Journal of The Korea Society of Computer and Information
Vol. 20, No. 4, April 2015
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.4.017

모바일 앱 트렌드를 고려한 2단계 군집화 방법

허 정 만*, 박 소 영*

Two-Phase Clustering Method Considering Mobile App Trends

Jeong-Man Heo*, So-Young Park*

요 약

본 논문에서는 단어 군집을 사용하여 모바일 앱을 군집화하는 방법을 제안한다. 모바일 앱 트렌드의 빠른 변화를 고려하여, 제안하는 방법은 미리 정의된 분류체계를 사용하지 않고, 모바일 앱 집합에 군집화 기술을 적용하여 의미적으로 유사한 모바일 앱을 묶는다. 짧은 모바일 앱 소개 글의 자료 부족 문제를 완화하기 위해서, 각 단어에 대해 unigram 뿐만 아니라, bigram, trigram, 단어 군집 정보를 추가적으로 확보하여 활용한다. 모바일 앱을 전체적으로 정확하게 군집화하기 위해서, 제안하는 방법은 단어 군집을 활용하여 모바일 앱 군집의 크기가 지나치게 작거나 크지 않도록 관리한다. 실험결과 제안하는 방법은 단어 군집을 활용하여 전체 정확도를 57.48%에서 79.66%로 22.18% 개선시켰다.

▸ Keywords : 모바일 앱 군집, 단어 군집, 군집화방법, 텍스트 분석

Abstract

In this paper, we propose a mobile app clustering method using word clusters. Considering the quick change of mobile app trends, the proposed method divides the mobile apps into some semantically similar mobile apps by applying a clustering algorithm to the mobile app set, rather than the predefined category system. In order to alleviate the data sparseness problem in the short mobile app description texts, the proposed method additionally utilizes the unigram, the bigram, the trigram, the cluster of each word. For the purpose of accurately clustering mobile apps, the proposed method manages to avoid exceedingly small or large mobile app clusters by using the word clusters. Experimental results show that the proposed method improves 22.18% from 57.48% to 79.66% on overall accuracy by using the word clusters.

▸ Keywords : Mobile App Clustering, Word Clustering, Clustering Algorithm, Text Analysis

•제1저자 : 허정만 •교신저자 : 박소영
•투고일 : 2014. 11. 13, 심사일 : 2014. 12. 31, 게재확정일 : 2015. 3. 16.
* 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

모바일 생태계를 바꾸어 놓은 스마트폰의 사용이 크게 증가하면서 IT 비즈니스 패러다임이 PC기반에서 모바일 기반으로 급격히 전환되고 있다[1]. 스마트폰 이용자 중 6.1%가 최근 1개월 이내 모바일 앱을 다운로드 받았고, 이들 중 65.9%가 다운로드 받은 앱을 이용한 적이 있으며, 그 중에서도 17.8%는 '하루에도 여러 번' 이용하는 것으로 나타났다[2]. 방대한 양의 다양한 모바일 앱 중에서 사용자가 원하는 서비스를 제공하는 모바일 앱을 찾는 것은 쉽지 않을 수 있다.

따라서, 사용자가 원하는 모바일 앱을 쉽게 찾을 수 있도록 분류체계를 체계적으로 정의하고, 모바일 앱의 소개글을 분석하여 자동으로 분류하는 방법이 제안되었다[3,4]. 이러한 방법을 활용하여 [표1]의 모바일 앱에 대해 '뉴스/날씨'나 '여행/지도/교통'과 같은 분류체계를 기준으로 자동으로 분류할 수 있다. 그러나, 미리 정의된 분류체계를 바탕으로 모바일 앱을 분류하는 경우, 의미적으로 유사한 앱이 서로 다른 범주로 분류 될 수 있다. 예를 들어, [표1]에서 모바일 앱 '(b)바다날씨'는 '뉴스/날씨'의 같은 범주에 속하는 '(a)신문뉴스'보다 '(c)바다낚시가이드'와 내용상 더 유사하다고 할 수 있다. 하지만, (a)와 (c)가 분류체계상으로 다른 범주에 속하므로, 이를 효과적으로 표현하기가 어렵다. 따라서, 모바일 앱

표 1. 분류된 모바일 앱 예제
Table 1. Classified Mobile App Samples

구분	설명
(a)	 신문·뉴스 뉴스/날씨 | 한국신문,한국뉴스,연예뉴스,뉴스속보,스포츠신문,경제신문,해외신문 판매회원:cyebiz | 다운로드:7,049 | 업데이트:2014.07.08 유사상품 더보기
	[뉴스/날씨] 신문·뉴스: 한국 신문, 한국 뉴스, 연예 뉴스, 뉴스 속보, 스포츠 신문, 경제 신문, 해외 신문
(b)	 바다날씨(파고,낚시,태풍,기상청) 뉴스/날씨 | 파고 및 바다날씨를 제공하여 바다낚시에 유용한 정보를 제공. 판매회원:TEAM HJ | 다운로드:4,308 | 업데이트:2013.11.13 유사상품 더보기
	[뉴스/날씨] 바다날씨(파고,낚시,태풍,기상청): 파고 및 바다날씨를 제공하여 바다낚시에 유용한 정보를 제공.
(c)	 바다낚시 가이드 여행/지도/교통 | 바다낚시 가이드 (날씨/조수/수온 보기, 출조일/추천장소/채비 검색) 판매회원:koinlab | 다운로드:12,871 | 업데이트:2012.01.13 유사상품 더보기
	[여행/지도/교통] 바다낚시가이드: 바다낚시 가이드 (날씨/조수/수온 보기, 출조일/추천장소/채비 검색)

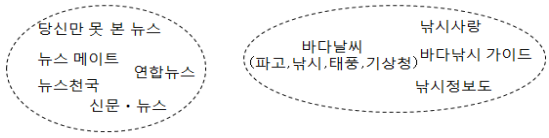

그림 1. 모바일 앱 군집 예시
Figure 1. Some Mobile App Cluster Examples

의 트렌드 흐름에 적절히 대응하고 그 특징을 효과적으로 분석하기 위해서는 [표1]과 같이 미리 정의된 분류체계에 모바일 앱을 분류하는 방식보다는 [그림1]과 같이 모바일 앱 집합에서 의미적으로 서로 밀접하게 관련이 있는 유사한 모바일 앱끼리 묶어주는 모바일 앱 군집화 방식이 필요하다.

따라서, 본 논문에서는 모바일 앱 집합에서 단어 군집 정보 추출하고 그 결과를 모바일 앱 군집화에 활용하는 2단계 군집화 방법을 제안한다. 본 논문은 다음과 같이 구성된다. 2장에서는 문서 분류 및 문서 군집화와 관련된 연구를 살펴본다. 3장에서는 제안하는 모바일 앱 2단계 군집화 방법을 설명하고, 4장에서는 실험을 통해 제안하는 방법의 성능을 평가한다. 마지막으로 5장에서는 결론 및 향후 연구에 대해 기술한다.

II. 관련연구

그동안 끊임없이 생겨나는 엄청난 양의 다양한 비정형 문서들을 체계적으로 정의된 분류체계로 자동으로 분류하기 위해서 다양한 문서 분류 기술이 연구되어 왔다. 이를 위해, 문서 분류 기술은 일반적으로 문서에서 분류에 필요한 자질 정보를 추출하는 단계와 이러한 자질을 바탕으로 분류하는 단계로 구성된다[4,5]. 먼저, 자질 정보 추출단계에서는 문서빈도, 정보이득률(Information Gain), tf-idf(Term Frequency-Inverse Document Frequency), 상호정보(Mutual Information), 카이제곱 통계량(Chi-square Statistics) 등을 계산하여 자질을 추출한다[5-8]. 그리고, 분류단계에서는 이러한 자질 정보를 신경망, 결정트리, Naive-Baysian, SVM 등의 분류 학습 방법에 적용하여 분류한다[9-11]. 그러나, 이러한 지도 학습 방법은 분류 성능을 높이기 위해서 정답이 부착된 학습집합이 충분히 확보되어야 한다. 학습집합의 구축비용을 줄이기 위해서, 최근에는 비지도 학습방법을 활용한 문서 분류 방법[12,13]도 제안되고 있다. 그럼에도 불구하고, 문서 분류 기술은 미리 정의한 분류체계를 사용하므로, 문서를 자동으로 분류하는 도중에 분류체계를 수정하는 것이 불가능하다. 따라서, 급격하게 변하는 모바일 앱의 트렌드 흐름에 적절히 대응할 수 없다는 한계가 있다.

반면에, 문서 군집화 기술은 미리 정의된 분류체계나 학습 집합을 사용하지 않고 문서에 포함된 단어를 기준으로 유사도를 계산해서 의미적으로 유사한 문서를 군집화한다. 주어진 문서집합에 따라서 군집화 결과도 달라질 수 있으므로, 최신 트렌드의 흐름에 적절히 대응할 수 있다. 이러한 군집화 방법은 크게 계층적 군집화 방법[14]과 비계층적 군집화 방법[15,16]으로 나누어 볼 수 있다. 계층적 군집화 기법은 매단계마다 가장 유사한 문서 쌍을 선택하여 군집을 형성하므로 정확하지만 속도가 느리다. 최근 군집화 속도를 개선하기 위해서 단어 대신 개체명을 활용하는 방법[17-19]이 제안되었다. 반면, 비계층적 군집화는 속도는 빠르지만 검색효율이 떨어지고 문서의 입력 순서에 따라 군집화의 결과가 달라진다는 단점이 있다. 한편, 새로운 문서의 추가나 기존 문서의 삭제로 인하여 문서집합이 변화하는 환경을 고려한 군집화 방법[20,21]도 제안되었다.

이렇게 제안된 군집화 방법은 단어를 충분히 포함하고 있는 신문기사 등을 잘 분류할 수 있도록 설계되었다. 그러나, 모바일 앱은 글보다는 그림으로 소개하려는 경향이 강해서 텍스트 정보의 길이가 다소 짧은 경향이 있다[4]. 게다가, 모바일 앱 텍스트 정보는 군집화에 노이즈로 작용하는 설치사양, 지원언어, 공지사항, 광고글, 사용법 등을 포함한다[4,22]. 이와 같이 모바일 앱 텍스트 정보는 충분한 단어 정보를 포함하고 있지 않기 때문에, 기존 문서 군집화 방법을 그대로 활용하면 자료부족문제가 매우 심각하게 나타날 수 있다. 이러한 점을 고려하여 제안하는 방법은 모바일 앱 집합에서 단어 군집 정보를 추출하고 그 결과를 모바일 앱 군집화에 활용하는 2단계 군집화 방법을 제안한다. 제안하는 방법은 단어 군집 정보를 활용하여 기존 단어 정보를 확장하므로, 모바일 앱의 단어 정보가 부족하다는 점을 보완할 수 있다.

III. 2단계 모바일 앱 군집화 방법

제안하는 2단계 모바일 앱 군집화 방법은 [그림2]와 같이 단어 자질 획득 단계, 단어 군집화 단계, 모바일 앱 군집화 단계로 구성된다. 모바일 앱 텍스트 집합을 입력하면, 제안하는 방법은 모바일 앱 텍스트에서 얻은 단어 자질을 바탕으로 단어 군집을 생성하고, 그 군집결과를 활용하여 모바일 앱을 군집화한다. 모바일 앱 텍스트가 대체적으로 짧고 노이즈 정보가 많아서 유효한 단어 정보가 부족하다는 점을 개선하기 위해서, 제안하는 방법은 [표2]와 같이 각 단어의 unigram과 함께 bigram, trigram, 단어 군집 정보를 추가적으로 확보하여 활용함으로써 단어 정보를 증가시킨다. 이렇게 증가된

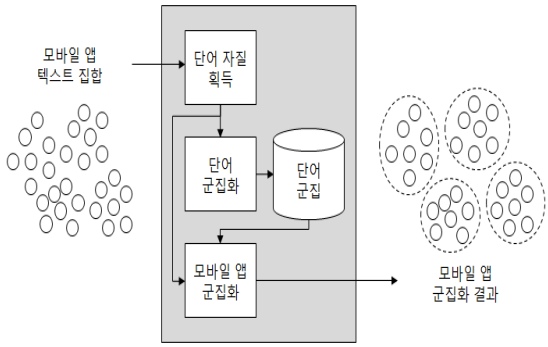

그림 2. 2단계 군집화 모델
Fig. 2. Two-Phase Clustering Model

단어 정보를 바탕으로 모바일 앱간 유사도를 계산하므로 유사도를 좀 더 정확하게 계산할 수 있다.

먼저, 단어 자질 추출 단계에서는 모바일 앱을 소개하는 텍스트 정보를 형태소 분석[23]하고, 품사가 명사, 동사, 영어에 해당하는 단어를 단어 자질로 선택하는데, 복합명사처럼 명사가 연달아 나타나면 unigram뿐만 아니라 bigram과 trigram도 함께 추출한다[4]. 예를 들어, 복합명사 "바다 낚시 가이드"에서 unigram "바다", "낚시", "가이드"과 함께 bigram "바다낚시", "낚시가이드", trigram "바다낚시가이드"를 추출한다. 모바일 앱 "바다낚시가이드"의 특징은 "바다"나 "가이드"보다 "바다낚시"이나 "낚시가이드"가 더 잘 표현한다고 할 수 있다. 따라서 제안하는 방법은 이와 같이 unigram, bigram, trigram을 모두 군집화에 활용하여 부족한 단어 정보를 보완한다.

표 2. 텍스트 정보 확장 예
Table 2. Some Text Information Expansion Examples

구분	unigram	bigram/trigram	단어군집
(a)	신문 뉴스 한국 신문 한국 뉴스 연예 뉴스 뉴스 속보 스포츠 신문 경제 신문 해외 신문	한국신문 한국뉴스 연예뉴스 뉴스속보 스포츠신문 경제신문 해외신문	W1 W1 W1 W1 W1 W1 W1 W1 W1 W1
(b)	바다 날씨 파고 낚시 태풍 기상청 파고 바다 날씨 제공 바다 낚시 유용 정보 제공	바다날씨 바다날씨 바다낚시	W2 W3 W2 W2 W2 W3 W3
(c)	바다 낚시 가이드 바다 낚시 가이드 날씨 조수 수온 보기 출조일 추천 장소 채비 검색	바다낚시 바다낚시가이드 바다낚시 바다낚시가이드 수온보기 추천장소 채비검색	W3 W3 W2 W2 W3 W3 W3

표 3 단어 군집 예시
Table 3. Some Word Cluster Examples

군집	포함 단어
W1	신문, 뉴스, 속보, 연예뉴스, 스포츠신문, …
W2	기상청, 날씨, 태풍, 수온, 온도, 기온, 파고, …
W3	낚시, 바다낚시, 낚시터, 조행기, 출조일, …

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

한편, 단어 군집화 단계 및 모바일 앱 군집화 단계에서는 수식(1)과 같이 K 평균 군집화 알고리즘(K-means Clustering Algorithm)[24]을 활용하여, [표3]과 같은 단어 군집과 [그림1]과 같은 모바일 앱 군집을 생성한다. 이를 위해, 단어 군집화 단계에서는 단어를 벡터 공간의 점 x_j로 표현하고, 모바일 앱 군집화 단계에서는 모바일 앱을 점 x_j로 표현한다. 생성할 군집의 개수를 n개라고 할 때, 군집 집합은 $S=\{S_1, S_2, \ldots, S_n\}$이며, i번째 군집 S_i의 중심을 μ_i로 설정하듯이 각 군집의 중심을 여러 점에서 선택하여 초기화한다. 군집 설정 단계에서는 각 점 x_j에 대해 n개의 중심 중에서 가장 거리가 가까운 μ_i를 선택하여 군집을 할당한다. 그리고, 군집 중심 재조정 단계에서는 각 점의 군집 할당 결과를 바탕으로 각 군집에 있는 단어들의 평균값으로 중심 μ_i를 재조정한다. 이와 같이 각 점 x_j의 군집을 설정하는 단계와 각 군집의 중심 μ_i을 재조정하는 단계를 반복하면서 전체분산을 최소화하는 방향으로 군집을 구한다. 이러한 과정이 반복되어 군집이 변하지 않는다면 반복을 중지한다[4,22]. 이러한 과정을 통해 [표3]과 같은 단어 군집 정보와 [그림1]과 같은 모바일 앱 군집 정보를 생성한다.

IV. 실험 및 평가

단어 군집 정보가 모바일 앱 군집화에 얼마나 기여하였는지를 살펴보기 위해서, 모바일 앱 군집화에 단어 군집 정보를 활용하는 방법과 그렇지 않은 방법으로 모바일 앱 텍스트 3,521개[3,4]에 대해 군집을 50개씩 생성하여 평가한다. 이때, 단어군집은 동일한 모바일 앱 텍스트에서 150개를 생성하여 활용한다. 군집결과에 대한 평가의 공정성을 고려하여, 군집결과를 수작업으로 평가하는 대신 기존에 분류된 25개의 범주 정보를 기준으로 모바일 앱 군집화 결과의 정확도를 평가하였다. 즉, 각 군집에서 가장 많이 포함된 범주를 기준으로, 각 모바일 앱이 해당 범주이면 그 군집에 맞는 모바일 앱으로 판단하고, 다른 범주이면 그 군집에 맞지 않는 모바일 앱으로 판단한다. 예를 들어, [그림1]의 오른쪽 군집은 "여행/지도/교통" 범주에 해당하는 모바일 앱이 다수를 차지하므로, "여행/지도/교통" 범주로 분류된 모바일 앱 "바다낚시가이드"는 해당 군집에 맞는 모바일 앱으로 판단할 수 있다. 반면에, 모바일 앱 "바다날씨"는 "뉴스/날씨" 범주로 분류되므로 해당군집에 맞지 않는다고 판단한다.

이러한 내용을 바탕으로, 각 군집의 정확도는 수식 (2)과 같이 군집에 포함된 모바일 앱중 맞게 군집화한 모바일 앱의 비율로 계산한다. 그리고, 전체 군집의 평균적인 정확도는 수식(3)과 같이 각 군집의 정확도에 대한 평균으로 계산한다[19]. 이때, 군집크기가 1인 경우는 군집정확도가 100%가 되므로, 크기가 작은 군집이 많을수록 평균정확도가 지나치게 높게 계산될 수 있다. 이를 보완하기 위해서, 수식(4)와 같이 전체 모바일 앱을 대상으로 각 군집에 맞는 모바일 앱의 비율을 전체 정확도로 계산한다.

$$\text{정확도}_i = \frac{i\text{번째 군집에 맞는 모바일 앱 수}}{i\text{번째 군집에 포함된 모바일 앱 수}} \quad (2)$$

$$\text{평균정확도} = \frac{\sum_{i=1}^{\text{군집수}} \text{정확도}_i}{\text{군집수}} \quad (3)$$

$$\text{전체정확도} = \frac{\sum_{i=1}^{\text{군집수}} i\text{번째 군집에 맞는 모바일 앱 수}}{\text{전체 모바일 앱 수}} \quad (4)$$

단어 군집이 모바일 앱 군집화에 어떤 영향을 주는지 살펴보기 위해서, 단어 군집의 가능한 조합을 모두 평가하고 전체정확도가 가장 높은 조합을 선택하여 평가하였다. [그림3]에 나타난 바와 같이, 단어 군집을 많이 활용한다고 해서 전체정확도가 계속 증가하지는 않는다. 이는 품질이 좋은 단어 군집을 활용하면 전체정확도를 높일 수 있지만, 품질이 다소 떨어

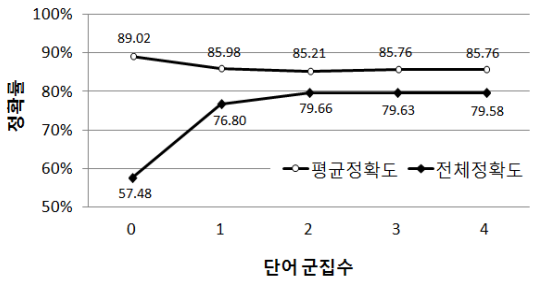

그림 3. 단어 군집 수에 따른 모바일 앱 군집 정확도
Fig. 3. Distribution of Accuracy per Number of Word Cluster

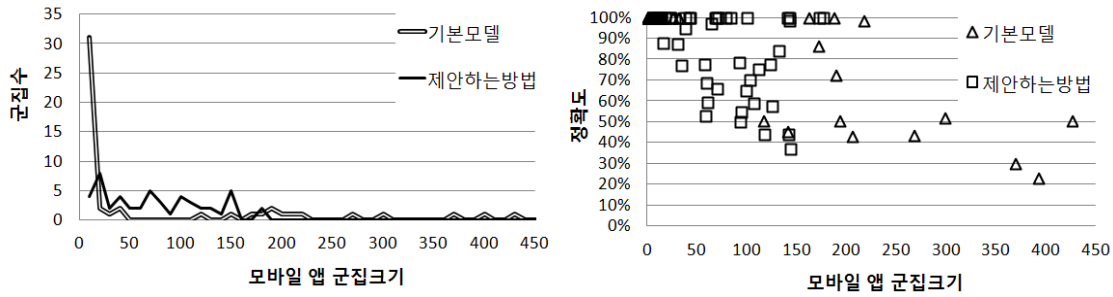

그림 4. 모바일 앱 군집크기에 따른 군집수 및 정확도 분포
Fig. 4. Distribution of Number of Clusters and Accuracy According to Mobile App Cluster Size

지는 군집을 활용할 경우 전체정확도가 오히려 떨어질 수 있다는 것을 나타낸다[4]. 모바일 앱 군집화에 활용되는 단어군집 개수가 2개 이상인 경우 전체정확도의 성능 개선 차이가 미비하므로, 효율성을 고려하여 2개 단어 군집을 활용하는 제안하는 방법과 단어 군집을 활용하지 않는 기본 모델을 중심으로 비교한다.

단어군집을 활용하지 않는 기본 모델에서는 평균 정확도(89.02%)와 전체 정확도(57.48%)의 편차가 심하다. 평균정확도는 군집 크기에 상관없이 군집별 정확도의 평균을 계산하므로, 작은 군집이 많을수록 유리하다. 그러나, 전체 정확도는 전체 모바일 앱을 대상으로 정확도를 계산하므로, 모바일 앱을 많이 포함하는 큰 군집에 영향을 많이 받는다. 따라서, 단어 군집 정보를 활용하지 않은 기본 모델은 군집 크기가 작은 경우가 많았고, 큰 군집의 경우 정확도가 낮은 편이었다. 반면, 단어 군집 정보를 활용하는 제안하는 방법은 상대적으로 작은 군집이 많지 않았고, 큰 군집중 일부의 정확도가 높게 나타났다. 따라서, 평균정확도(85.21%)와 전체정확도(79.66%)의 차이가 상대적으로 크지 않았다. 이러한 결과에 대해 군집별 정확도를 바탕으로 좀 더 자세히 분석하면 다음과 같다.

[그림4]의 왼쪽 그래프에 제시된 바와 같이, 단어군집을 활용하지 않는 기본 모델에서 생성한 군집 50개중에서 36개의 군집이 모바일 앱을 40개 이하를 포함하고, 31개의 군집이 모바일 앱을 10개 이하를 포함하고, 이 중 14개는 모바일 앱을 1개만 포함한다. 단어군집을 활용하지 않는 기본 모델에서 생성한 군집의 크기가 40이하로 작은 경우 해당군집의 정확도가 100%로 측정되어 평균 정확도를 높이는데 영향을 주었다. 정확도가 100%인 군집은 50개의 군집중에서 총 38개였다. 따라서, [그림3]과 같이 단어군집을 활용하지 않는 기본 모델에서 군집별 정확도의 평균은 89.02%로 나타났다.

반면에, 단어군집을 활용하지 않는 기본 모델에서 생성한 군집 50개중에서 12개 군집이 모바일 앱을 150개 이상 포함하고 있다. [그림4]의 오른쪽 그래프에 제시된 바와 같이 군집당 모바일 앱 수가 늘어날수록 정확도는 떨어지는 경향을 보인다. 전체 모바일 앱을 대상으로 하는 전체 정확도에서는 모바일 앱을 많이 포함하는 군집의 정확도에 영향을 많이 받는다. 예를 들어, 모바일 앱 426개를 포함하는 가장 큰 군집의 정확도는 50%인데, 평균 정확도에 대해 2%(=1/50)의 영향력이 있지만, 전체 정확도에 대해서는 12.10%(=426/3,521)의 영향력이 있다. 이를 바탕으로, 단어군집을 활용하지 않는 기본 모델은 [그림3]에 제시된 바와 같이 전체정확도가 57.48%로 나타났다.

한편, 단어군집을 활용하는 제안하는 방법에서 생성한 군집은 [그림4]의 왼쪽 그래프에 제시된 바와 같이 6개에서 177개 사이의 모바일 앱을 포함하고, 50개의 군집중에서 4개의 군집만 모바일 앱을 10개 이하로 포함한다. 이는 단어군집 정보를 활용하는 제안하는 방법이 군집의 크기가 지나치게 작거나 크지 않도록 관리하고 있다고 설명할 수 있다. 따라서, 정확도가 100%인 군집은 50개의 군집중에서 26개였다. 이와 같이, 단어군집을 활용하는 제안하는 방법은 정확도가 100%인 군집을 상대적으로 적게 생성하여 50개 군집의 평균정확도는 85.21%로 단어군집을 활용하지 않는 기본 모델에 비해 3.82% 낮았다. 반면에, 단어군집을 활용하는 제안하는 모델에서 생성한 단어군집은 크기편차가 크지 않고, 모바일 앱을 177개 포함하는 가장 큰 군집의 정확도로 100%였다.

결과적으로, [그림3]에 제시된 바와 같이 단어군집을 활용하지 않고 모바일 앱을 군집화하는 기본 모델의 전체정확도는 57.48%인 반면, 단어군집을 2개 활용하여 모바일 앱을 군집화하는 제안하는 모델의 전체 정확도는 79.66%이다. 즉, 제안하는 방법은 단어군집을 활용하여 모바일 앱 군집 결과에 대한 전체정확도를 22.18%까지 높일 수 있었다.

V. 결 론

본 논문에서는 모바일 앱 트렌드를 고려한 2단계 군집화 방법을 제안한다. 제안하는 방법은 모바일 앱 텍스트에서 얻은 단어 자질을 바탕으로 단어 군집을 생성하고, 그 군집 결과를 활용하여 모바일 앱을 군집화한다. 즉, 제안하는 방법은 단어와 모바일 앱에 대해서 군집화를 각각 실시한다. 제안하는 방법의 특징은 다음과 같다.

첫째, 제안하는 방법은 의미적으로 유사한 모바일 앱끼리 묶을 수 있으므로, 모바일 앱 트렌드의 빠른 변화에 적절히 대응할 수 있다. 미리 정의된 분류체계를 바탕으로 모바일 앱을 분류하는 방식은 모바일 앱 트렌드의 변화에 따라 분류체계를 수정하기가 쉽지 않다. 반면 제안하는 방법은 주어진 모바일 앱 집합에 군집화 기술을 적용하여 유사한 모바일 앱 군집을 생성하므로 분류체계를 정의할 필요가 없다.

둘째, 제안하는 방법은 모바일 앱 텍스트의 부족한 단어정보를 보완할 수 있다. 모바일 앱 텍스트가 대체적으로 짧고 노이즈 정보가 많아서 유효한 단어 정보가 부족하다. 이를 고려하여 제안하는 방법은 각 단어에 대해 unigram 뿐만 아니라, bigram, trigram, 단어 군집 정보를 추가적으로 확보하여 활용함으로써 단어 정보를 증가시킨다.

셋째, 제안하는 방법은 모바일 앱 군집의 크기가 지나치게 작거나 크지 않도록 관리할 수 있다. 실험결과 단어군집을 활용하지 않는 방법은 한 군집에 포함된 모바일 앱의 수가 40개보다 작은 경우나 150개보다 큰 경우가 대부분이고 중간크기의 군집이 거의 없었다. 반면에 단어군집을 활용하는 방법은 한 군집에 포함된 모바일 앱의 수가 180개 이상인 큰 군집은 없었고, 중간크기의 군집이 대부분이었다.

넷째, 제안하는 방법은 모바일 앱을 전체적으로 정확하게 군집화한다. 전체 모바일 앱을 대상으로 하는 정확도를 비교한 결과 단어군집을 활용하지 않는 방법의 정확도가 57.48%인데 반해 단어군집을 활용하는 방법의 정확도는 79.66%로 나타났다. 이는 단어 군집 정보가 모바일 앱 군집의 크기뿐만 아니라 정확도를 적절하게 관리한다는 것을 보여준다.

REFERENCES

[1] S. S. Kim, K. S. Han, B. S. Kim, S. K. Park and S. K. Ahn, "An Empirical Study on Users' Intention to Use Mobile Applications", Journal of Korean Institute of Information Technology, Vol. 9, No. 8, pp. 213-228, Aug. 2011.

[2] J. M. Lim, J. Y. Yu, S. J. Jang, J. H. Lee and J. M. Yu, "Survey on the Internet Usage", Korea Internet & Security Agency, pp. 284, Dec. 2013.

[3] 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.

[4] J. Heo, S. Y. Park, "Word Cluster-based Mobile Application Categorization", Journal of The Korea Society of Computer and Information, Vol. 19, No. 3, pp.17-24, Mar. 2014.

[5] H. S. Lim, "Development Trends and Construction of an Automatic Document Classifier", Journal of Internet Computing and Services, Vol. 3, No. 3, pp. 48-56, Sep. 2002.

[6] Y. Yang, J. O. Pedersenm, "A Comparative Study on Feature Selection in Text Categorization", Proc. of the International Conference in Machine Learning, pp. 412-420, July. 1997.

[7] 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.

[8] C. Apte and F. Damerau, "Automated Learning of Decision Rules for Text Categorization", ACM Trans. on Information Systems, Vol. 12, No. 3, pp. 223-251, July. 1994.

[9] E. Weiner, J. O. Pedersenm and A. S. Weigned, "A Neural Network Approach to Topic Spotting", Proc. of the Annual Symposium on Document Analysis and Information Retrieval, pp.317-332, Apr. 1995.

[10] T. Joachims, "Text Categorization with Support Vector Machines : Learning with many relevant

features", Proc. of International Conference on Machine Learning, pp. 137-142, July. 1998.
〔11〕 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.
〔12〕 D. W. Noh, S. Y. Lee and D. Y. Ra, "Developing a Text Categorization System Based on Unsupervised Learning Using an Information Retrieval Technique", Journal of KIISE : Computer Systems and Theory, Vol. 34, No. 2, pp. 160-168, Feb. 2007.
〔13〕 P. Liang, D. Klein, "Online EM for unsupervised models", Proc. of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 611-619, Jun. 2009.
〔14〕 O. Zamir, "Fast and Intuitive Clustering of Web Documents," Proc. of the International Conference on Knowledge Discovery and Data Mining, pp. 287-290, Aug. 1997.
〔15〕 O. Zamir and O. Etzioni, "Web Document Clustering: A Feasibility Demonstration," Proc. of ACM SIGIR, pp.46-54, Aug. 1998.
〔16〕 O. Zamir and O. Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results," Proc. of the International World Wide Web Conference, pp.1361-1374, May. 1999.
〔17〕 G. Wei, "Named Entity Recognition and An Apply on Document Clustering," MSCs thesis, Dalhousie University, Oct. 2004.
〔18〕 H. Toda and R. Kataoka, "A Search Result Clustering Method Using Informatively Named Entities," Proc. of ACM International workshop on WIDM, pp.81-86, Nov. 2005.
〔19〕 K. Y. Sung and B. H. Yun, "Topic based Web Document Clustering using Named Entities", Journal of the Korea Contents Association, Vol. 10, No. 5, pp. 29-36, May. 2010.
〔20〕 D. H. Kim, K. H. Joo and J. T. Choi, "An Effective Content Clustering Method for the Large Documents", Proceedings of KIIT Summer Conference, Hanbat National University, Korea, pp. 289-297, Jun. 2006.
〔21〕 J. C. Shin and C. Y. Ock, "Search Results Clustering In Real-time", Korea Computer Congress 2009, Mokpo National Maritime University, Korea, pp. 474-479, Jun. 2009.
〔22〕 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.
〔23〕 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.
〔24〕 J. A. Hartigan, and M. A. Wong, "A K-means Clustering Algorithm", Applied. Statistics, Vol. 28, No. 1, pp.100-108, Mar. 1979.

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