Journal of The Korea Society of Computer and Information
Vol. 19, No. 3, March 2014
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.3.017

단어 군집 기반 모바일 애플리케이션 범주화

허 정 만*, 박 소 영*

Word Cluster-based Mobile Application Categorization

Jeongman Heo*, So-Young Park*

요 약

본 논문에서는 단어 군집 정보를 활용하여 모바일 애플리케이션의 범주를 분류하는 방법을 제안한다. 제안하는 방법은 모바일 애플리케이션 설명이 짧을 수 있다는 점을 고려하여, 모바일 애플리케이션 설명에 포함된 단어 정보 뿐만 아니라 각 단어의 단어 군집 대표 정보를 범주화 자질로 활용한다. 그리고, 모바일 애플리케이션의 카테고리가 세분화되어 있으므로, 제안하는 방법은 범주별 단어 발생 빈도를 K 평균 군집화 알고리즘에 적용하여 단어 군집을 생성한다. 모바일 애플리케이션 설명이 설치사양과 같이 범주와 관련없는 내용이 있을 수 있다는 점을 반영하여, 제안하는 방법은 단어 군집 중에서 범주화에 유용한 일부 단어 군집만을 선별하여 활용한다. 실험결과 제안하는 방법은 단어 군집 정보를 활용하여 모바일 애플리케이션 범주화 재현율을 5.65% 개선시켰다.

▸ Keywords : 모바일 애플리케이션, 범주화, 단어 군집화

Abstract

In this paper, we propose a mobile application categorization method using word cluster information. Because the mobile application description can be shortly written, the proposed method utilizes the word cluster seeds as well as the words in the mobile application description, as categorization features. For the fragmented categories of the mobile applications, the proposed method generates the word clusters by applying the frequency of word occurrence per category to K-means clustering algorithm. Since the mobile application description can include some paragraphs unrelated to the categorization, such as installation specifications, the proposed method uses some word clusters useful for the categorization. Experiments show that the proposed method improves the recall (5.65%) by using the word cluster information.

▸ Keywords : Mobile Application, Categorization, Word Clustering

•제1저자 : 허정만 •교신저자 : 박소영
•투고일 : 2013. 12. 19, 심사일 : 2014. 1. 22, 게재확정일 : 2014. 2. 20.
* 상명대학교 게임학과(Dept. of Game Design & Development, SangMyung University)
※ 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2012R1A1A3013405)

I. 서 론

최근 스마트폰이나 태블릿 PC 등의 모바일 기기가 빠른 속도로 확산되면서, 모바일 환경에서 다양한 서비스를 제공하는 모바일 애플리케이션이 기하급수적으로 늘어나고 있다. 이로 인해, 모바일 애플리케이션 스토어에서 사용자가 원하는 기능을 가진 모바일 애플리케이션을 찾는데 어려움이 따를 수 있다. 따라서, 사용자가 모바일 애플리케이션을 쉽게 찾을 수 있도록, 모바일 애플리케이션을 효율적이고 체계적으로 분류하고 관리할 필요성이 커지고 있다.

그동안 뉴스 기사나 웹 문서 등의 다양한 문서를 미리 정의된 범주 체계로 자동 분류해 주는 기술은 활발히 연구가 진행되고 있다. 대표적으로, 기계학습기반 분류 방법은 단순 베이지안 모델이나 SVM 등을 바탕으로 문서 분류에 필요한 정보를 학습집합에서 추출하여 사용하며, 스팸 분류 등에서 높은 성능을 보였다[1-4]. 이러한 방법은 분류할 종류가 많아지거나 학습집합의 크기가 작은 경우 성능을 보장하기 어렵다[4]. 한편, 지식기반 분류 방법은 워드넷과 같은 온톨로지를 활용하여 단어의 의미나 관계성과 같은 지식을 활용한다[5,6]. 그러나 온톨로지의 불완전성 때문에, 단어의 의미나 관계중의성 효과가 제한되어 있다[6]. 웹 기반 분류 방법은 웹을 돌아다니면서 특정 주제와 관련된 정보를 스스로 획득하여 사용할 수 있다[6-10]. 그러나, 매우 복잡하게 연결되어 있는 웹에서 정확한 정보를 찾기가 쉽지 않고, 이미 정확한 정답 정보를 포함하고 있는 학습집합이 어느 정도 확보된 경우는 그 효과가 제한적일 수 있다[7].

이러한 기존 문서 분류 방법을 모바일 애플리케이션 범주화 분야에 그대로 적용하기에는 다음과 같은 어려움이 있다.

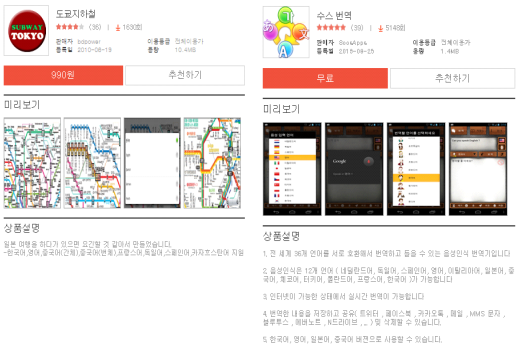

그림 1. 모바일 애플리케이션 설명 화면 예제
Fig. 1. Sample Screenshots of Mobile Application Descriptions

표 1. 모바일 애플리케이션 설명 예제
Table 1. Sample Mobile Application Descriptions

제목	설명	범주
도쿄 지하철	일본 여행을 하다가 있으면 요긴할 것 같아서 만들었습니다. - **한국어, 영어, 중국어(간체), 중국어(번체), 프랑스어, 독일어, 스페인어**, 카자흐스탄어 지원 필요하신 분만 받으세요.	여행지 도교통
수스번역	1. 전 세계 36개 언어를 서로 호환해서 번역하고 들을 수 있는 음성인식 번역기입니다. 2. 음성인식은 12개 언어(네덜란드어, **독일어, 스페인어, 영어**, 이탈리아어, 일본어, **중국어**, 체코어, 터키어, 폴란드어, **프랑스어, 한국어**)가 가능합니다. 3. 인터넷이 가능한 상태에서 실시간 번역이 가능합니다. 4. 번역한 내용을 저장하고 공유(트위터, 페이스북, 카카오톡, 메일, MMS 문자, 블루투스, 에버노트, N드라이브, ...) 및 삭제할 수 있습니다. 5. 한국어, 영어, 일본어, 중국어 버전으로 사용할 수 있습니다.	교육

첫째, 모바일 애플리케이션 설명은 길이가 일정하지 않다. 모바일 애플리케이션 등록자가 해당 애플리케이션에 대해서 [표1]과 같이 짧게 설명할 수도 있고, 4,000자 이내에서 충분히 길게 설명할 수도 있다. 모바일 애플리케이션에 대해 사용자에게 효과적으로 설명하기 위해서, [그림1]과 같이 주요 내용을 그림으로 표현하는 경우 텍스트로 작성된 설명은 더욱 짧아지게 된다. 이러한 점을 고려하여, 텍스트 정보 대신 하드웨어 접근권한 등을 이용하여 모바일 애플리케이션 군집화 할 수 있다[11]. 그러나, 이러한 방법은 모바일 애플리케이션 스토어에서 미리 정의한 범주 체계와 연관성을 찾기 어렵다.

둘째, 모바일 애플리케이션 설명은 설치사양, 지원언어, 공지사항, 광고글, 사용법 등을 포함하고 있는데, 이러한 내용은 모바일 애플리케이션의 범주를 분류하는데 별로 도움이 되지 않는다. 예를 들어, [그림1]의 텍스트 내용을 포함하는 [표1]에서 모바일 애플리케이션 "도쿄지하철"은 그 특징을 설명하는 내용보다는 지원하는 언어에 대한 내용이 많으므로, "여행/지도/교통"으로 범주화하기 쉽지 않다. 오히려, 언어교육용 모바일 애플리케이션으로 잘못 분류할 수 있다. 따라서, 모바일 애플리케이션 설명에 포함된 노이즈를 제거하여 문서 분류 성능개선에 도움을 줄 수 있다[12]. 그러나, 이러한 방법은 모바일 애플리케이션의 부족한 텍스트 정보를 더 부족하게 만들 수 있는 위험성이 있다.

셋째, 모바일 애플리케이션은 사용자가 원하는 애플리케이션을 찾기 쉽도록 범주가 세분화되어 있다. 신문기사는 전통적으로 정치, 경제, 사회, 문화 등의 10개 내외 범주로 구성되어 있다. 반면에, 모바일 애플리케이션 설명은 일반적으로

25개 내외의 범주로 구성되어 있다. 특히, 게임 범주의 경우 등록된 애플리케이션의 수가 상대적으로 많아 아케이드 게임, 액션/슈팅 게임, 시뮬레이션 게임, 스포츠게임, 퍼즐/보드 게임, RPG 게임 등으로 세분화될 수 있다. 이를 고려하여 사용자에게 상황에 맞는 모바일 애플리케이션 카테고리를 추천하는 방법으로 협력적 필터링 방법[13,14]이 제안되었다. 이러한 접근방법은 모바일 애플리케이션 설명 정보가 유용함에도 불구하고 활용하지 못한다는 한계가 있다[12].

본 논문에서는 단어 군집 정보를 활용하여 모바일 애플리케이션을 범주화하는 방법에 대해 살펴본다. 모바일 애플리케이션의 세분화된 범주를 바탕으로 단어 발생 빈도를 분석하여 단어군집을 생성한다. 그리고, 모바일 애플리케이션 설명이 지나치게 짧아서 범주화에 필요한 정보를 얻기 어려울 수 있다는 점을 고려하여, 제안하는 방법은 단어 군집 정보를 추가하여 모바일 애플리케이션 설명의 텍스트 정보량을 증가시킨다. 이때, 모바일 애플리케이션 설명에 설치 사양이나 지원언어처럼 범주 분류에 관련 없는 내용이 있을 수 있다는 점을 고려하여, 모바일 애플리케이션 범주화에 유용한 단어 군집을 일부 선별하여 활용한다.

본 논문은 다음과 같이 구성된다. 앞으로, 2장에서는 단어 군집화 기반 모바일 애플리케이션 범주화 방법을 설명하고, 3장에서는 실험을 통해 제안하는 방법의 성능을 평가한다. 마지막으로 4장에서는 결론 및 향후 연구에 대해 기술한다.

II. 단어 군집 기반 문서 분류

모바일 애플리케이션의 제목이나 설명처럼 텍스트 정보가 주어지면 제안하는 방법은 [그림2]와 같이 단어 자질 추출 단계, 단어 군집 대표 자질 추가 단계, 범주화 단계를 통해 모바일 애플리케이션의 범주를 분류한다. 이러한 과정에 활용되는 단어 군집 정보와 모바일 애플리케이션 범주화 학습 결과는 범주 체계에 따라 분류된 모바일 애플리케이션으로 구성된 학습집합을 바탕으로 단어 군집화 단계와 범주화 학습 단계를 통해 미리 생성한다. 단어 자질 추출, 단어 군집 대표 자질 추가, 범주화, 범주화 학습, 단어 군집화의 각 단계에 대해 자세히 살펴보면 다음과 같다.

단어 자질 추출 단계에서는 모바일 애플리케이션 텍스트 정보를 형태소분석[15]하고, 명사, 동사, 영어에 해당하는 단어를 바탕으로 unigram, bigram, trigram을 추출한다. 예를 들어, [표1]의 모바일 애플리케이션 "도쿄지하철"에서 단어 자질 리스트 "도쿄, 지하철, 도쿄지하철, 일본, 여행, 일본여행, 한국어, 영어, 중국어, 간체, 중국어, 번체, 프랑스어, 독일어, 스페인어, 카자흐스탄어, 지원, 필요"를 추출할 수 있다. 이와 같이, 단일명사는 unigram만 추출하지만, "음성인식 번역기"와 같은 복합명사는 "음성", "인식", "번역기"의 unigram 뿐만 아니라, "음성인식", "인식번역기"의 bigram과 "음성인식번역기"의 trigram을 모두 추출한다. 따라서, 단일명사인 "음성"이나 "인식"에 비해서 모바일 애플리케이션을 분류하는데 좀 더 유용한 정보를 많이 포함하는 복합명사 "음성인식번역기"도 함께 활용할 수 있다.

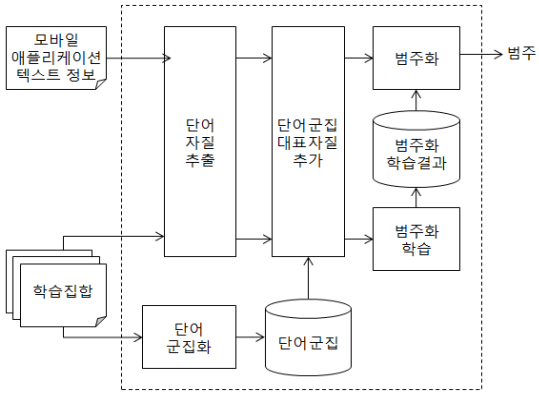

그림 2. 단어 군집 기반 문서 분류
Fig. 2. Word Cluster-based Document Classification

단어 군집 대표 자질 추가 단계에서는 모바일 애플리케이션 텍스트 정보에서 추출된 단어 자질중에서 단어 군집에 포함된 단어 자질이 있는 경우 해당되는 단어 군집 대표 정보를 범주화 자질로 추가한다. 예를 들어, [표1]의 모바일 애플리케이션 "도쿄지하철"은 텍스트 정보에 "지하철"과 "여행"이라는 단어를 포함하며, 이 단어는 [표2]의 단어 군집에서 "C3"과 "C4"에 각각 해당한다. 따라서, 모바일 애플리케이션 "도쿄지하철"은 해당 단어군집을 나타하는 "C3"과 "C4"를 단어 군집 대표 자질로 추가한다. 결국, 범주화에 활용되는 자질은 "도쿄, 지하철, 도쿄지하철, 일본, 여행, 일본여행, 한국어, 영어, 중국어, 간체, 중국어, 번체, 프랑스어, 독일어, 스페인어, 카자흐스탄어, 지원, 필요, C3, C4"가 된다.

범주화 단계는 각 자질별 발생빈도를 기준으로 전체 단어 자질 및 군집 자질로 구성된 벡터를 생성하고, 이를 최대엔트

표 2. 단어 군집의 일부 예
Table 2. Some Word Cluster Examples

구분	포함 단어
C1	문제, 퍼즐, 숫자, 두뇌, 퍼즐게임, 맞고 …
C2	판매, 회원, 이벤트, 인증, 쿠폰, 코드, …
C3	지역, 경로, 찾기, 지하철, 운전, 차량, …
C4	위치, 여행, 주변, 지도, GPS, …

로피 모델[16,17]에 적용하여 모바일 애플리케이션의 범주를 분류한다[7]. 모바일 애플리케이션을 나타내는 벡터 $\vec{d}$를 범주 c_i로 분류하는 조건부 확률식은 수식(1)과 같이 정의된다. 이때 $f_j(\vec{d},c_i)$는 j번째 자질과 관련된 자질함수이고, λ_j는 자질함수 $f_j(\vec{d},c_i)$의 가중치이며, k는 자질 개수이고, $Z(\vec{d})$는 $\Sigma_c P(c_i|\vec{d}) = 1$을 만족시키는 상수값을 나타낸다[16,17].

$$P(c_i \mid \vec{d}) = \frac{1}{Z(\vec{d})} \exp\left(\sum_{j=1}^{k} \lambda_j f_j(\vec{d},c_i) \right) \quad (1)$$

범주화 학습 단계에서는 모바일 애플리케이션의 텍스트 정보와 이에 대응하는 범주의 쌍으로 구성된 학습집합에서 범주화 단계에서 사용할 확률 분포를 학습한다. 이때, 학습집합에 포함된 모바일 애플리케이션과 그 범주의 쌍은 이미 실세계에 존재하는 자료이므로, 이들이 실세계에서 나타날 확률이 최대가 되도록 수식(1)에서 자질함수 $f_j(\vec{d},c_i)$, 자질함수 가중치 λ_j, 상수값 $Z(\vec{d})$를 학습한다[16].

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{w_j \in S_i} | w_j - \mu_i |^2 \quad (2)$$

단어 군집화 단계에서는 K 평균 군집화 알고리즘(K-means Clustering Algorithm)[18]을 이용하여 유사한 단어끼리 하나의 그룹으로 묶어서 [표2]와 같이 단어 군집을 생성한다. 이를 위해, 단어 w_j는 각 범주에서 그 단어가 몇 번 나타났는지를 바탕으로 벡터로 표현한다. 그리고, 각 단어 w_j를 유사한 단어끼리 묶어 $S=\{S_1, S_2, \ldots, S_k\}$와 같이 k개의 단어 군집으로 나누기 위해서, K 평균 군집화 알고리즘은 k개의 중심점 μ_i를 설정하고 각 단어에 대해 가장 가까운 중심점 μ_i를 찾아 군집을 할당한다. 할당된 결과를 바탕으로 각 군집에 있는 단어들의 평균값으로 중심점 μ_i를 재조정한다. 이와 같이 각 단어 w_j의 단어 군집을 설정하는 단계와 각 단어 군집의 중심점 μ_i을 재조정하는 단계를 반복하면서 전체 분산을 최소화하는 단어군집을 찾는다[18].

III. 실험 및 평가

제안하는 방법으로 생성된 단어 군집 결과가 모바일 애플리케이션 범주화 문제에서 유용하게 활용될 수 있는지를 살펴보기 위해서, 모바일 애플리케이션 스토어에서 25개의 범주로 분류된 모바일 애플리케이션 텍스트 정보 3,521개를 수집

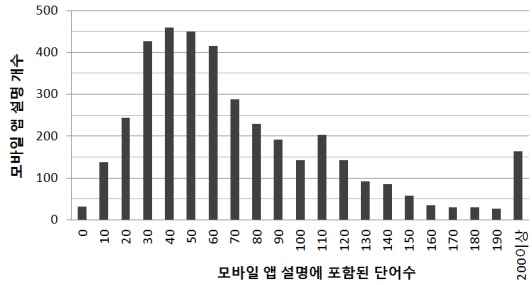

그림 3. 길이별 문서 분포
Fig. 3. Document Distribution by Length

하여 평가 말뭉치를 구축하였다[7]. 이 평가 말뭉치를 학습집합 90%와 실험 집합 10%로 구분하여, MALLET 문서 분류기[19]에 적용하였다.

평가 말뭉치의 범주별 모바일 애플리케이션 분포는 [표3]과 같이 전체집합에서 반 이상의 모바일 애플리케이션이 유틸리티, 퍼즐/보드게임, 건강/생활안전, 일정/메모, 아케이드게임 등의 특정 범주에 집중되어 있다. 특히, 유틸리티 범주는 전체 집합의 21%의 모바일 애플리케이션을 포함한다[7]. 한편, 평가 말뭉치에서 텍스트 정보의 길이별 모바일 애플리케이션 분포는 [그림3]과 같이 대부분 50단어 내외로 작성되어 있고, "[제목]긴급싸이렌, [설명]긴급싸이렌"과 같이 9단어 이하로 짧게 작성된 모바일 애플리케이션 설명도 31개가 있다.

제안하는 단어 군집 기반 모바일 애플리케이션 범주화 방법이 얼마나 정확하게 모바일 애플리케이션을 분류하는지 평가하기 위해서, 수식 (1)과 같이 제안하는 방법의 정확률을 측정한다. 그리고, 제안하는 방법이 전체 모바일 애플리케이션에 대해 올바르게 범주화한 경우가 얼마나 많은지를 평가하기 위

표 3. 범주별 모바일 애플리케이션 수
Table 3. Number of Mobile Applications per Categories

범주	문서수	범주	문서수
아케이드 게임	214	유틸리티	745
액션/슈팅 게임	106	건강/생활안전	276
시뮬레이션 게임	80	오디오	100
스포츠게임	89	뮤직Apps	10
퍼즐/보드 게임	357	스포츠/연예	88
RPG 게임	111	쇼핑	64
비즈니스/금융	168	사진	72
여행/지도/교통	186	음식	47
정보검색/상식	105	SNS	61
방송/영화	28	유형테스트	53
뉴스/날씨	84	운세	88
교육	99	종교	31
일정/메모	259	총계	3,521

표 4. 단어 군집을 활용한 범주별 성능 개선
Table 4. Performance Improvement per Categories by Utilizing Word Clusters

범주	단어 군집 정보 배제			단어 군집 정보 활용			개선폭		
	정확율	재현율	f-값	정확률	재현율	f-값	정확률	재현율	f-값
아케이드 게임	33.73	13.27	19.05	34.34	16.75	22.52	0.61	3.48	3.47
액션/슈팅 게임	56.36	28.18	37.58	58.93	30.56	40.24	2.56	2.37	2.67
시뮬레이션 게임	36.36	13.95	20.17	35.00	8.33	13.46	-1.36	-5.62	**-6.71**
스포츠게임	68.75	32.04	43.71	69.05	29.90	41.73	0.30	-2.14	-1.98
퍼즐/보드 게임	21.42	76.38	33.46	44.15	77.75	56.32	**22.73**	1.36	**22.86**
RPG 게임	83.16	74.53	78.61	86.52	73.33	79.38	3.36	-1.19	0.77
비즈니스/금융	54.46	42.07	47.47	58.77	45.27	51.15	4.31	3.20	3.67
여행/지도/교통	56.43	43.65	49.22	52.38	37.29	43.56	-4.05	-6.36	-5.66
정보검색/상식	47.22	17.53	25.56	36.17	16.83	22.97	-11.05	-0.69	-2.59
방송/영화	57.89	36.67	44.90	71.43	27.78	40.00	13.53	**-8.89**	-4.90
뉴스/날씨	32.08	20.48	25.00	36.84	28.38	32.06	4.77	7.90	7.06
교육	60.00	18.00	27.69	42.86	15.96	23.26	**-17.14**	-2.04	-4.44
일정/메모	71.96	49.82	58.87	70.00	46.49	55.88	-1.96	-3.32	-3.00
유틸리티	46.89	56.37	51.19	40.28	72.86	51.88	-6.61	**16.49**	0.69
건강/생활안전	55.14	48.91	51.84	54.61	55.17	54.89	-0.54	6.27	3.05
오디오	52.63	18.87	27.78	49.02	22.94	31.25	-3.61	4.07	3.47
뮤직Apps	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
스포츠/연예	67.74	53.85	60.00	72.31	54.65	62.25	4.57	0.81	2.25
쇼핑	78.95	22.06	34.48	70.00	22.58	34.15	-8.95	0.52	-0.34
사진	19.05	5.33	8.33	30.00	10.91	16.00	10.95	5.58	7.67
음식	53.57	31.91	40.00	65.22	31.91	42.86	11.65	0.00	2.86
SNS	56.52	21.67	31.33	76.92	31.25	44.44	20.40	9.58	13.12
유형테스트	8.33	1.92	3.12	27.27	6.00	9.84	18.94	4.08	6.71
운세	42.31	11.22	17.74	45.00	21.69	29.27	2.69	10.46	11.53
종교	75.00	46.15	57.14	61.11	44.00	51.16	-13.89	-2.15	-5.98
총계	42.05	42.05	42.05	47.70	47.70	47.70	5.65	5.65	**5.65**

해, 수식 (6)과 같이 재현율을 측정하였다. 수식 (7)는 정확률과 재현율의 조화평균인 f-값을 나타낸다.

$$\text{정확률} = \frac{\text{올바르게 자동분류한 문서수}}{\text{자동분류한 문서수}} \quad (5)$$

$$\text{재현율} = \frac{\text{올바르게 자동분류한 문서수}}{\text{실험집합에 포함된 전체 문서수}} \quad (6)$$

$$f-\text{값} = \frac{2 \times \text{정확률} \times \text{재현율}}{\text{정확률} + \text{재현율}} \quad (7)$$

〔표4〕는 모바일 애플리케이션 범주화의 성능이 단어 군집 정보를 활용하여 전체적으로 5.65% 개선되었다는 것을 보여준다. 범주별로 자세히 살펴보면, 일부 범주는 단어 군집 정보의 도움을 받아 성능이 개선되었지만, 일부 범주는 오히려 성능이 떨어진 경우도 있었다.

〔표3〕에 제시된 바와 같이 퍼즐/보드 게임은 게임 범주 중에서 가장 많은 357개의 모바일 애플리케이션을 포함하고 있으므로, 특정 범주로 분류하기 애매한 모바일 애플리케이션이 퍼즐/보드 게임 범주로 분류되는 경향이 있었다. 단어 군집 정보를 바탕으로 부족한 텍스트 정보를 보완하여, 퍼즐/보드 게임 범주로 분류되는 애매한 모바일 애플리케이션의 수가 줄어들면서, 퍼즐/보드 게임 범주의 정확률이 22.73%까지 향상되었다. 반면에, 〔표3〕에서 모바일 애플리케이션의 수가 상대적으로 적은 시뮬레이션 게임이나 종교 범주의 경우 단어 군집 정보를 활용하여 오히려 성능이 떨어졌는데, 이는 해당 범주의 학습집합의 크기가 80개와 31개로 크지 않아서 적절한 단어 군집을 생성하기 어려웠다는 것을 보여준다.

모바일 애플리케이션 분야의 특징과 전통적으로 범주화 실험에 많이 사용되는 뉴스기사 분야의 특징을 비교하기 위해서, 〔표5〕와 같이 뉴스 기사 13,129개를 수집하여 평가 말뭉치를 구축하고 이를 바탕으로 실험하였다. 모바일 애플리케이션 평가말뭉치는 범주가 25개로 세분화되어 있는 반면에, 신문기사 평가말뭉치는 범주가 10개로 단순하며 평가말뭉치의 크기도 약 3.7배 크다. 따라서, 모바일 애플리케이션 평가말뭉

표 5. 범주별 뉴스기사 수
Table 4. Number of News Articles per Categories

범주	문서수	범주	문서수
정치	460	스포츠	1,874
경제	1,100	게임	1,750
사회	2,783	청소년	325
법	300	여행레져	1,357
해외	659		
연예	2,521	총계	13,129

치에서 단어 군집 정보를 전혀 활용하지 않는 경우 〔그림4〕과 같이 42.05%로 상대적으로 낮은 재현율을 보이지만 뉴스기사 평가말뭉치에서는 76.64%의 재현율을 보인다.

범주화에 활용되는 단어 군집의 개수에 따라서 성능이 어떻게 변하는지를 살펴보기 위해서, 단어 군집의 가능한 조합을 모두 평가하고 가장 높은 재현율을 보이는 조합을 선택하여 그 결과를 〔그림4〕에 제시하였는데, 그 결과 단어 군집을 많이 사용한다고 해서 성능이 항상 향상되지는 않았다. 이는 품질이 좋은 단어 군집을 추가할 경우 성능이 올라갈 수 있지만, 품질이 다소 떨어지는 군집을 사용할 경우 재현율이 오히려 떨어질 수 있다는 것을 나타낸다.

제안하는 방법은 〔표6〕에 제시된 바와 같이 모바일 애플리케이션 평가 말뭉치에서 단어 군집 4개를 활용하여 범주화할 때 재현율이 44.70%로 가장 높았고, 이는 단어 군집 정보를 배제한 방법의 재현율 42.05%에 비해 5.65% 높다. 그리고, 신문 기사 평가 말뭉치에서는 단어 군집 1개를 활용하여 범주화할 때 재현율이 80.23%로 가장 높았고, 이는 단어 군집 정보를 배제한 방법의 재현율 76.64%에 비해 3.59% 높다. 즉, 모바일 애플리케이션 범주화 방법이 신문 기사 범주화 방법에 비해 더 많은 단어 군집을 활용하고 그 결과 성능 개선폭도 더 크다. 이는 모바일 애플리케이션 설명이 뉴스기사에

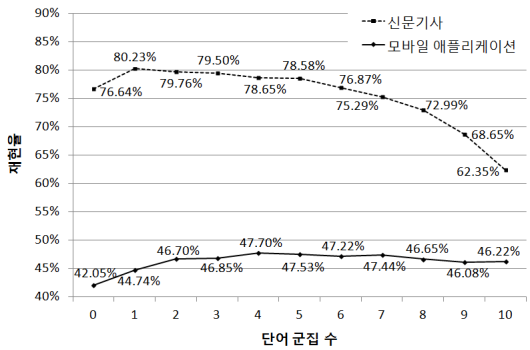

그림 4. 단어 군집 개수에 따른 성능 변화
Fig. 4. Performance Variation by Number of Word Clusters

표 6. 단어 군집을 활용한 성능 개선
Table 1. Performance Improvement by Utilizing Word Clusters

	단어 군집 배제	단어 군집 활용	개선폭
모바일 애플리케이션	42.05%	47.70%	5.65%
신문기사	76.64%	80.23%	3.59%

비해 짧고 범주 분류와 무관한 내용을 포함하고 있어서 텍스트 정보가 매우 부족한데, 이를 단어 군집 정보가 보완할 수 있다는 것을 나타낸다.

IV. 결 론

본 논문에서는 단어 군집화 정보를 활용하여 모바일 애플리케이션을 범주화하는 방법을 제안한다. 제안하는 방법은 단어 자질 추출 단계, 단어 군집 대표 자질 추가 단계, 범주화 단계를 통해 모바일 애플리케이션의 범주를 분류한다. 이 때, 범주화에 필요한 단어 군집 정보와 범주화 확률분포는 단어 군집화 단계와 범주화 학습 단계를 통해 학습집합에서 미리 생성한다. 제안하는 방법은 다음과 같은 특징이 있다.

첫째, 제안하는 방법은 단어 군집 정보를 추가하여 모바일 애플리케이션의 텍스트 정보량을 증가시키므로, 모바일 애플리케이션의 부족한 텍스트 정보를 보완한다. 실험결과 단어군집 정보를 활용하지 않았을 때에 비해 단어 군집 정보를 추가하였을 경우 모바일 애플리케이션 범주화 재현율을 5.65% 개선하였다.

둘째, 단어 군집은 모바일 애플리케이션 설명의 세분화된 범주를 바탕으로 생성한다. 즉, 단어 w_j는 각 범주에서 그 단어가 몇 번 나타났는지를 바탕으로 벡터로 표현하고, 이를 K 평균 군집화 알고리즘에 적용하므로, 범주 분류와 관련성이 높은 단어를 중심으로 단어군집을 생성할 수 있다.

셋째, 제안하는 방법은 모바일 애플리케이션의 범주 분류에 유용한 일부 단어 군집만 선별하여 사용한다. 이와 같이, 제안하는 방법은 부적절하게 생성된 단어 군집을 배제할 수 있으므로, 지원언어나 설치사양과 같이 범주 분류와 관련이 적은 내용의 영향력을 다소 감소시킬 수 있다.

참고문헌

〔1〕 B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," Journal of Advances in Information Technology, Vol. 1, No. 1, pp. 4-20, Feb. 2010.

[2] J. P. Moon, W. S. Lee, J. H. Chang, "A proper folder recommendation technique using frequent itemsets for efficient e-mail classification," Journal of the Korea Society of Computer and Information, Vol. 16, No. 2, pp. 33-46, Feb. 2011.

[3] Y. S. Hwang, J. C. Moon, S. J. Cho, "Classification of malicious Web pages by using SVM," Journal of the Korea Society of Computer and Information, Vol. 17, No. 3, pp. 77-83, Mar. 2012.

[4] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," Machine Learning, Vol. 88, No. 1-2, pp. 157-208, Dec. 2011.

[5] S. Y. Park, J. Chang, and T. Kihl, "Document classification model using Web documents for balancing training corpus size per category," Journal of Information and Communication Convergence Engineering, Vol. 11, No. 4, Dec. 2013.

[6] G. Lu, P. Huang, L. He, C. Cu, and X. Li, "A new semantic similarity measuring method based on Web search engines," WSEAS Transactions on Computers, Vol. 9, No. 1, pp. 1-10, Jan. 2010.

[7] B. K. Sun, D. H. We, K. R. Han, "A Study on Paper Retrieval System based on OWL Ontology," Journal of the Korea Society of Computer and Information, Vol. 14, No. 2, pp. 169-180, Feb. 2009.

[8] S. Samarawickrama, and L. Jayaratne, "Automatic text classification and focused crawling," in Proceeding of the 6th International Conference on Digital Information Management, Melbourne, Australia, pp. 143-148, Sept. 2011.

[9] de Groc, C. "Babouk: focused web crawling for corpus compilation and automatic terminology extraction," In Proceeding of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 497-498, Aug. 2011.

[10] H. Liu, and E. Milios, "Probabilistic models for focused Web crawling," Computational Intelligence, Vol. 28, No. 3, pp. 289-328, Aug. 2012.

[11] Y. R. Lee, and E. G. Im, "A Study on the smart phone application clustering using information of android permissions," in Proceeding of the Conference on the Korean Institute of Communication Science, pp. 812-813, Feb. 2012.

[12] H. G. Yoon, S. Kim, and S. B. Park, "Noise elimination in mobile app descriptions based on topic model," in Proceeding of the Conference on Human & Cognitive Language Technology, pp.64-68, Oct. 2013.

[13] W. H. Rho, S. B. Cho, "A mobile app category recommendation system with contexts using bayesian network," in Proceeding of Korea Computer Congress, pp.1408-1410, Jun. 2013.

[14] B. Yan, and G. Chen, "AppJoy: personalized mobile application discovery," in Proceedings of the 9th international conference on mobile systems, applications, and services, pp. 113-126, Jun. 2011.

[15] S. Z. Lee, J. I. Tsujii, and H. C. Rim, "Hidden Markov model-based Korean part-of-speech tagging considering high agglutinativity, word-spacing, and lexical correlativity," in Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp. 384-391, Oct. 2000.

[16] A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra, "A maximum entropy approach to natural language processing," Computational Linguistics, vol. 22, no. 1, pp. 39-71, Mar. 1996.

[17] J. H. Lim, Y. S. Hwang, S. Y. Park, and H. C. Rim, "Semantic role labeling using maximum entropy model," in Proceeding of the Conference on Computational Natural Language Learning, Boston: MA, pp. 122-125, May. 2004.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P.

Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, "Scikit-learn: machine learning in python," Journal of Machine Learning Research, Vol. 12, pp. 2825-2830, Oct. 2011.

[19] A. K. McCallum, MALLET: a machine learning for lan-guage toolkit, http://mallet.cs.umass.edu.

저 자 소 개

허 정 만

2013: 상명대학교
디지털미디어학부 이학사.
현 재: 상명대학교
게임학과 석사과정 재학중
관심분야: 컴퓨터과학
Email : vngofgof@naver.com

박 소 영

1997: 상명대학교
전자계산학과 이학사.
1999: 고려대학교
컴퓨터과학과 이학석사.
2005: 고려대학교
컴퓨터과학과 이학박사
현 재: 상명대학교 게임학과 부교수
관심분야: 지식정보처리
Email : ssoya@smu.ac.kr