Journal of Korea Multimedia Society Vol. 11, No. 9, September 2008(pp. 1195-1203)

음향 DB 구축을 위한 한국어 의성어 군집화

김명관[†], 신영석[††], 김영래[†††]

요 약

한글 문서에서 의성어는 자연적 혹은 인공적 소리를 인간의 언어로 표현하는 것으로서, 대상과 가장 가깝게 느껴지는 의성어 단어로 표현할 수 있으며 또한 음향 도서관구축 등 멀티미디어 데이터를 분류하는 기준으로 활용할 수 있다. 이 연구에서 우리는 말뭉치에서 의성어들의 출현빈도를 구하고, 실험에서 사용할 의성어 100개를 선별하였다. 의성어의 관계를 분류하기 위하여 유사도 및 거리 매트릭스의 특징을 추출하고, 이후에 주성분 분석 방법(PCA)을 사용하여 의성어 특성의 차원을 낮추었으며 의성어들의 관계를 벡터 공간에 표현하였다. 비계층적 클러스터링 방법 들을 비교하여 k-means 알고리즘을 사용하였다. 결과로 의성어를 분류하였고 분류 결과를 통해 의성어들의 특성을 반영할 수 있었다.

Korean Onomatopoeia Clustering for Sound Database

Myung-Gwan Kim[†], Young-Suk Shin[††], Young-Rye Kim[†††]

ABSTRACT

Onomatopoeia of korean documents is to represent from natural or artificial sound to human language and it can express onomatopoeia language which is the nearest an object and also able to utilize as standard for clustering of Multimedia data. In this study, We get frequency of onomatopoeia in the experiment subject and select 100 onomatopoeia of use to our study. In order to cluster onomatopoeia's relation, we extract feature of similarity and distance metric and then represent onomatopoeia's relation on vector space by using PCA. At the end, we can clustering onomatopoeia by using k-means algorithm

Key words: Sound Library(음향도서관), K-means Clustering Algorithm(K-평균 클러스터링), Onomatopoeia(의성어), Principal Component Analysis(PCA)(주성분 분석)

1. 서 론

오늘날 인터넷과 컴퓨터의 보급으로 대량의 이미지, 사운드, 동영상 등과 같은 멀티미디어 데이터들이 발생하였다. 그래서 대량의 멀티미디어 데이터들을 보다 효율적으로 분류하고 관리할 수 있는 방법에 대한 연구들이 곳곳에서 이루어지고 있다. 그 중 의성어는 음향 데이터를 분류하는 유용한 도구로 이용될 수 있다. 많은 멀티미디어 데이터들은 일반적으로 소리나 영상 등을 포함하며 이러한 소리나 영상 속에는 많은 종류에 의성어들이 포함되어 질 수 있다. 따라서 의성어를 올바로 분류한다면 대량의 멀티미디어 데이터들의 효율적 분류와 관리가 가능하다.

음향 데이터를 분류하기 위하여 진행된 연구로는 다음과 같은 것들이 있다. 소리의 특징을 표현한 의성어를 사용하여 음향 데이터를 분류하는 연구가 있었다[1]. 또한, Nearest feature line (NFL)방법을 사용하여 내용 기반의 음향 데이터베이스를 효과적으

※ 교신저자(Corresponding Author) : 김명관, 주소 : 경기도 성남시 수정구 양지동 212(461-713), 전화 : 031)740-7312, FAX : 031)740-7366, E-mail : binsum@culji.ac.kr
접수일 : 2008년 2월 27일, 완료일 : 2008년 7월 7일
[†] 정회원, 을지대학교 의료산업학부 부교수
[††] 준회원, 을지대학교 의료산업학부졸업
(E mail : xzerostone@bmc.hanyang.ac.kr)
[†††] 준회원, 을지대학교 의료산업학부 졸업
(E-mail : kissgood@konkuk.ac.kr)

로 분류 및 검색하는 연구가 있었으며 이 시스템의 예는 http://www.musclefish.com과 http://www.comparisonics.com 에서 확인할 수 있다[2]. 그리고 단어의 설명을 사용한 음향 분류의 분석에 대한 연구가 있었다[3].

본 논문에서는 멀티미디어데이터 분류 및 다양한 분야에서 활용을 위해 주성분 분석을 통한 한국어 의성어의 분류를 실험하였다. 실험 대상은 한말 연구학회의 박동근님이 작성해놓은 『의성어 의태어 어휘목록』과 『한국어의 의성어와 의태어 : 서울대학교 출판부』에서 총 26,480개에 있는 의성어를 바탕으로 말뭉치(600권 분량의 소설)에서 출현 빈도수가 높은 의성어 단어 100개를 선정하였다. 선정된 의성어 단어 100개간의 관계를 위해 tf/idf로 유사도를 구하고 유사도에 따른 각 단어의 거리 값을 구한다. 그리고 주성분분석을 이용하여 2차원 맵으로 표현하고 k-means 알고리즘을 사용하여 한국어 의성어 단어를 클러스터링 하였다.

2. 실험 환경 구성 및 시스템 구조

2.1 의성어의 관계 분석을 위한 말뭉치의 구성

실험에 사용한 의성어 빈도 및 유사도 특징 추출을 위한 말뭉치는 인터넷 자료실인 "CLUB BOX"에서 이용자들이 가장 선호하는 소설로 선별하였다. 소설은 다양한 장르로 분류할 수 있는데, 우리는 먼저 6가지 장르를 선별하고 나머지 장르는 기타에 포함하여 총 7개의 분류로 나누어 실험 환경을 구성하였다. 7가지 분류의 범위는 무협 20%, 추리 15%, 판타지 15%, 고전 10%, 역사 15, 로맨스 5%, 기타 5%로 조사되었고 표 1과 같다. (소설 600권 분량의 데이터)

2.2 의성어의 관계 분석을 위한 대표 의성어 추출

의성어 빈도 및 유사도 추출을 위한 말뭉치를 대상으로, 의성어, 의태어 어휘목록의 총 26,480개에서 출현발생수를 구하여 출현발생수가 가장 높은 상위 100개의 의성어를 구성하였다. 다음 표 2와 표 3은 출현발생수가 높았던 상위 100가지 의성어 예이다.

표 1. 의성어 추출 실험에 사용된 말뭉치의 비율

분류	판타지 소설	무협 소설	추리 소설	역사 소설	고전 소설	로맨스 소설	기타
양(%)	30	20	15	15	10	5	5

2.3 시스템의 구성

의성어분류를 하기위하여 제안된 시스템은 다음과 같이 구성하였다. 첫째로 인터넷 자료실인 "CLUB BOX"에서 말뭉치를 구성하기 위하여 여러 장르의 소설을 수집하였다. 다음으로 의성어 목록을 만들기 위하여 의성어 의태어 어휘목록 총 26,480개에서 각 의성어의 출현발생수를 구하여 출현발생수가 높은 상위 100개의 의성어를 선정하였다.

그리고 선정된 의성어 단어 100개간의 관계를 구하기 위해 tf/idf로 유사도를 구하고 유사도에 따른 각 단어 사이의 거리 값을 구한다. 다음으로 특징이 추출된 대표 의성어를 주성분 분석 방법(PCA)을 사용하여 2차원 맵에 표현한다. 마지막으로 군집화를

표 2. 말뭉치에서 의성어 출현 횟수의 사례

의성어	출현 횟수	의성어	출현 횟수	의성어	출현 횟수
중얼	22846	하하	7826	딱딱	4664
똑똑	4394	훌쩍	3198	후후	3080
껄껄	2296	벌컥	2197	꿀꺽	2153
줄줄	2016	툭툭	1968	으르렁	1958
앗	1924	질질	1825	호호	1817
호호	1804	펄럭	1790	펑	1724
뚝뚝	1645	헤헤	1552	쯧쯧	1383
톡톡	1341	빽빽	1251	깔깔	1204
덜컥	1163	출렁	1101	주르륵	908
끙끙	903	쩌렁쩌렁	874	달랑	869

표 3. 대표 의성어 추출 목록

꿀꺽 꿀꿀 쿵쿵 쾅쾅 쿵쾅 똑딱 딱딱 똑똑 야옹 둥둥 덜컥 딸꾹 철썩 찰칵 멍멍 짹짹 뚜벅 칙칙 빵빵 툭탁툭탁 컥컥 휙휙 쌔앵 풍덩 우르릉 쨍그랑 꼬끼오 삘릴리 푸드득 따르릉 뽀드득 부엉부엉 똑딱똑딱 삐뽀삐뽀 삐약삐약 재깍재깍 뚜벅뚜벅 웅성웅성 와글와글 음메 사각사각 중얼 하하 껄껄 벌컥 줄줄 툭툭 으르렁 호호 호호 펄럭 뚝뚝 헤헤 주르르 출렁 쩌렁 톡톡 주르륵 깔깔 쯧쯧 낄낄 찰그랑 쿵당 드르렁드르렁 빠드득 딸랑 스르륵 찰랑 쿨럭 쩌렁쩌렁 까르르 히히 왈칵 빽빽 끙끙 달랑 아하하 웅웅 콜록 와르르 으드득 쩝쩝 와장창 달그락 꿀꺽꿀꺽 어이쿠 쿵쾅쿵쾅 빠지직 끅끅 후후 훌쩍 질질 퍽퍽 꼬르륵 우적우적 으악 앗 펑 꽝 쿵

하기 위하여 k-means 알고리즘을 사용하여 클러스터링하고, 5개의 클러스터로 구분하여 의성어들을 분류하였다. 우리의 시스템의 구조는 그림 1과 같다.

3. 실험에 사용된 알고리즘

3.1 주성분 분석법 (PCA)

주성분 분석은 고차원 데이터로부터 데이터의 구조를 밝히거나, 데이터의 차원을 낮추는데 많이 이용되는 다변량 통계 분석 방법이다. 이는 상관행렬(correlation matrix)의 고유벡터(eigenvectors)를 찾아내는 문제로 행렬 연산으로 찾아내는 방법과 신경망 등을 사용하여 반복적으로(iteratively) 찾아내는 방법 등이 있다. 즉, 주어진 데이터를 분산이 최대가 되는 축으로 변환하는 것으로, 이 새로운 차원에서의 데이터의 벡터들을 주성분 (principal components)이라고 한다. 이 때 분산이 작은 성분을 제거함으로써 데이터의 차원을 줄이는 동시에 데이터에 포함되어 있던 잡음(noise)을 제거할 수 있다. 데이터 행렬 X 의 차원을 k 로 낮추는 식이 다음과 같다[4].

$$X \cdot V^k$$

여기서 V 는 X 의 상관행렬의 고유벡터를 해당하는 고유값 (eigenvalue)의 내림차순으로 정렬한 행렬이고, k 는 이 중 k 개의 열을 사용하겠다는 의미이다.

3.2 단어 클러스터링

단어 클러스터링은 용도에 따라 비슷한 특성을 갖는 단어들을 같은 클래스로 병합하는 과정을 말한다. 언어학적인 관점에서 클러스터링을 수행할 경우, 문법적으로 유사한 쓰임을 갖거나, 의미적으로 연관된 개념을 가지는 단어들을 클러스터링하게 된다. 일반적으로 문법적인 범주에 따라 나눠진 품사집합의 경우도 단어를 문법적인 쓰임이 유사한 단어들로 클러스터링 한 경우로 볼 수 있다[5].

본 논문에서는 단어의 클래스를 의성어와 주변 의성어의 연관성을 기준으로 클러스터링을 하는 방법을 제시한다. 즉 각 단어에 대하여 클러스터링 한 기준이 되는 특성 벡터(feature vector)를 정의한 후 특성 벡터들 사이의 유사도가 높은 단어들을 같은 클러스터로 할당하는 방법을 제안한다.

3.3 k-means Clustering Algorithm

k-means 은 군집화 (Clustering) 문제를 해결하는 간단한 자율학습(Unsupervised Learning) 알고리즘 중 하나이다. 사전에 정해진 어떤 수의 클러스터를 통해서 주어진 데이터 집합을 분류하는 간단하고 쉬운 방법이다. 데이터 이외에 클러스터의 수 k를 입력으로 하며 이때 k를 seed point 라고 한다. seed point는 임의로 선택되며 바람직한 클러스터구조에 관한 어떤 지식들이 seed point를 선택하는 데에 사용될 수 있다. 하나의 샘플이 하나의 클러스터에 합류하자마자 곧 클러스터의 centroid 가 다시 계산된다. 데이터 집합에서 단지 두 번 만에 통과가 이루어진다. 그 과정은 다음과 같다.

(1) 처음에 k 클러스터로서 시작한다. 남아있는 n-k 샘플들에 대해서는 가장 가까이 있는 centroid를 찾는다. 이것에 가장 가까이 있는 centroid를 가지는 것이 확인된 클러스터에 샘플을 포함시킨다. 각각의 샘플들이 할당된 후에 할당된 클러스터의 centroid 가 다시 계산된다.

(2) 그 데이터를 두 번 처리한다. 각 샘플에 대하여 가장 가까이 있는 centroid를 찾는다. 가장 가까이 있는 centroid를 가진 것으로 확인된 클러스터에 샘플을 위치시킨다. (이 단계 에서는 어떤 centroid 도 다시 계산하지 않는다.)

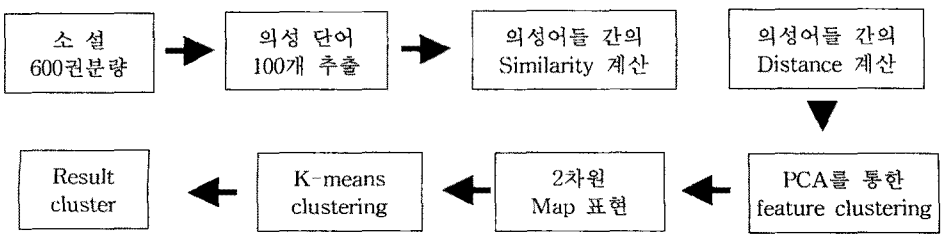


그림 1. 제안된 시스템의 구조

4. 실 험

4.1 유사도(similarity)/거리 매트릭(distance metric) 과 2차원 맵의 표현

실험에 사용한 데이터는 600권 분량의 여러 장르의 소설이다. 한말 연구학회의 박동근님이 작성해놓은 『의성어 의태어 어휘목록』과 『한국어의 의성어와 의태어 : 서울대학교 출판부』에서 총 26,480개에 있는 의성어를 바탕으로 말뭉치(600권 분량의 소설)에서 출현 빈도수가 높은 의성어 단어 100개 선정하여 각 페이지에서 의성의 발생 수 및 페이지를 조사하였다. 출현 빈도수가 높은 100개의 의성어를 클러스터링하기 위하여 먼저 의성어를 데이터 벡터로 만들어 준다. 따라서 tf/idf(term frequency/inverse document frequency)를 이용하여 의성어간에 유사도(similarity) 및 거리 매트릭(distance metric)을 구하였다. 이 100×100 거리 매트릭스를 PCA를 사용하여 저차원 공간으로 바꾸어 2차원 맵에서 표현하기 위하여 Self-Organizing Map Toolbox2.0 [6]을 가지고 저차원 공간에서 의성어간의 관계를 표현한 2차원 map을 생성하였다.

4.1.1 2차원 맵의 표현

무수히 많은 데이터들을 수치에 따라 그들 사이에 위치관계를 파악하는 것은 불가능하다. 따라서 이들을 시각화하여 그래프나 그림 등으로 표현해주면 그 결과에 대해 알아보기 쉽고 파악하기에 유용할 것이다. 우리는 의성어간의 관계를 2차원 맵에서 구분하기 쉽게 표현하고자 Teuvo Kohonen등에 의해 만들어진 Self-Organizing Map Toolbox2.0이라는 툴을 사용하여 얻고자하는 수치를 시각화하여 표현하였다. 먼저 유사도와 거리공식으로부터 얻어진 100×100 거리 매트릭스에 각 column과 row에 대응하는 의성어를 라벨링(labeling)하였다. 그림 2는 의성어가 라벨링된 100×100 거리 매트릭스의 일부분이다.

다음으로 그림 2처럼 라벨링한 100×100의 거리매트릭스를 가지고 데이터벡터들을 정규화 한다. 정규화는 변수들의 분산을 이용하여 계산하며, 분산 x′은 다음과 같이 구할 수 있다.

$$x' = (a(x) - x) / \sigma x$$

여기서, a(x) 는 변수 x의 평균이며, σx는 표준편차이다. 분산을 이용한 정규화 과정을 통해 그림 3과 같은 결과를 얻을 수 있었다.

이렇게 생성된 100×100 데이터 벡터들을 사용하여 맵의 생성을 위하여 학습과 초기화를 수행한다. 이 과정을 거쳐 48×100형태의 codebook이라고 하는 데이터 벡터들이 생성된다. 그리고 각 대응하는 100개의 라벨들도 48개로 축소되며 의성어간의 관계가 매우 밀접하거나 이웃하는 정도가 거의 일치하는 데이터의 라벨은 그 중심이 되는 라벨만을 표시하고 그 이 외의 52가지 의성어 라벨들은 별도로 기억하게 된다. 생성된 codebook 데이터의 일부분은 그림 4에서 볼 수 있다.

이렇게 생성된 codebook 데이터를 이용하여 PCA에 투영시켜 2차원 맵에서 그 관계를 의성어 라벨로 표현하였고 이 관계는 그림 5에서 확인할 수 있다.

그림5의 2차원 맵에 나타난 결과를 보면 분포된 결과가 인간의 육성의 소리(중얼, 껄껄, 훌쩍 · · ·) 또는 사물과 관련된 소리(“달랑, 찰랑, 와르르 ·

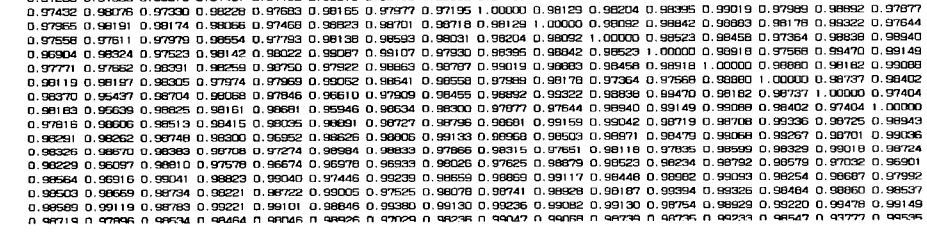


그림 2. 100×100 거리매트릭스의 각 column과 row에 해당하는 의성어 라벨링 결과의 일부분

	'중얼'	'하하'	'딱딱'	'똑똑'	'꼴찍'	'후후'	'컬컬'	'벌컥'	'꿀꺽'	'을을'	'투
'중얼'	0.8038	-3.3334	-5.6061	-4.2596	-3.1606	-2.2762	-1.4036	-2.564	-1.2788	-2.1245	
'하하'	-4.1846	0.7009	-2.1761	-3.3265	-1.5549	-6.7715	-4.6785	-1.7599	-0.7893	-1.771	
'딱딱'	-4.0203	-1.1362	0.8762	-2.3424	-1.1062	-0.9424	-0.6471	-1.6114	-1.3583	-1.7976	
'똑똑'	-3.2697	-1.9968	-2.5598	0.8903	-1.9713	-0.7064	-2.2163	-2.317	-0.6738	-1.9865	
'꼴찍'	-2.5469	-0.959	-1.3944	-2.1301	0.9837	-0.734	-2.8684	-2.0266	-1.088	-2.9018	
'후후'	-2.0774	-5.1849	-1.2423	-0.8112	-0.7142	0.6696	-0.084	-0.3015	-0.7217	-0.7825	
'컬컬'	-0.9748	-2.8999	-0.6838	-2.1971	-2.6953	-0.0473	0.7748	-3.3352	-0.8646	-0.9733	
'벌컥'	-1.6267	-0.9203	-1.515	-1.9669	-1.5692	-0.2346	-2.8475	0.9927	-1.4589	-0.9467	
'꿀꺽'	-1.5019	-0.6962	-2.345	-1.121	-1.5893	1.0149	-1.4498	-2.6928	0.6729	-1.868	
'을을'	-1.0233	-0.6126	-1.3268	-1.3049	-1.8281	-0.4108	-0.6537	-0.6917	-0.749	1.0585	
'똑똑'	-1.3887	-1.0338	-1.5621	-0.751	-1.4671	-1.0397	-0.7724	-1.5944	-0.692	-1.9258	
'으르렁'	-1.9759	-0.5161	-2.1122	-1.2187	-1.2128	-0.1685	-0.2072	-1.7271	-0.5469	-0.7528	
'앗'	-1.1975	-0.9968	-1.065	-1.0958	-0.4044	-1.2379	-0.4755	-0.6011	-0.0726	-0.6886	
'질질'	-0.885	-0.6083	-1.1688	-1.4094	-1.2717	-0.1914	-0.7196	-0.9019	-0.8554	-1.7913	
'호호'	-0.6597	-2.6123	-0.6874	-1.3027	-1.4083	-2.4423	-1.5246	-1.0373	-0.1692	-0.002	
'호호'	-0.8276	-2.4657	-0.5414	-1.1971	-0.481	3.0519	-0.7273	-1.2409	-0.9406	-2.6266	
'펄럭'	-1.1571	-0.3113	-0.9178	-0.9088	-1.1984	-0.3484	-0.6251	-0.5892	-0.3295	-0.2569	
'펑'	-0.7306	-0.5611	-0.6343	-1.0393	-2.4011	-0.5917	-0.5382	-0.1465	-0.1114	-1.293	
'똑똑'	-0.6992	-0.2648	-1.0747	-0.5762	-2.0435	-0.263	-0.5085	-1.8111	-0.6077	-2.6156	
'헤헤'	-0.7863	-2.1331	-0.5595	-1.9588	-2.7098	-2.1045	-2.5979	-1.6009	-1.1321	-0.6949	
'풋풋'	-0.4855	-1.5384	-0.2808	-0.4457	-0.0824	-1.6749	-0.062	-0.7692	-0.1866	-0.3226	
'툭툭'	-0.5403	-0.2656	-0.6512	-1.129	-0.4355	-0.2498	-1.9469	-1.5326	-0.2839	-0.6183	
'뻐뻐'	-0.4631	0.0612	-0.5921	0.0061	-0.0146	-0.3897	0.093	-0.1504	0.0923	-0.3774	
'깔깔'	-0.3463	-0.8268	-0.8925	-0.8532	-1.1862	-0.3163	-2.4923	-1.325	-0.0514	-0.3993	

그림 3. 정규화 과정을 거쳐 나온 100×100 매트릭스의 일부분

구분	'중얼'	'하하'	'딱딱'	'똑똑'	'꼴찍'	'후후'	'컬컬'	'벌컥'	'꿀꺽'	'을을'
'으르렁'	-0.6829	-0.2003	-0.883	-0.7015	-0.9556	-0.1602	-0.3908	-0.813	-0.3579	-0.8759
'펄럭'	-0.7063	-0.2374	-0.8078	-0.7233	-0.8818	-0.2472	-0.5188	-0.6663	-0.4071	-0.8878
'출렁'	-0.4202	-0.0109	-0.5039	-0.3672	-0.5435	-0.1806	-0.2394	-0.2636	-0.2555	-0.6853
'뻐뻐'	-0.1943	0.0831	-0.243	-0.1936	-0.375	-0.0719	-0.193	-0.1266	-0.0991	-0.541
'꿀꺽'	0.0086	0.1659	-0.0281	0.0206	-0.1999	0.0475	0.0826	0.0852	0.0629	-0.2951
'웅웅'	0.1803	0.2477	0.1564	0.124	-0.0046	0.153	0.2199	0.2592	0.1764	-0.0246
'으밋'	0.3623	0.3445	0.3704	0.3259	0.2846	0.2983	0.4342	0.4585	0.3054	0.1996
'와르르'	0.4534	0.4492	0.4639	0.4212	0.3945	0.3904	0.502	0.4856	0.3652	0.3141
'으드득'	0.5404	0.5188	0.5676	0.5405	0.5071	0.458	0.5599	0.5288	0.4394	0.4032
'어이구'	0.6235	0.5766	0.6522	0.6455	0.6803	0.5142	0.5772	0.6567	0.4609	0.569
'몸살웅성'	0.6716	0.5923	0.6996	0.6833	0.7768	0.5016	0.6165	0.7072	0.4617	0.6246
'퍽퍽'	0.7091	0.6296	0.7338	0.7404	0.8505	0.5716	0.6932	0.7409	0.486	0.8212
'똑똑'	-1.0431	-0.5345	-1.1058	-1.0446	-1.1015	-0.4156	-0.884	-0.9548	-0.5694	-1.1227
'질질'	-1.0799	-0.7505	-1.0576	-1.1335	-1.0751	-0.6168	-1.0529	-0.9581	-0.591	-1.1067
'쩌렁'	-0.5103	-0.1836	-0.5393	-0.5791	-0.6571	-0.2461	-0.5417	-0.4621	-0.3131	-0.7639
'끙끙'	-0.2484	-0.0839	-0.2754	-0.3784	-0.4782	-0.1533	-0.4662	-0.3474	-0.1322	-0.578
'왈칵'	-0.0204	0.1507	-0.0687	-0.0928	-0.2885	0.0257	-0.1093	-0.0367	0.0293	-0.3455

그림 4. 생성된 codebook 데이터벡터의 일부분

그림 5. 실험에 사용된 의성어들의 관계를 2차원 맵에서 표현한 결과이다.

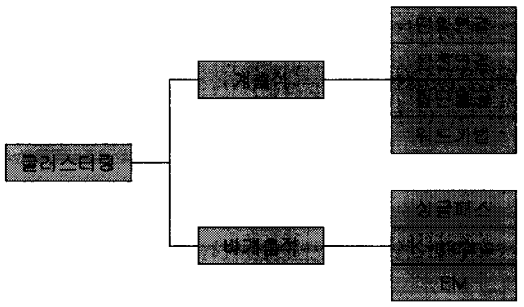


그림 6. 클러스터링의 분류

・・), 동물의 소리(삐약삐약, 야옹・・・)등 과 같은 유사한 것들이 근처에 분포되어 있음을 볼 수 있다.

4.2 클러스터링

4.2.1 클러스터링 방법의 선택

클러스터링은 문서 등을 자동으로 분류하기 위한 비지도학습 방법이다. 클러스터링은 그림 6에서와 같이 계층적방법과 비계층적 방법으로 나뉜다. 이와 같이 다양한 클러스터링 방법이 존재하기 때문에 방법의 선택에 앞서 다음과 같은 것을 생각해 보아야한다.

첫째는 선택하고자 하는 알고리즘이 이론적인 근거를 살펴보아야 하고 둘째로 클러스터링의 효율 즉, 소요되는 시간과 컴퓨터 기억용량을 얼마나 사용하는가를 검토하여야 한다.

이론적인 근거는 다음과 같은 기준을 만족시켜야 한다.

- 시스템에 있는 문헌들의 변동, 삭제, 수정, 첨가가 있어도 전체 클러스터링에 크게 영향이 있어서는 안 된다.

- 문헌의 기술에 있어 작은 실수가 있더라도 전체

클러스터링에 크게 변화를 주어서는 안 된다.

- 최종적인 클러스터링 구성이 초기 클러스터들과 독립적이어야 한다.

표 4에서 비교한 것과 같이 대규모 문서를 클러스터링할 때는 보통 비계층적 클러스터링을 선호한다. 비계층적 클러스터링에는 싱글패스, K-means, EM(Expectation Maximization) 등의 방법들이 있다[7]. 싱글패스 방법은 K-means에 비해 목표 범주에 대한 클러스터 개수의 불일치 및 성능 평가로 사용하는 클러스터 인덱스에서 불리한 것으로 알려져 있다[8]. 또한 EM 방식은 주어진 데이터 집합에 대하여 적합한 가우시안을 사용하는 방법으로서 각 가우시안은 평균값과 공분산 행렬을 가지고 있다. 그러나 낮은 차원(예, 2차원)의 데이터에 대한 클러스터링에 있어 보통 계산 횟수와 클러스터 구성이 K-means가 EM 보다 너 나은 것으로 알려져 있다[9]. 따라서 본 논문에서는 의성어 클러스터링을 위해서 K-means 방법을 채택하였다.

4.2.2 의성어 분류

클러스터링 문제를 해결하는 가장 간단한 자율학습(Unsupervised Learning) 알고리즘 중 하나인 k-means[1]를 사용하여 의성어 단어의 자율적 분류(Unsupervised Clustering)를 시도 하였다. K-means 클러스터링은 data mining의 군집화 작업에 주로 사용된다[10]. 이 기법은 N개의 속성으로 구성되는 각각의 레코드를 벡터로 표시하여 N차원의 데이터 공간(space)에 나타낼 때, 유사한 특성을 갖는 레코드들은 서로 근접하여 위치한다는 가정에 근거하고 있다. 여기에서 영문자 'K'는 K개의 군집을 의미한다.

표 4. 클러스터링 기법들의 비교

비교 요소	계층적 클러스터링	비계층적 클러스터링
클러스터링 성능	높다	낮다
속도	늦다	빠르다
문서규모	작은 문서	큰 문서
예외사항	단일연결기법 (적당하지 않음)	-

1) K-means Clustering Algorithm(MacQueen 1967) <"http://www.aistudy.com/pattern/k-means_clstering.htm">

본 연구에서 K-means 클러스터링 기법을 2차원 공간에서 적용하는 예로서, 각 의성어들 사이의 거리에 따른 2차원 좌표들로 구성된 레코드들을 5종류(K=5)의 군집(부류)으로 분류하는 작업을 단계별로 설명한다.

첫 단계는 각 레코드들 중에서 5개의 레코드를 임의로 선택하여 각 군집의 중심값으로 지정한 후, 나머지 레코드들이 소속될 군집을 결정하고 군집간의 경계선을 긋는다.

단, 소속 군집의 결정 기준은 각 레코드와 5개 중심값과의 직선 거리 중에서 가장 짧은 중심값으로 한다. 예를 들어 첫 번째 군집의 중심값이 (80, 25), 두 번째 군집의 중심값이 (100, 40), 세 번째가 (125, 30)이고, 첫째 레코드의 값이 (90, 45)이라면 이 레코드와 첫 번째 군집의 중심값과의 직선거리는 다음과 같이 계산된다.

$$\text{직선거리} = \sqrt{(90+80)^2+(45-25)^2} = 22.7$$

마찬가지 방법으로 두 번째, 세 번째 군집의 중심값과의 직선거리를 계산하면 각각 11.2와 38.1이 된다. 따라서 첫째 레코드는 일단 두 번째 군집으로 분류된다. 또한 두 군집간의 경계선은 두 중심값과 같은 거리에 위치한 좌표들의 집합, 즉 직선이 된다. 그림에서는 첫 번째, 두 번째, 세 번째 군집에 속한 레코드들을 각각 삼각형, 원, 사각형으로 표시하고 있다.

다음 단계로는 각 군집에 속한 레코드들의 중심값을 재측정한다. 중심값이 구해지면 첫 단계에서와 마찬가지로 각 레코드에 대해 군집의 중심값과 직선거리를 측정하여 가장 근접한 군집에 포함시킨 후, 군집간의 경계선을 표시한다. 특히 첫 번째 군집에 속한 레코드(삼각형으로 표시) 중 하나는 중심값이 이동함에 따라 새롭게 이 군집에 포함된 것을 알 수 있다. 새로운 군집이 형성되면 이전 단계의 과정을 중심값의 이동이 미비할 때까지, 즉 경계선의 변화가 거의 없을 때까지 반복적으로 실행한다.

표 5에서는 k-means 클러스터링의 장점과 단점을 보여주고 있으며[11] k-means의 클러스터링 방법은 총 5단계로 요약해 볼 수 있으며 다음과 같다.

1단계 : random하게 cluster centroid를 선택한다.

2단계 : 각 vector들을 가장 가까운 cluster centroid에 연결한다.

표 5. K-means 클러스터링의 장점과 단점

장 점	단 점
○ K-means 클러스터링 기법을 지원하는 상용화된 제품이 많으며, 사용 또한 쉽고 간편하다. 따라서 사용자가 자신이 원하는 사양을 갖춘 제품을 이용하여 손쉽게 문제의 영역에 적용할 수 있다. ○ 군집분석 이외에도 분류·예측을 위한 선행작업, 특이 오류값이나 결손값 처리작업 등 다양한 분석에 사용할 수 있다.	○ 속성들의 형태가 다르거나 같은 형태의 속성이라도 값의 범위가 다양할 경우 거리 측정기준을 설정하는데 어려움이 따른다. ○ K-means 클러스터링 기법은 사용자가 지정한 K값에 따라 데이터를 K개의 군집으로 나눈다. 그러나 실제 데이터의 구조가 이 값보다 작거나 큰 수의 군집 특성을 갖고 있다면 좋은 결과를 기대하기 어렵다. 군집결과에 대한 해석이 용이하지 않다.

3단계 : 각 vector들과 cluster centroid 사이의 값을 전부 더한 후 평균을 낸다.

4단계 : 3단계에서 나온 결과로 cluster centroid 갱신한다.

5단계 : 2~4단계를 특정 임계값이 만족할 때까지 반복한다.

우리는 이런 과정으로 분류하는 k-means알고리즘을 사용하여 우리가 얻어낸 codebook의 수치들을 가지고 의성어를 분류하였다.

4.2.3 클러스터링 결과

codebook의 의성어 데이터벡터를 가지고 k-means를 사용하여 총 5개의 군집으로 클러스터링 하였다. 5개의 군집에 포함된 개체의 수의 결과는 그림 7과 같다.

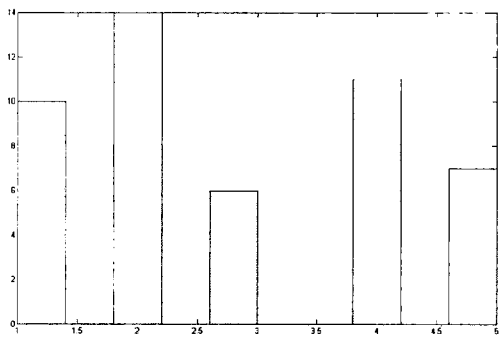


그림 7. 각 클러스터에 포함된 의성어의 누적 수의 그래프

그림6에서 그래프는 각 클러스터에 포함된 대표 의성어의 개수를 누적시킨 히스토그램 그래프이다. 히스토그램 그래프를 보면 Cluster 1에는 10개가, Cluster 2에는 14개, Cluster 3에는 6개, Cluster 4에는 11개 Cluster 5에는 7개가 포함되어 총 48개의 대표 값들이 클러스터링 된 것을 확인할 수 있다. 다음은 5개의 군집을 구성하고 있는 데이터벡터의 수치 일부분이며 그림 8과 같다.

다음은 5개의 군집을 형성한 데이터 벡터 값을 가지고 5개의 군집에 실제로 어떤 의성어들이 포함되는지를 테이블로 만들어 분류하여 표 6에 표현하였다.

각 클러스터에는 표 6에서 볼 수 있는 것과 같은 의성어들이 포함 되었다. '**웅웅**(콜록)'을 보면, 먼저 '**웅웅**'은 진하게 표시하였는데 이는 각 48개 항목 중 Cluster1에서 대표되는 의성어이며 '(콜록)'은 '웅웅'에 일치되어 따로 기억되었던 52개 항목 중 연관 있는 의성어이다.

통계적 사실에 근거한 데이터를 가지고 k-means를 통해 클러스터링 한 결과를 통해 우리는 다음과 같은 의미부여를 할 수 있었다. 먼저 Cluster 1은 대체적으로 자연 현상이나 상황에서 발생하는 소리 등에 관한 의성어가 군집을 이루었으며, Cluster 2는 동물이나 기계소리(야옹, 부엉부엉, 꿀꿀, 따르릉, 찰그랑, 찰칵)등과 같은 종류의 의성어가 군집을 이루었다. Cluster 3은 사물이나 환경이 부서지거나 파괴되는 종류의 의성어가 군집을 이루었고, Cluster 4는 주로 웃음소리등과 같은 사람에게서 나타나는 인간

구분	줄얼	하하	딱딱	똑똑	폴짝	후후	껄껄	빕쿡	꿀꿀	줄줄	툭툭
Cluster 1	-0.1673	-0.0346	-0.2141	-0.2473	-0.3574	-0.1127	-0.2454	-0.1996	-0.0687	-0.431	-0.3907
Cluster 2	0.6903	0.6241	0.712	0.7422	0.8223	0.5723	0.65	0.7514	0.4278	0.814	0.8356
Cluster 3	0.4391	0.4405	0.4409	0.4122	0.3891	0.3746	0.4728	0.4733	0.3135	0.3351	0.3467
Cluster 4	-1.0453	-0.9797	-1.0309	-1.1223	-1.089	-0.8185	-1.0165	-1.0969	-0.5108	-1.0427	-1.1099
Cluster 5	0.2841	0.2684	0.2148	0.2769	0.1981	0.1937	0.2125	0.1746	-0.2245	0.2227	0.1623

그림 8. 5개의 Cluster에 대한 데이터벡터 수치의 일부분.

표 6. 각 Cluster에 클러스터링된 의성어 단어(괄호 안의 단어는 같은 값을 갖는 의성어)

Cluster #	onomatopoeia words
Cluster 1	웅웅(콜록) 쩌렁 출렁(주르륵) 왈칵(뚜벅) 쿨럭(주르르, 쿵쾅) 빽빽(둥둥) 끙끙(쩌렁쩌렁)
Cluster 2	야옹(와글와글, 사각사각, 똑딱, 부엉부엉, 삘릴리, 똑딱똑딱) 꿀꿀(짹짹) 웅성웅성 꼬르륵 끅끅(음메) 뽀드득(쌔앵) 따르릉(꼬끼오, 드르렁드르렁, 재깍재깍, 툭탁툭탁, 쿵탕, 찰그랑, 삐약삐약) 컥컥(쿵쾅쿵쾅, 우적우적, 빠지직) 어이쿠 딸랑(푸드득, 달그락) 찰칵
Cluster 3	쨍그랑 와르르(쾅쾅) 으악(퍽퍽, 멍멍, 풍덩) 으드득(우르릉,와장창)
Cluster 4	훌쩍(벌컥, 줄줄) 앗(톡톡, 히히, 낄낄) 으르렁(펑, 덜컥, 꽝, 쿵쿵) 중얼(하하, 딱딱, 똑똑, 후후, 꿀꺽, 툭툭) 질질 껄껄(호호, 호호, 헤헤) 뚝뚝 펄럭 쯧쯧(깔깔)
Cluster 5	칙칙(스르륵) 찰랑(뚜벅뚜벅) 달랑(휙휙, 아하하, 철썩, 까르르) 쩝쩝(빵빵, 빠드득, 딸꾹, 꿀꺽꿀꺽) 쿵

의 육성의 소리(중얼, 껄껄, 훌쩍, 호호, 낄낄)들이 한 군집을 이루었다.

이처럼 통계적 데이터를 기반으로 자동화 분류를 수행하였더니 일반적으로 우리가 알고 있는 사실관계 뿐만 아니라 미처 알지 못했던 단어들 사이의 관계성을 파악할 수 있었다.

5. 결 론

오늘날 소리나 동영상등과 같은 멀티미디어 데이터들의 범람 속에서 멀티미디어 데이터들을 효과적으로 분류하고 관리하기 위한 방법 대해 무수히 많은 연구들이 이루어졌고 이 연구들은 각 각의 특징과 효과들을 가지고 있었다. 그렇지만 소리나 동영상과 같은 멀티미디어 데이터는 모두 음향적 효과를 지니고 있으며 따라서 음향을 표현한 의성어 단어의 사용을 통한 멀티미디어 데이터의 분류는 분명히 매력적이고 유용한 요소임에 틀림이 없다. 본 논문에서 우리는 한국어 의성어 단어라는 요소를 사용하여 의성어 단어의 분류를 실험하였다. 먼저 한국어 의성어 단어의 목록 100여개를 선정하여[12] 소설 7가지 분류로 구성한 말뭉치에서 tf/idf를 통해 의성어 단어 사이에 유사도 및 거리 값을 구하였다. 이렇게 얻어진 실험용 데이터를 가지고 주성분 분석 방법(PCA)을 통해 2차원 맵에서 각 의성어간의 분포 관계를 시각적으로 표현하였으며, 표현된 의성어를 k-means알고리즘을 사용하여 분류하였다. 우리의 실험을 통해 대부분 의성어의 의미가 비슷한 것들끼리의 관계와 분류를 확인할 수 있었다.

이렇게 얻은 클러스터들을 가지고 기존 음향 라이브러리 유사한 소리의 데이터들을 연결하여 음향 디렉토리를 구성할 수 있다. 따라서 향후 주성분 분석을 통한 의성어의 분류를 사용하여 음악, 동영상 등과 같은 멀티미디어 데이터의 분류에 응용이 가능하리라 생각되며 그 외에도 보다 다양한 분야에서 의성어의 활용이 가능할 것이다. 더 정확하고 유용한 결과를 위해 기존 k-means를 개선한 새로운 클러스터링 방법 및 의성어 분석과 분류에 대한 추가적인 연구가 필요할 것이다.

참 고 문 헌

[1] S. Sundaram, and S. Narayanan, "Vector-based representation and clustering of audio using onomatopoeia words," In Proc. of AAAI 2006 Fall Symposia, Arlington, VA, Oct. 2006.

[2] V. Rice, "Audio and video retrieval based on audio content," Comparisonics TM, Grass Valley, CA, USA, <http://www.comparisonics.com/WhitePaper.html>, Apr. 1998.

[3] S. Sundaram, and S. Narayanan, "Analysis of Audio Clustering using Word Description," Hawaii Convention Center Honolulu, pp. 230-237, Apr. 2007.

[4] 신형주, 장병탁, 김영택, "대용량 문서분류에서의 비선형 주성분 분석을 이용한 특징 추출," 한국정보과학회 학술발표회, pp. 146-148, 2000.

[5] 신중호, 한선화, "단어의 유사성 척도와 클러스터링 알고리즘," 제2회 디지털도서관컨퍼런스 논문집., pp. 78-81, Nov. 1999.

[6] J. Vesanto, and J. Himberg, and E. Alhoniemi, and J. Parhankangas, "SOM Toolbox for Matlab 5," Helsinki University of Technology SOM Toolbox Team, pp. 36-36, 2005.

[7] 정영미, 이재윤, "한국어 텍스트 내 용어연관성

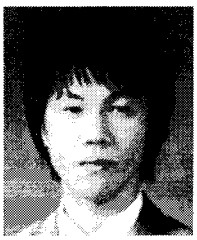
분석을 위한 기초 연구," 한국정보관리학회 학술대회 논문집, 제5권, pp. 243-246, 1998.
[8] 조태호, "가변적 클러스터 개수에 대한 문서 군집화 평가 방법," 한국정보과학회 가을학술발표 논문집, Vol.33, No.2(B), pp. 233-236, 2006.
[9] N. Alldrin and A. Smith and D. Turbull, "Clustering with EM and K-means," www.neilalldrin.com/research/w03/cse253/project1.pdf, CSE 253, 2001.
[10] 이준, 이종태, "신경망을 이용한 군집화 기법의 개선과 데이터 마이닝의 기능 향상 에 관한 연구," 한국경영과학회/대한산업공학회 춘계공동학술대회, pp. 324-327, 2001.
[11] 고영중, 서정연, "문서관리를 위한 자동문서범주화에 대한 이론 및 기법," 정보관리연구, Vol. 33, No.2, pp. 19-32, 2002.
[12] 채완, 한국어의 의성어와 의태어, 서울대학교 출판부, 서울, pp. 40, 2003.

김 명 관

1981년 3월~1985년 2월 숭실대학교 전자계산학과 학사
1985년 3월~1987년 2월 숭실대학원 전자계산학과 석사
1996년 9월~2004년 2월 숭실대학원 컴퓨터학과 박사
1989년 8월~1993년 2월 한국전자통신연구소 인공지능연구실 연구원
1993년 3월~2007년 2월 서울보건대학 컴퓨터정보과 부교수
2007년 3월~현재 을지대학교 의료산업학부 의료전산전공 부교수
관심분야 : 인공지능, 자연어처리, 질의응답시스템, 시멘틱웹

신 영 석

2007년 2월 을지대학교 컴퓨터정보과 졸업
2007년 9월 한양대학원 의용생체공학과 입학
관심분야 : 의용생체공학, 3D 그래픽



김 영 래

2007년 2월 을지대학교 컴퓨터정보과 졸업
2007년 9월 건국대학원 컴퓨터학과 압학
관심분야 : HCI, 유비쿼터스 헬스케어