http://dx.doi.org/10.5392/JKCA.2014.14.04.407

정보 알고리즘 기반 아리랑의 계통도 및 상관관계 분석

Correlation Analysis of the Arirangs Based on the Informatics Algorithms

김학용
충북대학교 자연과학대학 생화학과

Hak Yong Kim(hykim@cbnu.ac.kr)

요약

우리 민족의 대표적인 민요이면서 동시에 유네스코 인류무형문화유산인 아리랑을 정보알고리즘 기법을 도입하여 후렴구를 중심으로 계통도를 분석하고 아리랑들 사이의 상관관계는 본문 단어중심으로 분석하였다. 아리랑의 계통도 분석은 생명체의 진화관계를 분석하는 알고리즘인 다중서열정렬 기법을 사용하였다. 분석한 아리랑 106개 중에서 38개 아리랑이 빠른 템포를 가지고 있었으며, 나머지 68개 아리랑이 느린 템포를 가지고 있었다. 이를 바탕으로 후렴구 기반 아리랑 계통도를 완성하였다. 아리랑 본문 단어는 아리랑에 있는 단어와 아리랑 제목을 노드로 하는 bipartate 네트워크를 구축하고 이들로부터 73개 아리랑 및 104개의 핵심 단어를 추출하였다. 먼저, 이 데이터를 바탕으로 쌍대비교분석 기법을 사용하여 아리랑들 사이의 상관관계를 분석하였다. 또한, 네트워크 연결계수가 1인 노드를 단계적으로 제거하여 핵심네트워크를 구축한 다음 네트워크 기반으로 아리랑들 사이의 상관관계를 분석하였다. 그동안 아리랑을 어원 중심의 인문과학이나 음률적인 접근을 통하여 아리랑의 어원, 계통도, 상관관계를 분석하려는 연구가 있었다. 본 연구에서는 이러한 시도를 벗어나 과학적 접근방법인 정보알고리즘을 사용하여 아리랑을 분석함으로써 세계적인 문화유산의 위상을 한층 더 높이고 객관적인 결과를 통해서 아리랑의 대중화 및 세계화의 기틀을 마련함에 있어 그 방법론을 제시하였다.

■ 중심어 : | 아리랑 | 후렴구상관관계 | 아리랑가사 상관관계 | 정보 알고리즘 |

Abstract

An arirang is the most famous Korean folk song and was registered in UNESCO(Unitied Nations Educational, Scientific and cultural Organization) as an intangible cultural heritage in 2012. Most arirangs are composed of text and refrain parts. Genealogy of the arirang was classified in refrain patterns by using multiple sequence alignment algorithm. There are two different refrain patterns, slow and fast melodies. Of 106 arirangs, 38 and 68 arirangs contain fast and slow melodies, respectively. 73 arirangs and 104 their key words were extracted from bipartate arirang network that composed of arirangs, text works, and their relationships. The correlation among the arirangs was analyzed from the selected arirangs and key words by using pairwise comparison matrix. Also, analysis of correlation among the arirnags was performed by stepwise removal of the single degree nodes from the bipartate arirang network. In this study, arirangs were analyzed in genealogy and correlation among arirangs by using informatic algorithm and network technology, in which arirang research will be constructed a stepping stone for the popularization and globalization of the arirangs.

■ keyword : | Arirang | Correlation of Burden | Correlation of Arirang Text | Information Algorithm |

* 본 연구는 2013학년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 수행되었다.
* 본 논문은 한국콘텐츠학회 ICCC2013 국제학술대회 우수논문입니다.
접수일자 : 2014년 02월 17일 심사완료일 : 2014년 03월 24일
수정일자 : 2014년 03월 17일 교신저자 : 김학용, e-mail : hykim@cbnu.ac.kr

I. 서 론

아리랑은 대표적인 한국의 민요로써 한민족 구성원이라면 누구나 아리랑을 알고 즐겨 부르는 것이 현실이다. 아리랑은 "아리랑" 또는 "아라리"와 같은 유사한 구절이 후렴에 들어있는 민요의 총칭으로 지역과 시대에 따라 다양한 선율과 내용이 진화되어 전승되고 있다[1]. 현재까지 전승된 아리랑의 종류와 곡수를 정확하게 알 수는 없지만 최소한 약 60여 종(version) 3천6백여 수(variation)에 이르는 것으로 추산하고 있다. 이 아리랑이 2012년 12월 유네스코(UNESCO) 인류무형문화유산으로 등재되었다.

아리랑 단어의 유래에 대해서는 다양한 설이 있는데, '아리랑 고개를 넘어간다'는 말을 근거로 지명인 것에 무게가 실리고 있다[2]. 이병도는 북조선에서 남쪽으로 내려오는 교통로인 자비령에 있는 하선령 고개를 지칭한다고 하였다[3]. 자비령은 낙랑의 남계를 이루는 중요한 곳으로 낙랑은 흔히 '악랑' 또는 '락랑'으로 발음되고 다시 '알라' 또는 '아라'를 소리 나는 그대로 적은 것이다. 따라서 아리랑 또는 '아라'는 '알라' 또는 '낙랑'으로 보아야한다고 주장하고 있다[3].

이에 대해 국어학자인 양주동은 고어의 진화 관점에서 아리랑의 유래를 분석하였는데, 아리랑이 고개를 지칭하는 지명임은 이병도의 주장과 일치한다. 그러나 양주동은 아리랑의 원음은 '아리령'이며 '령'자는 '嶺'자의 음이며 '아리'는 국내의 아리와 유사한 음을 가진 령이 다수 존재한다고 주장한다[4]. 대동여지도의 지명을 보면, 평안도 분산의 아이진(阿耳鎭), 경상도 기장의 아이봉수(阿爾熢燧), 충청도 영동의 어리산(於里山), 강원도 원주의 어로현(於路峴), 평안도 철산의 어랑산(於郎山), 평안도 삭주의 오리동(吾里洞), 함경도 함경의 오로촌(吾老村)은 모두 '아리', '어리', '오리'를 지칭하는 지명이다. 따라서 아리랑은 '아리'와 같은 지명을 가진 모든 고개를 지칭하는 말이라고 주장한다[4]. 이와 같이 연구결과를 볼 때 '아리랑'은 우리나라 도처에 있는 '아리'와 유사한 고개를 가진 지명을 일컫는 말이라고 결론을 지울 수 있을 듯하다.

대부분 아리랑은 본문과 후렴구로 구성되어 있는데, 후렴구의 형태는 3.3.4구, 3.4.6구, 5.5.4구, 또는 5.5.6구 등과 같이 그 낱말의 수를 중심으로 구분하고 있다[5]. 예를 들면, 본조아리랑의 후렴구는 '아리랑 아리랑 아라리요'이므로 3.3.4구의 대표적인 예이며, 긴 아리랑의 후렴구는 '아리랑 아리랑 아라리로구려'이므로 3.3.6구의 대표적인 예이다. 이외에도 강원도 아리랑과 진도아리랑은 각각 5.5.4구(아리아리랑 쓰리쓰리랑 아라리요)와 5.5.6구(아리아리랑 쓰리쓰리랑 아라리가낫네)를 대표한다[5].

아리랑은 원래 한 지방의 노래였던 것이 경기, 서도, 강원, 영남 등지에 걸쳐 다양한 내용과 곡조를 가지고 전국으로 확산된 노래이며 곧 민요이다[6]. 아리랑은 본문의 내용에 따라 크게 4가지로 분류할 수 있는데, 생활성, 사회성, 염정성(艶情性), 풍류성이 그것이다[7]. 생활성을 노래한 아리랑은 본조아리랑과 정선아리랑, 사회성을 노래한 아리랑은 본조아리랑과 별조아리랑, 염정성을 노래한 아리랑은 밀양아리랑, 강원도 아리랑, 본조아리랑, 풍류성을 노래한 아리랑에는 정선아리랑과 진도아리랑이 있다.

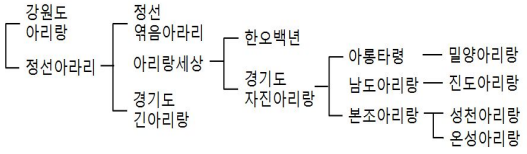

그림 1. 아리랑 곡조를 기반으로 한 아리랑의 음악적 계보[8]

1997년 한국민요집 제5집에 수록된 "아리랑 소리의 근원과 그 변천에 관한 음악적 연구"에서 이보형은 아리랑의 근원과 변천을 음악적 계보를 통해서 알아보았다[8]. 비록 본문과 전승지역이 다르더라도 곡조가 같은 것끼리는 하나의 범주로 묶어서 음악적 계보를 정리하였는데 아리랑의 근원은 '정선아라리'라고 하였다[그림 1]. '정선아라리'를 근원으로 '아리랑세상', '경기도 자진아리랑'을 거쳐 '아롱타령'을 연결고리로 경상도아리랑의 중심인 '밀양아리랑'으로 계보가 연결되었으며, '남도아리랑'을 거쳐 전라도 아리랑의 대표적인 '진도아리랑'으로 이어졌다. 끝으로 '본조아리랑'은 '성천아리랑'과 '온조아리랑'으로 그 계보를 잇고 있다. 최근에 우리

표 1. 한국문화콘텐츠닷컴에 실린 121개 아리랑 및 지역적 분류

지역	약어	아리랑 종류
서울 · 경기	SK1 ~ SK12	1)본조아리랑, 2)긴아리랑, 3)강원도아리랑, 4)한오백년, 5)여주아리랑, 6)양평아리랑, 7)수원 아리랑타령, 8)안성 아리랑타령, 9)양평(자탄)아리랑, 10)양평(자탄)아리랑2, 11)서울제 정선아리랑, 12)구아리랑
강원도	KW1 ~ KW23	1)정선아라리, 2)정선아라리2, 3)정선아라리3, 4)정선 엮음아라리, 5)정선 팔경아라리, 6)태백 아라레이, 7)태백아라리 8)횡성 어러리, 9)횡성 어리랑타령, 10)평창아리랑, 11)평창아라리, 12)명주 자진 아라리, 13)인제아라리, 14)인제 뗏목아리랑, 15)영월아라리, 16)강릉자즌아라리, 17)강릉자즌아라리2, 18)양양아라리, 19)소양강 뗏목아리랑, 20)양구얼러지, 21)홍천아리랑, 22)원주아리랑, 23)아리랑 뽕 따러 가세
충청도	CC1 ~ CC12	1)중원아라성, 2)단양아라리, 3)제천어러리, 4)증평아라리, 5)여주어리랑, 6)음성어러리성, 7)충주 아리랑타령, 8)제천 아리랑타령, 9)옥천아리랑, 10)서산아리랑, 11)대전 콩밭매는소리, 12)해방가 아리랑
경상도	KS1 ~ KS14	1)밀양아리랑, 2)경상도아리랑, 3)예천아리랑, 4)광복군아리랑, 5)울릉도 아리랑, 6)양산아리랑, 7)부산아리랑, 8)창원아리랑, 9)하동 사리랑타령, 10)문경새재 아리랑, 11)문경아리랑, 12)영천아리랑, 13)대구아리랑, 14)상주아리랑
전라도	JR1 ~ JR13	1)진도아리랑타령, 2)진도아리랑, 3)보길도 아리랑, 4)아리롱 서리롱, 5)아롱타령, 6)화개아리랑, 7)장흥아리랑, 8)해남아리랑, 9)무등아리랑, 10)완주아리랑, 11)무주아리랑타령, 12)신안아리랑, 13)고흥아리랑
제주도	JJ1 ~ JJ3	1)제주 조천아리랑, 2)제주도 아리랑(조천), 3) 제주 아리리리리 동동
북한	NK1 ~ NK22	1)영천아리랑, 2)경상도아리랑, 3)서도아리랑, 4)해주아리랑, 5)안주아리랑, 6)청진아리랑, 7)태평아리랑(성진), 8)이북아리랑, 9)아리랑, 10)잦은아리랑, 11)아라리요, 12)신아리랑, 13)삼일포 아리랑, 14)북간도 벌판, 15)이 내 신세, 16) 살피막 신세, 17)이 땅 이 거리, 18)단천아리랑, 19)온성아리랑, 20)무산아리랑, 21)통천아리랑, 22)고성아리랑
일본	JA1 ~ JA2	1)아리랑, 2)이츠키 자장가
중국	CH1 ~ CH13	1)쪽박의 아리랑, 2)기쁨의 아리랑, 3)아리랑 연곡, 4)장백의 새 아리랑, 5)아리랑 망향가, 6)정선아라리(중국), 7)광복군 아리랑, 8)얼쑤 아리랑, 9)아리랑, 10)강원도 아리랑, 11)강원도 아리랑2, 12)새 아리랑, 13)연변아리랑
미국	US1	1)상항아리랑
러시아	RU1 ~ RU3	1)사할린본조아리랑, 2)사할린본조아리랑2, 3)아리랑 엮음
독일	GR1 ~ GR3	1)아리랑 쓰리랑, 2)하리랑 노래, 3)아리랑가

민족 전체에 가장 널리 알려진 본조아리랑에서 아리랑의 근원을 찾으려는 노력이 있었으나 이 아리랑은 1920년대 당시에 가장 유행했던 아리랑인 경기 자진아리랑을 편곡하여 만든 것으로 알려졌는데[8], 음악적 계보를 통해서 분석한 결과와도 일치하고 있다.

지금까지 살펴본 바와 같이 대부분의 연구는 아리랑을 인문학적 접근방법에 따라 계통을 분류하고 아리랑 사이의 상관관계를 분석하였다. 유네스코 인류무형문화유산으로 등재된 것을 계기로 또 다른 접근 방법인 정보과학 및 기술적인 접근방법을 통해서 아리랑 사이의 상관관계를 분석하여 정보를 제공한다면 세계적인 문화유산으로서의 위상을 높이는데 크게 기여할 것으로 사료된다.

본 연구에서는 이러한 측면을 고려하여 정보과학에서 주로 사용하는 다중서열정렬(multiple sequence alignment) 알고리즘을 도입하여 후렴구를 중심으로 아리랑 사이의 계통도를 분석하여 아리랑들 사이의 진화관계를 분석하고자 하였다. 또한 분문에 등장하는 내용을 중심으로 네트워크 및 쌍대비교행렬(pairwise comparison matrix) 기법을 사용하여 아리랑 사이의 연관성을 분석하고자 하였다. 인류무형문화유산인 아리랑을 객관적이고도 과학적인 방법을 통해서 분석함으로써 아리랑의 가치를 한층 더 높이는 계기를 만드는데 본 연구의 의의가 있다고 하겠다.

II. 연구자료 및 연구 방법

1. 데이터베이스 구축

본 연구에 사용된 121개 아리랑가사는 한국문화콘텐츠닷컴(http://www.culturecontent.com/main.do)에 있는 음악분야 중에서 '겨레의 노래 아리랑'으로부터 얻었다[표 1]. 본 연구에서 사용한 121개 아리랑을 지역별로 정리하였으며 긴 제목의 아리랑을 표시하기 어려워 약어를 사용하였다. 예를 들어, 서울경기 지역의 아리랑은 약어인 SK(서울경기의 약어)를 앞에 쓰고 아리랑 종류에서 사용한 순서에 따라 번호를 부여하였다. SK1은 서울·경기지역 아리랑인 '본조아리랑'을 의미하며, NK22는 북한아리랑 중에서 '고성아리랑'을 의미한다[표 1].

121개 아리랑 본문에 등장하는 모든 단어 2,741개를 추출하여 아리랑 가사 네트워크를 구축하기 위한 노드

(node)로 사용하였다.

후렴구는 121개 아리랑가사 중에서 '한오백년'과 같이 후렴구가 없는 아리랑 15개를 제외한 106개 아리랑으로부터 선별하였다. 106개 아리랑의 후렴구에 등장하는 단어 중에서 최소한 5번 이상 등장하는 등장횟수 상위 20개의 단어를 추출하였다. 가장 많이 등장한 단어는 '아리랑'으로 221번이었으며 두 번째 많이 등장한 단어는 79번으로 '고개'였다. 5번 등장하는 단어는 '쓰리', '가세', '주게'등 3단어였다. 이 상위 20개 후렴구 단어의 집합을 기반으로 생명체의 진화를 분석하는 다중서열정렬 알고리즘을 적용하여 아리랑 계통 분류 체계 또는 아리랑 진화 계보도를 만드는데 활용하였다[그림 2].

2. 후렴구 분석 알고리즘 및 시각화

두 개 이상의 생명체를 DNA 염기 서열이나 단백질 아미노산 서열을 기반으로 진화적 또는 발생학적 관계를 분석하는 알고리즘은 생명정보학 분야에서 널리 이용되고 있다. 이 다중서열정렬 알고리즘은 비교하고자 하는 생명체의 염기서열을 가능한 모든 상관관계를 비교 분석하여 실제 진화적 관계를 가장 잘 설명하는 계통수를 찾아주는 알고리즘이다[9]. 후렴구에 최소한 5번 이상 등장하는 후렴구 단어를 중심으로 서열정렬한 결과를 다중성열정렬 알고리즘을 사용하여 도출하였으며 이 결과를 시각화하기 위하여 EBI(European Bioinformatics Institute)에서 제공하는 ClustalW2-Phylogeny 프로그램을 사용하였다(http://www.ebi.ac.uk/Tools/phylogeny/clustalw2_ phylogeny/).

3. 본문 내용 분석 알고리즘 및 시각화

아리랑 본문을 중심으로 상관관계 분석을 위해서 네트워크와 쌍대비교행렬 알고리즘을 도입하였다. 네트워크는 노드라 불리는 구성요소와 그들 사이의의 상호관계를 링크로 연결함으로써 복잡계에서 일어나는 복잡한 현상을 쉽게 이해할 수 있다. 아리랑 가사 네트워크는 아리랑에 등장하는 단어와 아리랑 제목을 노드로 하고, 각 아리랑 제목에 등장하는 단어를 연결하는 링크로 하여 bipartate 네트워크를 구축하였다. 따라서 아리랑 가사 네트워크는 두 개의 서로 다른 노드 (아리랑 제목과 그 아리랑에 등장하는 단어)와 링크로 구성되어 있다. 네트워크의 시각화는 싸이토스케이프(cytoscape) 프로그램을 사용하였다[10].

아리랑 본문을 중심으로 각 아리랑 사이의 상관관계를 분석하기 위해서는 하나의 아리랑에만 등장하는 단어는 아무런 의미가 없다. 이 단어는 네트워크상에서 링크가 하나 뿐인, 즉 연결계수(degree)가 1인 노드를 의미한다. 복잡한 아리랑 가사 네트워크[그림 3A]를 단순화하기 위하여 연결계수가 1인 노드를 제거하였다[그림 3B]. 추출한 핵심 아리랑과 단어를 쌍대비교행렬 알고리즘을 이용하여 상관관계를 분석하였다[11]. 쌍대비교행렬이란 집단 속에서 모든 가능한 짝을 비교하는 사용하는 방법으로 두 집단과 평균 간의 차이를 비교하는 알고리즘이다.

아리랑 CC1과 CC12에 동시에 등장하는 단어는 1개였기에 1로 표시하였고, JR1과 KS1에 함께 등장하는 단어는 12개였기에 12로 표시하였다[그림 4]. 그림 4에서 위의 가로축과 좌측의 세로축에 나열한 아리랑의 순서는 알파벳과 번호의 순으로 나타냈다. 따라서 쌍대비교행렬에서 먼저 CC1, CC10, CC11, CC12, CC2의 순으로 나열하였으며 마지막에는 SK8을 배열하였다. 배열한 총 아리랑 제목은 73개였다. 가장 많은 단어를 공유하는 것은 16이었으며 가장 낮은 단어를 공유한 것은 1이였다. 따라서 그림 4는 1에서부터 16까지 공유한 단어의 수를 직접 표시하였으며 동시에 검은색의 밝기를 16 단계로 나누어 상대적인 크기로 표시하였다.

네트워크 기반으로 아리랑들 사이의 상관관계를 분석하였는데, 구축한 아리랑 biparate네트워크에서 연결계수가 1인 노드를 단계적으로 제거하다보면 점진적으로 네트워크의 크기는 작아지고 핵심 단어와 핵심 아리랑으로 연결되는데, 이를 통해서도 아리랑들 사이의 상관관계를 분석하였다.

Ⅲ. 연구결과

후렴구에 의한 아리랑의 계통도를 분석해보면 크게 두 종류로 분류된다[그림 2]. 하나는 음률의 템포가 비

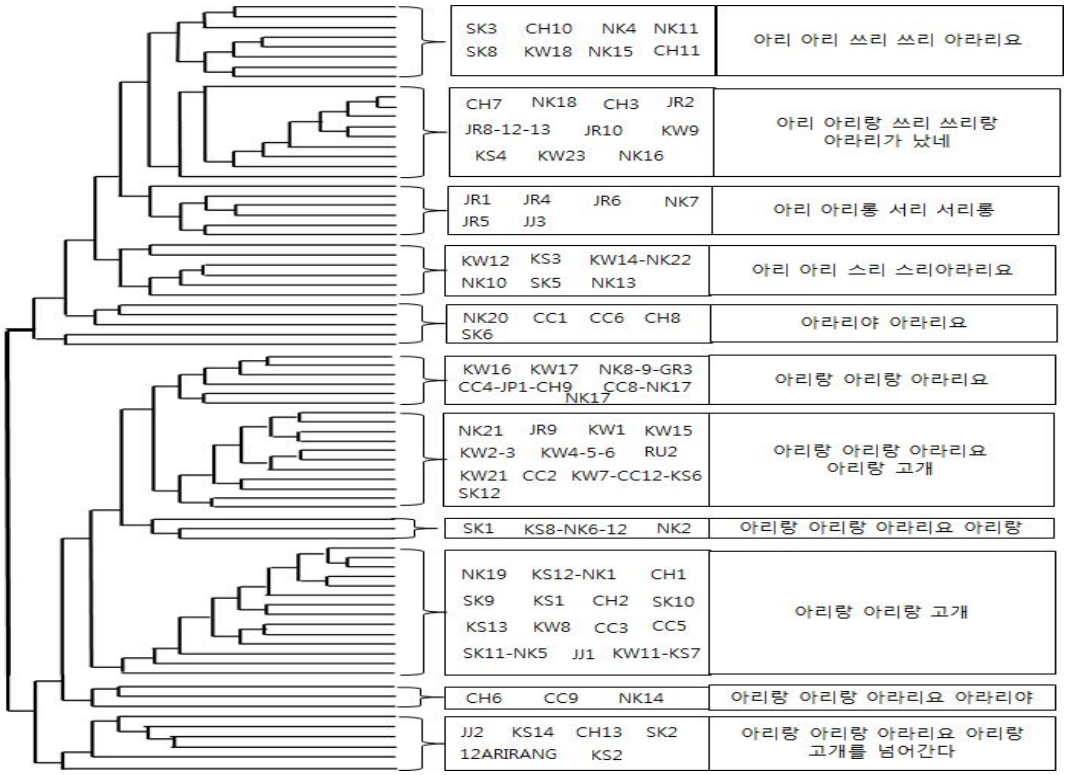

그림 2. 후렴구를 중심으로 계통 분류 알고리즘을 사용하여 분류한 아리랑의 계통도

교적 빠른 “아리 아리 쓰(스)리 쓰(스)리” 및 “아리 아리랑 쓰리 쓰리랑”이며, 다른 하나는 템포가 느린 “아리랑 아리랑 아리랑 아라리요”이다. 후렴구가 있는 아리랑 106개 중에서 34%인 38개 아리랑이 빠른 템포를 가지고 있었으며 나머지 66%인 68개 아리랑은 느린 템포를 가지고 있었다.

10개 이상의 아리랑을 보유하고 있는 지역은 서울·경기도, 강원도, 충청도, 경상도, 전라도, 북한, 및 중국이었다. 템포를 기준으로 분류해보면 크게 3부류로 나눌 수 있는데, 지역별 아리랑 노래 중 65%이상이 느린 템포를 가진 지역은 서울·경기도, 강원도, 충청도, 경상도였으며, 북한과 중국은 느린 템포와 빠른 템포가 약 50%였다. 유일하게 전라도 지역에 있는 11개 아리랑 중에서 9개 아리랑이 빠른 템포였으며, 나머지 2개 아리랑이 느린 템포를 가지고 있었다.

북한과 중국 아리랑은 다른 지역의 어느 아리랑으로부터 전해졌느냐에 따라 분포가 달라지기 때문에 전라도와 나머지 지역으로부터 고르게 받아드린 것으로 추정할 수 있다. 왜냐하면 북한과 중국에는 다른 지역의 이름을 가지고 있는 아리랑을 그대로 가지고 있기 때문이다. 예를 들어, 북한에는 경상도의 ‘영천아리랑’과 ‘경상도아리랑’을 가지고 있으며 중국에도 ‘강원도 아리랑’과 ‘정선아라리’를 가지고 있다.

전라도 지역의 아리랑의 대부분이 빠른 템포를 가지고 있는 아리랑(82%)인 이유가 이 지역이 비교적 넓은 평야를 가지고 있기 때문에 아리랑이 일을 하면서 빠르게 또는 신명나게 부르는 노동가 성격의 아리랑일 가능성을 알아보았다. 본문 내용을 조사한 결과 전라도 지역의 아리랑들이 노동가에 집중되어 있는 것은 아니었다. 따라서 전라도 지역의 이러한 분포가 노동가 성향을 띄고 있는 것은 아니라고 할 수 있다. 현재까지 3,600여 수가 알려져 있는데, 121개만을 대상으로 한 결과이

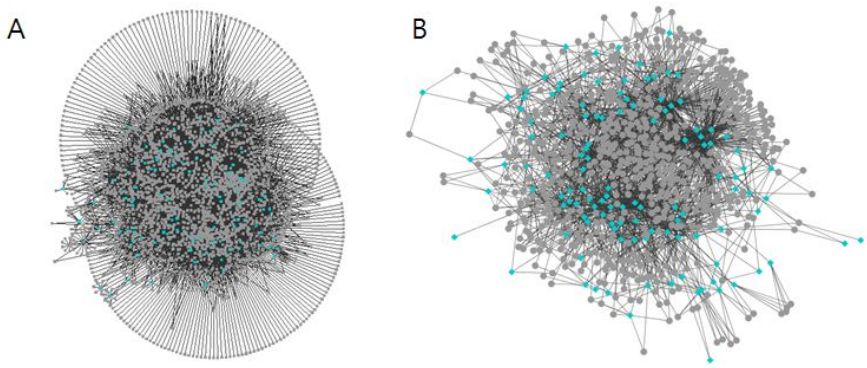

그림 3. 아리랑과 아리랑에 포함된 단어로 구성된 bipartate 네트워크. (A) 전체 네트워크는 2,862개 노드와 8,308개의 링크로 구성되어 있음. (B) 연결계수가 1인 노드를 제거한 축약 네트워크는 723개 노드와 4,030개 링크로 구성되어 있음.

기 때문에 아리랑 수를 확장하여 분석하더라도 같은 결과가 나온다면 다른 측면에서 원인을 찾을 수 있을 것이다.

본 연구에서 분석한 총 아리랑 수는 121개였으며, 본문에서 추출한 단어는 무려 2,741개였다. 본문에 등장하는 단어를 중심으로 아리랑 사이의 상관관계를 알아보기 위해 먼저 아리랑 제목과 그 아리랑에 등장하는 단어를 노드로 하고 아리랑에 포함하는 단어가 있는 경우를 링크로 연결하여 bipartate 네트워크를 구축하였다[그림 3].

한 아리랑에 단 한번 등장하는 단어가 있다(연결계수가 1인 노드)[그림 3A]. 이 경우 이 단어는 단 하나의 아리랑에만 존재하기 때문에 다른 아리랑과의 상관관계 분석에는 도움이 되지 않는다. 따라서 연결계수가 1인 노드를 모두 제거하고 비교적 단순하게 축약된 네트워크를 다시 구축하였다[그림 3B].

구축한 축약 네트워크의 노드 수는 723개였는데, 이는 121개의 아리랑과 602개의 본문 단어로 구성된 것이다. 이와 같이 단순화시켰음에도 불구하고 네트워크는 여전히 복잡하다. 이를 극복하기 위하여 한 단어가 최소한 5개의 아리랑에 나타나는 단어만을 선별하기 위하여 노드의 연결계수가 4이하인 단어를 제거하였다. 104개 본문 단어가 최소한 5개 이상의 아리랑에 등장하는 것으로 나타났다. 가장 많이 등장한 단어는 '임'으로 20개의 아리랑에 등장하였으며 그 다음으로 '큰애기'와 '님'이었다. '임'과 '님'은 동일한 의미를 나타내는 단어이지만 데이터베이스 있는 표현에 충실하고자 합치지 않았다. 최소 5개 아리랑에 동시에 등장하는 단어는'고향'을 포함한 31개 단어가 있었다. 이 조건에 충족한 본문 단어는 총 104개 단어였으며 이에 해당하는 아리랑은 73개였다. 73개 아리랑에 등장하는 104개 단어를 그룹(104개 그룹)으로 묶어 아리랑 사이의 상관관계를 쌍대비교행렬 알고리즘를 사용하여 분석하였다[그림 4].

쌍대비교행렬을 통해서 분석한 결과, 가장 많은 16개 단어를 공유하고 있는 아리랑은 '정선아라리'와 '진도아리랑타령'이었다[표 2]. 그 다음으로 공통단어가 14번 공유하고 있는 '정선아라리'와 '밀양아리랑'이 있으며 '인제아라리'와 '정선아라리'도 같은 수만큼의 단어를 공유한다. 표 2에는 두 아리랑 사이에 최소한 공유하는 단어가 9개 이상인 아리랑 쌍을 보여주고 있다. 이에 속하는 아리랑은 총 13개 아리랑으로 강원도 지역 아리랑이 대부분이며, 경산도 지역 아리랑은 '밀양아리랑'과 '예천아리랑'이 있으며, 전라도지역 아리랑은 '진도 아리랑타령'이 이에 속한다.

13개 아리랑 중에서'원주아리랑'을 제외하고는 이들 아리랑 가사가 다른 아리랑에 비해 비교적 길다. 아리랑 사이의 친밀한 정도를 분석함에 있어서 긴 아리랑 가사를 가지고 있기 때문에 공통으로 존재하는 단어가 많을 수밖에 없는 한계가 분명이 있다. 본 연구에서 아리랑 가사의 길이를 보정하지 않고 있는 가사 그대로를

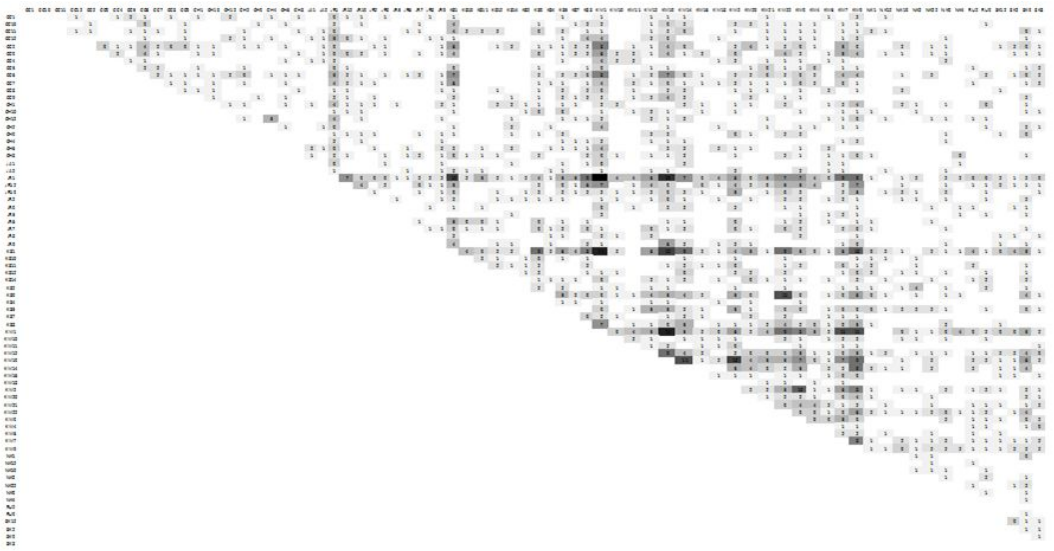

그림 4. 쌍대비교행렬을 사용하여 분석한 73개 아리랑에 등장하는 공통단어 기반 아리랑 사이의 상관관계 비교.

사용하여 분석하였다. 그 이유는 가사가 길기 때문에 서로 공유하는 단어가 많기는 하지만 공유하는 단어가 많다는 것은 같은 심정이나 생활을 노래한 것이고 결국 두 아리랑 사이에 친밀도(또는 밀접도) 정도가 높다는 것을 의미하기 때문이다.

비록 표 2에서는 제시하지 않았지만, 그림 4에서 제시한 비교적 낮은 밀접도를 가진 아리랑 사이의 상관관계도 역시 의미 있다고 하겠다. 이에 속하는 대부분 아리랑은 가사가 짧은 편이다. 가사가 짧기 때문에 공유하는 단어는 적지만 두 아리랑 사이의 상관관계를 분석하는데 충분하다. 예를 들어 그림 4에서 왼쪽의 CC2 및 상단의 CC3은 각각 '단양아라리'와 '제천어러리'를 표시한 것인데, 이 들 두 아리랑 사이에는 3개의 단어를 공유하고 있다. 따라서 밀접도가 높은 아리랑이 아니더라도 그림 4의 쌍대비교행렬 결과를 통해서 한 아라랑과 다른 아리랑들 사이의 상관관계가 어느 정도인지 분석할 수 있다.

비록 특정 지역이름으로 되어 있는 북한의 경상도아리랑(NK2)과 경상도의 경상도아리랑(KS2) 사이에는 공유하고 있는 단어가 "4"인데, 그 만큼의 밀접도가 있는 것을 의미한다. 이는 북한에서 즐겨 부르는 경상도 아리랑은 경상도 지역에서 유래했거나 경상도 지역 사람이 북한으로 이주하여 살면서 그 지역 특성에 맞게 변형시켜 부른 것으로 추정할 수 있다. 또 다른 예가 서울·경기도 지역의 강원도아리랑(SK3)은 3개 단어 이상을 고유하고 있는 아리랑의 종류만 14개다[그림 4]. 이 아리랑은 강원도 지역의 '정선아라리'를 포함하여 가장 많은 6개 아리랑과 관계가 있으며, 충청도 지역의 3개, 경상도 지역의 3개 아리랑과 관계가 있는 것으로 나타났다. 이외에도 중국의 '아리랑 연곡'과 북한의 '영천아리랑'과도 연관성이 있는 것으로 나타났다. 오히려 북한의 '영천아리랑'은 경상도의 '영천아리랑'과는 공유하는 단어가 2개로 이들보다 낮았다. 비록 후렴구는 정확하게 일치하지만[그림 2], 본문 가사에 공유하는 단어가 낮다는 것은 경상도 아리랑이 북한 지역으로 이동한 후 그 지역에서의 생활상이 가사에 새롭게 반영된 결과라고 할 수 있다.

또 다른 본문 내용을 중심으로 아리랑들의 상관관계를 분석하기 위해 핵심 네트워크를 사용하였다. 그림 3B에서 보는 바와 같이 연결계수가 1인 노드를 제거하더라도 여전히 복잡하다. 연결계수 1인 노드를 제거하면 또 다른 연결계수가 1인 노드가 새롭게 생성된다. 이는 아리랑 네트워크의 뭉침계수(clustering coefficient)가 0(zero)이기 때문이다. 이러한 구조적 특징을 가진

네트워크는 연결계수가 1인 노드를 단계적으로 제거한다면 비교적 쉽게 핵심 네트워크를 구축할 수 있다.

그림 5는 11단계를 거쳐 연결계수가 1인 노드를 제거하여 구축한 핵심네트워크를 보여주고 있다. 그림에서 보는 바와 같이 '쪽박의 아리랑'은 '님', '돈', 및 '산'을 노래하고 있으며, '음성어러리성'은 '꽃', '정든님', '낭군'을 노래하고 있다. 동시에 아리랑들 사이의 상관관계 분석이 가능한데, '제천 아리랑타령', '쪽박의 아리랑', '무등아리랑'은 '산'을 노래하고 '긴아리랑', '보길도아리랑', '음성어러리성'은 다 같이 '정든님'을 노래하고 있다. 만약 검색하고자하는 아리랑 제목이 없다면 연결계수 제거 단계를 앞으로 당겨 9 또는 10단계에서 분석할 수도 있으며 만약 현 단계가 너무 복잡하여 분석이 어려우면 더 많은 노드를 제거하여 12 또는 13단계에서 분석할 수 있을 것이다.

표 2. 본문으로부터 추출한 아리랑 사이의 밀접도 순위.

밀접도 순위	아리랑
1(16)*	정선아라리-진도 아리랑타령
2(14)	정선아라리-밀양아리랑
	인제아라리-정선아라리
4(13)	인제아라리-진도 아리랑타령
5(12)	밀양아리랑-진도 아리랑타령
	인제아라리-밀양아리랑
	정선아라리2-인제아라리
8(11)	인제 뗏목아리랑-인제아라리
	원주아리랑-예천아리랑
	태백아라리-정선아라리
	횡성 어리랑타령-정선아라리
12(10)	횡성 어리랑타령-밀양아리랑
	정선아라리2-정선아라리3
13(9)	창원아리랑-진도 아리랑타령
	태백아라리-진도 아리랑타령
	횡성 어이랑 타령-진도 아리랑타령
	예천아리랑-밀양아리랑
	인제 뗏목아리랑-밀양아리랑
	원주아리랑-밀양아리랑
	원주아리랑-정선아라리
	인제아라리-명주 자진 아라리
	횡성 어리랑타령-인제아라리
	횡성 어리랑타령-인제 뗏목아리랑

*괄호안의 숫자는 두 아리랑 사이에 공존하는 단어의 수를 의미

따라서 뭉침계수가 0(zero)인 이러한 네트워크에서는 연결계수가 1인 노드를 단계적으로 제거하여 원하는 아리랑을 네트워크 상에서 무엇을 노래하고 어느 아리랑과 밀접도가 높은지를 분석하는 도구로 활용할 수 있다.

Ⅳ. 결 론

지금까지 전승된 아리랑은 약 60여 종 3천6백여 수에 이르는 것으로 추산되나 정확하게 알 수는 없다. 이 모든 아리랑을 수집하여 데이터화한다면 그 용량은 빅데이터(big data) 수준일 것이다. 따라서 유네스코 문화유산으로 등재된 것을 계기로 빅데이터인 우리의 아리랑을 수집 정리하는 작업과 연구는 매우 중요하다.

향후 3,600여수의 아리랑이 수집되어 DB화 된다면, 아리랑을 중심으로 어원 중심의 근원 문제의 합리적인 정착, 음률 중심의 아리랑 사이의 계통도 분석, 아리랑 본문을 중심으로 아리랑 사이의 상관관계 분석 등에 관한 연구를 인문, 사회, 예술적인 접근 방법을 통해서 분석하는 것도 매우 중요한 연구 중의 하나이다. 동시에 수집될 아리랑을 세계문화유산에 걸맞고 세계인이 공감할 수 있는 정보과학 및 기술적인 방법으로 분석하는 연구도 매우 중요하다고 판단된다.

본 연구에서는 3천6백여 수에 이르는 아리랑이 수집되기 전에 먼저 한국문화콘텐츠닷컴에 수록된 121개 아리랑을 중심으로 정보과학 및 기술적인 방법을 도입하여 아리랑 사이의 계통도 및 상관관계 분석을 시도하였다. 이를 위해 도입한 방법은 복잡계를 분석하는 기술의 하나인 네트워크 분석[12], 생명체의 진화 관계를 찾아주는 다중서열정렬 알고리즘[9], 및 아리랑 본문 내용 중심의 상관관계를 분석하는 쌍대비교행렬 알고리즘[11]을 도입하였다.

아리랑 121개 중에서 후렴구가 있는 106개 아리랑에 나타나는 후렴구의 배열을 중심으로 생명체의 진화관계를 분석하는 알고리즘을 도입하여 분석하였다[그림 2]. 생명체의 진화를 분석하는 알고리즘은 아미노산 수인 20개 단어의 배열만을 분석할 수 있는 약점이 있다. 다시 말해, 20개 이상의 후렴구를 가지고 있는 아리랑의 계통 관계를 분석하기에 적절하지 못한 측면이 있다. 그러나 본 연구에서는 우선적으로 106개 아리랑만을 분석하고 향후 빅데이터로써 아리랑을 분석할 수 있는지에 대한 가능성을 타진하고자 하였기에 후렴구에 상대적으로 많이 등장하는 상위 20개 단어를 선발하고 배열하여 아리랑 사이의 계통도를 분석하였다.

그림 5. 11단계를 거쳐 노드를 단계적으로 제거하여 얻은 핵심 네트워크

흥미 있는 것은 아리랑은 후렴구를 중심으로 분류하였을 때 빠른 템포의 아리랑과 느린 템포의 아리랑으로 확연히 분류된다는 점이다. 빠른 템포의 아리랑은 다시 '아리 아리 쓰이 쓰이 아라리요'와, '아리 아리랑 쓰리 쓰리랑 아라리가 났네'로 나눌 수 있다[그림 2]. 당시 생활의 한과 애잔함이 묻어나는 느린 템포의 아리랑이 다수를 차지했으며 이러한 아리랑 템포는 지역적인 분포 특성을 보여주었다.

본 연구에서 분석한 106개 아리랑의 후렴구는 가능하지만 향후 DB화된 아리랑 수가 많아진다면 핵심 단어 20개로만 비교하기에는 충분하지 못할 것이다. 따라서 빅데이터로써 아리랑을 분석하기 위해서는 후렴구 단어 20개가 아닌 이 보다 더 많은 후렴구의 단어를 도입하더라도 분석할 수 있는 확장된 아리랑 분석 알고리즘의 개발을 요구하고 있다. 그러나 생명체 진화에 매우 유용하게 상용되고 수 십 년 동안 검정된 알고리즘이기에 보완한다면 향후 빅데이터로써의 아리랑도 충분히 분석할 수 있을 것이다.

본문 가사 중심의 상관관계 분석을 위해서 복잡계 현상을 분석하는 기술인 네트워크 기법을 먼저 적용하여 아리랑에 등장하는 단어를 중심으로 아리랑과의 관계를 분석하고자 하였다. 특히, 121개의 아리랑과 그 안에 등장하는 단어가 무려 2,741개이기 때문에 단순화할 필요가 있다. 먼저 아리랑과 그에 등장하는 단어를 노드로 하고 어떤 아리랑에 단어가 등장하면 링크로 연결하는 네트워크를 구축하였다[그림 3A]. 상관관계를 분석하는데 영향을 미치지 않는 연결계수가 1인 노드를 제거하여 비록 단순하지만 축약된 네트워크 구축[그림 3B]하여 본문 단어 중심으로 아리랑 사이의 상관관계를 분석하였다. 최소한 5개 이상의 아리랑에 등장하는 단어 104개를 선택하고 그 단어가 있는 아리랑들 사이의 관계를 쌍대비교행렬 알고리즘으로 표한하여 시각화 하였다[그림 4].

이 방법을 통하여 두 가지 상관관계 정보를 도출할

수 있었는데, 하나는 밀접도 상위 아리랑 사이의 상관관계이며 다른 하나는 최소한 연관이 있는 이리랑 사이의 연결고리를 찾을 수 있었다. 밀접도가 가장 높은 아리랑는 '정선아라리', '진도 아리랑타령', '인제아라리', '밀양아리랑'이었다. 우리나라 아리랑의 기원을 '정선아라리'에 있다고 볼 때[13], 이 아리랑을 중심으로 전라도 지역의 대표 아리랑인 '진도 아리랑타령'과 경상도 지역의 대표 아리랑인 '밀양아리랑'이 연결되는 것은 이미 잘 알려져 있으며[8] 본 연구 결과에서도 이를 뒷받침하고 있다.

본 연구에서는 두 아리랑 사이의 밀접도가 낮더라도 두 아리랑 또는 한 아리랑과 연관이 있는 다수의 아리랑들 사이의 상관관계에 관한 정보를 추가적으로 제공하고 있다[그림 4]. 특히, 공유하고 있는 단어의 수를 제시함으로써 상관관계 빈도를 정량적으로도 분석할 수 있도록 하였다.

복잡계 분야에서 복잡한 현상을 분석하는 방법으로 네트워크 방법을 도입하고 있다[14]. 데이터의 양이 많은 경우 너무 복잡하여 분석이 어려울 때 네트워크를 단순화시켜 핵심 네트워크를 구축하거나[15] 모듈 구조를 도출하여 복잡한 현상을 분석하려는 연구가 진행되고 있다[16]. 이러한 핵심 네트워크나 모듈을 도출하는 방법으로 널리 사용되는 알고리즘이 'K-코어' 알고리즘[17], MCODE(Molecular COmplex DEtection) 기법[18], SNN(Shared Near Neighbor) 알고리즘[19]등이 있다. 본 연구에서 뭉침계수가 0(zero)인 네트워크에서 연결계수가 1인 노드를 단계적으로 제거함으로써 핵심 네트워크를 구축할 수 있는 새로운 방법을 제시하였다.

2012년 12월에 등재된 세계문화유산인 아리랑이 향후 세계 속의 아리랑과 세계인의 아리랑으로 발돋움하기 위해서는 아직도 발굴하지 못한 아리랑을 발굴하여 빅데이터로써의 아리랑 자료 축적과 인문학적 분석이 필요하다. 동시에 본 연구에서 제시한 방법을 수정 보완하여 빅데이터를 분석할 수 있을 정도로 끌어 올리는 연구가 추가적으로 진행되어야 할 것이다. 본 연구는 그 동안 어원 중심, 근원 중심, 음율 중심의 아리랑 분석을 탈피하여 정보과학 및 기술적인 기법을 도입하여 분석할 수 있는 새로운 방법을 제시하고 관련 정보를 도출하고 제공하는데 의의가 있다.

참 고 문 헌

[1] 조규익, 조용호, *아리랑 연구총서 1*, 도서출판 학고방, 2010.

[2] 임동권, "아리랑의 기원에 대하여", 한국민속학, 제1집, 1969.

[3] 이병도, "아리랑 곡의 유래", 개척, 제86호, p.5, 1956.

[4] 양주동, "도령과 아리랑", 민족문화, 제4권, 제2호, 민족문화사, 1959.

[5] 최재억, *한국민요연구-아리랑 민요고*, 민족문화, 광운대, 1970.

[6] 고정옥, *조선민요연구*, 수선사, 1949.

[7] 임경화, "민족에서 민족으로 가는 길", 동방학지, 제163집, pp.261-288, 1969.

[8] 이보형, "아리랑소리의 근원과 그 변천에 관한 음악적 연구", 한국민요학 제5집, pp.81-120, 1997.

[9] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson, "Multiple sequence alignment with the clustal series of programs," Nucl. Acids Res. Vol.31, No.13, pp.3497-3500.

[10] http://cytoscape.org

[11] Y. M. Wang, Y. Luo, and Y, S. Xu, "Cross-weight evaluation for pairwise comparison matrices," Group Decision and Negotiation, Vol.22, No.3, pp.483-497, 2013.

[12] Linton Freemann, *The Development of Social Network Analysis,* Empirical Press, 2006.

[13] 이용식, "강원도 아라리의 음악적 특징과 원형적 특질", 한국민요학, 제25집, pp.225-251, 2009.

[14] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," Nature, Vol.393, No.6684, pp.409-410, 1998.

[15] 김학용, "조선왕조 가계 인물 네트워크", 한국콘

텐츠학회논문지, 제12권, 제4호, pp.476-484, 2012.
[16] M. E. J. Newman, "Modularity and community structures in networks," Vol.103, No.23, pp.8577-8582, 2006.
[17] J. I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani, "K-core decomposition: a tool for the visualization on large scale networks," eprint cs.NI/0504107, 2005.
[18] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," BMC Bioinformatics, Vol.4, No.1, p.2, 2003.
[19] R. A. Jarvis and E. A. Patrick, "Clustering using a similarity measure based on shared near neighbors," IEEE Trans. Comp, Vol.C-22, No.11, pp.1025-1034, 1973.

저 자 소 개

김 학 용(Hak Yong Kim) 종신회원

- 1985년 2월 : 충북대학교 농화학과(농학사)
- 1987년 2월 : 충북대학교 화학과(이학석사)
- 1994년 5월 : 미국 코네티컷대학교, 분자세포생물학과(이학박사)
- 1998년 3월 ~ 현재 : 충북대학교 생화학과 교수

<관심분야> : 시스템생물학, 단백질네트워크, 생체동역학, 사회네트워크