 보다 많을수록 판매자가 원하는 카테고리를 상대적으로 검색하기 수월함을 보여주었다. 또한 카테고리 검색 시 같은 그룹 간에서 정확도가 약 66%로 보여지고, 다른 그룹 간에서 정확도가 82~84%로 보여지는데, 이는 비슷한 카테고리 수를 가지는 오픈마켓일지라도 카테고리 관계도가 유사하지 않음을 보여주고 있으며, 반면에 다른 카테고리 수를 가지는 오픈마켓 간에 카테고리 관계도의 유사 여부를 떠나 보다 향상된 정확도를 보여주고 있다.

결과적으로 색인어기반 검색방법은 단순 색인어 검색으로 카테고리 관계도가 상이할수록 카테고리가 조회되지 않거나, 의미가 다른 카테고리만 조회되는 경우가 빈번히 발생하였다.

하지만 본 논문에서 제안한 의미분석 검색방법은 원본카테고리를 의미적으로 분석하는 단계를 거쳐 색인어를 추출하기 때문에 카테고리 관계도가 상이하더라도 색인어 기반 검색방법보다 정확도가 높아 성능이 우수하였음을 보였다.

VI. 결론 및 향후 연구

국내 대표적인 오픈마켓들을 살펴보면 상품/배송업무 프로세스의 기본 절차는 같지만 서로 간의 데이터가 표준화 되어 있지 않아 여러 오픈마켓에서 상품을 판매하려는 판매자의 입장에서는 어려운 점이 많다. 그 중 오픈마켓 카테고리의 수정이 한 해 평균 65회 정도로 빈번하게 발생되므로 카테고리 관리 에 어려움이 있었으며, 오픈마켓에 상품을 등록 할 때도 카테고리 선택에 많은 시간이 소요되는 문제가 있다.

따라서 본 논문에서는 오픈마켓에서 대량의 상품을 판매하는 판매자들, 그리고 써드파티 업체들이 좀 더 효율적으로 카테고리를 검색하고, 상품을 등록하기 위한 방안으로 WordNet 기반의 오픈마켓 카테고리 검색시스템을 제안하였다. 이를 위해 오픈마켓 카테고리의 특성을 분석하였고, 그 특성에 맞게 카테고리에서 색인어를 추출하였다. 추출된 색인어를 의미적으로 분석하기 위해 언어적 측정값과 구조적 측정값을 계산하여 조회하는 시스템을 설계 및 구현해 보았다. 그 결과로 색인어기반 검색방법 보다 의미기반 검색방법이 원본카테고리에 의미적으로 가까운 카테고리를 추천하여 제안하는 방법의 유용성을 확인 할 수 있었다.

향후 연구로는 향상된 카테고리 검색을 위해 보다 정확하게 동일성을 파악하기 위해서는 WordNet의 계층관계를 이용하기 보단, 공개적으로 이용할 수 있는 한국어 워드넷이 필요로 하다. 그리고 한글의 동의어, 유의어에 대한 조사가 이루어지지 않았는데 향후에는 국립국어원에서 구축한 21세기 세종계획〔14〕을 이용하여 추출된 색인어 목록에 동의어와 유의어를 추가한다면 매칭 정확도가 보다 높아질 것이다. 마지막으로 오픈마켓 카테고리가 상품의 모든 속성을 표현할 수 없기 때문에 사용자 의도를 정확히 판단하여 카테고리 매칭하기 위해 개인화 연구가 필요하다.

참고문헌

〔1〕 Korean Statistical Information Service, "E-commerce and cyber-shopping trends", http://kostat.go.kr, 2008-2011.

〔2〕 D. K. Kim, J. B. Kim, S. G. Lee, "Catalog integration for electronic commerce through category-hierarchy merging technique", Proceedings of the 12th International Workshop on Research Issues in Data Engineering: Engineering e-Commerce/ e-Business Systems, pp. 28-33, 2002.

〔3〕 D. W. Cha, Y. J. Park, "Articles : A Study on the Personalized Information Service Method in Internet Shopping Malls", Journal of Korea Corporation Management Association, Vol. 14, pp. 231-247, 2001.

〔4〕 N. Guarino, C. Masolo, G. Vetere, "OntoSeek: Content-Based Access to the Web", IEEE Intelligent Systems and their Applications, Vol. 14, No. 3, pp. 70-80, 1999.

〔5〕 W. J. Kim, N. H. Choi, D. W. Choi, "An Ontology-Driven Mapping Algorithm between Heterogeneous Product Classification Taxonomies", Journal of Intelligent Information Systems, Vol. 12, No. 2, pp. 33-48, 2006.

〔6〕 G. A. Miller, "WordNet: a lexical database for English", Communications of the ACM, Vol. 38, No. 11, pp. 39-41, 1995.

〔7〕 K. Robkop, S. Thoongsup, T. Charoenporn, V. Sornlertlamvanich, H. Isahara, "WNMS: Connecting the Distributed WordNet in the Case of Asian WordNet", Proceedings of the 5th International Conference of the Global WordNet Association, 2010.

〔8〕 S. Hwang, Y. Jung, A. Yoon, H. C. Kwon, "Building Korean Classifier Ontology Based on Korean WordNet", Proceedings of the 9th International Conference on Text, Speech and Dialogue, Vol. 4188, pp. 261-268, 2006.

〔9〕 K. S. Choi, H. S. Bae, "A Korean-Japanese-Chinese Aligned Wordnet with Shared Semantic Hierarchy", Lecture Notes in Computer Science, Vol. 2911, pp. 91-96, 2003.

〔10〕 S. S. Kang, "Korean Morphological Analysis and Information Retrieval", Hongrung Publishing Company, 2002.

〔11〕 M. C. Chimato, "It's a Wiki Wiki World", Medical Reference Services Quarterly, Vol. 26, No. 1, pp. 169-190, 2007.

〔12〕 I. H. Kwon, C. O. Kim, K. P. Kim, C. J. Kwak, "Recommendation of e-commerce sites by matching category-based buyer query and product e-catalogs", Computers in Industry, Vol. 59, No. 4, pp. 380-394, 2008.

〔13〕 S. Castano, A. Ferrara, S. Montanelli, "H-match: an algorithm for dynamically matching ontologies in peer-based systems", Proceedings of the 1st International Workshop on Semantic Web and Databases at VLDB, pp. 1-20, 2003.

〔14〕 H. G. Kim, B. M. Kang, J. H. Hong, "21st Century Sejong Modern Korean Corpora: Results and Expectations", Proceedings of the 19th Conference on Hangul and Korean Information, pp. 311-316, 2007.

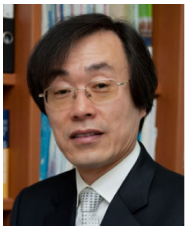
저 자 소 개

홍 명 덕

2008 : 서울디지털대학교
컴퓨터공학과 공학사
2011 : 인하대학교
컴퓨터정보공학과 공학석사
2011~현재 : 인하대학교
컴퓨터정보공학과 박사과정
관심분야 : 추천 시스템, 시맨틱 웹, 군집 지능, 메타휴리스틱
Email : hmdgo@eslab.inha.ac.kr

김 장 우

2008 : 평생교육진흥원
컴퓨터공학과 공학사
2011 : 인하대학교
컴퓨터정보공학과 공학석사
2007~현재 : Esellers R&D Center
관심분야 : 온톨로지, 데이터베이스, 전자상거래
Email : jwheat21c@naver.com



조 근 식

1982 : 인하대학교
전자계산학과 공학사
1985 : Queens Colleg/CUNY M.A.
컴퓨터공학과 공학석사
2004 : City University of New York
컴퓨터공학과 공학박사
1991~현재 : 인하대학교
컴퓨터정보공학과 교수
관심분야 : 인공지능, 시맨틱 웹, 전문가시스템, 지능형에이전트시스템
Email : gsjo@inha.ac.kr