위키피디아 인물 아카이브 서비스 개선을 위한 분석 연구*

Improving the Biography Archive Service of Wikipedia

최 상 희 (Sanghee Choi)**

목 차

초 록

인물에 대한 기록정보는 사회의 주요 분야에서 특정기준에 맞는 유명한 인물에 한정하여 정보를 수집, 가공, 제공하는 인물데이터베이스 형태가 일반적이었으나 최근 위키피디아와 같이 이용자들이 참여하여 다양한 인물에 대하여 자유롭게 서술하며 디지털 아카이브로 축적하는 체제가 활성화되고 있다. 이 연구는 위키피디아 바이오그래피 포털에서 범죄자, 교수, 영화감독 카테고리에서 인물 유형별로 500건의 데이터를 각각 수집하여 서술된 내용간 유형별 차별성이 있는지 계량적으로 분석하였다. 용어의 빈도 분석과 차별지수 분석을 수행한 결과 차별지수가 각 유형별로 특화되어 있는 내용을 표현하는데 효과적인 것으로 나타났다. 이 연구에서는 차별지수값이 높은 상위 100개의 용어와 세 유형에 공통적으로 출현한 용어 고빈도어 100개를 워드 클라우드 형태로 활용하여 특정 유형의 인물에 대하여 서술하는 이용자와 이를 승인하는 에디터가 참조할 수 있는 가이드를 제시하고자 하였다.

ABSTRACT

Biographical information about people is usually collected and provided by a company or an institute which has a specific standard to select people for service. Recently, user oriented contents service like Wikipedia has started biographical information service, Wikipedia Biography Portal, in which users select people and freely describe about them. This study collected 500 biographical data from three categories of Wikipedia biography portal such as criminals, faculty, and directors. The contents of data from each category were analyzed with the word frequency and the divergence indicator to identify the characteristics of each category. As a result, divergency indicator is effective to represent the differential factors of each category. This study provides word clouds of top 100 word with divergence indicator and top 100 common words of three categories with word frequency as a guide for users to write about a person in these categories and for editors to accept and monitor the biography from users.

키워드: 인물기록, 위키피디아, 내용분석, 계량분석, 차별지수
Biographical Information, Wikipedia, Content Analysis, Bibliographic Analysis, Divergence Indicator

---

* 이 논문은 2015년도 대구가톨릭대학교 교내연구비 지원에 의한 것임.
** 대구가톨릭대학교 도서관학과 부교수(shchoi@cu.ac.kr)
논문접수일자: 2018년 1월 31일 최초심사일자: 2018년 1월 31일 게재확정일자: 2018년 2월 12일
한국문헌정보학회지, 52(1): 447-467, 2018. [http://dx.doi.org/10.4275/KSLIS.2018.52.1.447]

1. 서 론

인물기록은 사회와 역사를 연구하는데 기초적인 자료로서 매우 중요한 사회적 가치를 가지고 있는 기록자료이다. 초창기 인물기록은 전기나 출생기록같이 독립된 단위로 생산되고 축적되어 왔었으나 데이터베이스 기술이 적용되면서 인물 데이터베이스 형태로 발달 되어 왔다. 인물데이터베이스는 인쇄형태로 발행했던 인명록을 데이터베이스로 구축한 사례가 일반적이서 인물에 대하여 기술하는 항목이나 서술방식이 한정적이며 특정 유형의 인물만이 가지는 특화된 요소를 반영하고 있지 못하는 경우가 많다. 예를 들어 야구선수와 같은 인물 유형에 대하여 사람들이 알고 싶어 하는 요소와 국회의원에 대하여 사람들이 알고 싶어 하는 요소는 차이가 많은데 대부분 인물데이터베이스에서는 두 집단의 공통적인 요소 중심으로 정보가 제공되고 있어 이용자는 특정 인물이 가지고 있는 특화된 정보를 제공받지 못하고 있다. 또한 인물데이터베이스는 인물을 선정하는 기준도 다양하지 못하고 데이터를 쉽게 수집할 수 있는 집단으로 편중되어 있어 사회의 다양한 분야에서 이슈가 되는 인물기록을 찾기 쉽지 않고 업데이트가 제대로 이루어지고 있지 못한 점도 문제점으로 지적되고 있다(한상길 2008).

이러한 문제점을 보완하는 방안으로 기록정보 관리를 중심으로 하는 아카이브 체제에서 이용자가 직접 참여하여 기록을 기술하고 공유하는 형태로 전환하는 아카이브 2.0이 언급되기 시작했다. 아카이브 2.0은 기록정보를 보존하고 관리하는 체제에서 이용자가 기록을 생산하고 소비하는 체제로 전환하는 것이 핵심개념이며 이용자 중심성과 개방성이 가장 중요한 요소이다. 이러한 요소를 적용하여 가장 효과적으로 운영될 수 있는 형태의 아카이브 2.0 서비스는 위키 형태의 서비스인데 이는 위키 형식의 기록정보서비스가 이용자 참여를 극대화할 수 있기 때문이다(김유승 2010).

인물 디지털 아카이브 서비스에서도 위키 형식의 서비스가 시행되고 있는데 대표적인 것이 위키피디아의 바이오그래피 포털(Biogrpahy portal)이다. 기존의 위키피디아 데이터에서도 인물이 차지하는 비중은 20%가 넘어 인물은 이용자들이 자발적으로 서술하는 중요한 항목이었다. 위키피디아에서는 인물기록을 따로 분류하여 포털의 형태로 제공하고 있다. 바이오그래피 포털에서는 인물기록이 가지고 있는 중요한 요소인 국적, 성별, 직업 등 주요 패싯을 대분류로 하여 인물기록을 분류, 유형별로 인물기록을 제공하고 있다. 그러나 이용자들이 다양한 인물을 자유롭게 기술하는 형태로 인물기록 서비스가 운영되고 있어 특정 유형의 인물기록에서 중점적으로 다루어져야 할 요소나 공통적으로 서술되어야 할 요소들이 무엇인지 파악하기 매우 어려운 문제점이 있다. 따라서 인물기록을 서술하는 이용자나 또는 이용자가 서술한 인물기록을 승인하는 에디터도 특정 유형의 인물에서는 어떠한 사항들이 중점적으로 기술되어야 하는지 알 수 없어 내용을 보완하거나 평가하는데 어려움이 따르고 있다. 특히 이와 같은 문제점은 항목별로 인물을 기술하는 형태가 아닌 서술형 인물기록에서 나타나고 있다. 서술형 인물기록은 위키피디아 인물기록과 같이 특정 인물에 대하여 다양한 이용자들이 자발적으로 자유롭게 기술하는 형태의 인물기

록이라고 할 수 있다.

이 연구에서는 위키피디아 바이오그래피 포털의 3개 카테고리인 범죄자, 교수, 영화감독에서 각각 500건의 인물기록을 수집한 후 각 유형별로 특화되어 있는 서술 요소들이 있는지를 파악하기 위하여 용어빈도와 차별지수를 적용하여 분석하였다. 또한 세 유형에서 공통적으로 나타나는 용어를 분석하여 인물기록에 대하여 서술할 때 보편적으로 활용할 수 있는 요소가 있는지도 파악하였다. 분석한 결과는 워드클라우드 형태로 제공하여 특정 유형의 인물에 대하여 서술하는 이용자와 이를 평가하는 에디터 모두 참조할 수 있는 도구를 제시함으로써 이용자가 참여하는 서술형 인물기록 서비스를 개선하는 방안을 제안하였다.

2. 연구배경

2.1 선행연구

인물기록에 관하여 수행된 국내 연구는 인물기록서비스의 전반적인 현황보다는 특정 인물이나 특정 사건에 관련하여 수행된 경우가 많았다. 특정 인물에 대한 기록을 지역의 로컬리티 콘텐츠 개발의 측면에서 발굴하여 컬렉션으로 발전시키는 연구가 수행되었는데 부산지역의 로컬리티를 대표하는 인물로 박기종이라는 특정인을 선정하여 관련된 사건과 공간 분석을 통해 설정한 주제 영역을 기반으로 지역사 인물 콘텐츠를 설계한 사례이다(현문수, 김동철 2013). 이 연구에서는 특정 인물의 생애사를 기술하는 것에서 확장하여 주요 활동을 선별하여 콘텐츠로 개발하고 연계하는 안을 제시하였다. 이와 유사하게 지역사에 대한 기록을 인물 중심으로 개발하는 연구로서는 경기지역 역사인물의 문화 콘텐츠화를 연구한 사례가 있었으며(김흥식, 김진형 2011) 부산지역의 로컬리티를 기록화하기 위해 부산의 인물 기록을 시대와 주제로 분석한 사례가 있었다(송정숙 2012). 해외에서는 2차 세계대전과 같은 특정 사건에 참여한 군인 직군으로 한정하여 인물기록에 대하여 연구를 수행한 사례가 있었으며(Leskinen 2017) 지역사를 이해하기 위해 인물을 선정하여 기록을 연구한 사례가 많았다(Keith 2017; Thomson 2016; Connor 2014).

상업적 인물데이터베이스를 비교 평가한 연구는 인물데이터베이스가 인터넷을 통해 본격적으로 서비스 되기 시작했던 시기인 2000년대 초 수행되었는데 당시 한국 데이터베이스 진흥센터에서 발행하는 한국의 데이터베이스 목록에서 인물/기관 정보로 분류되어 있는 데이터베이스 중 인물 데이터베이스 15종과 조선일보, 연합뉴스, 여성인명사전, KOLIS 법조인명록 4종을 합하여 총 19종의 인물데이터베이스를 최신성, 포괄성, 완전성, 다양성 등으로 평가하였다(장혜란 2001). 이 형태의 인물데이터베이스는 특정 기관이 기준을 가지고 인물을 선정한 후 인물에 대하여 수집한 인쇄기록을 전자화하는 과정에서 운영되고 있는 형태이며 가장 큰 문제로는 내용의 업데이트에 대한 이슈와 인물 선정의 다양성인 것으로 나타났다.

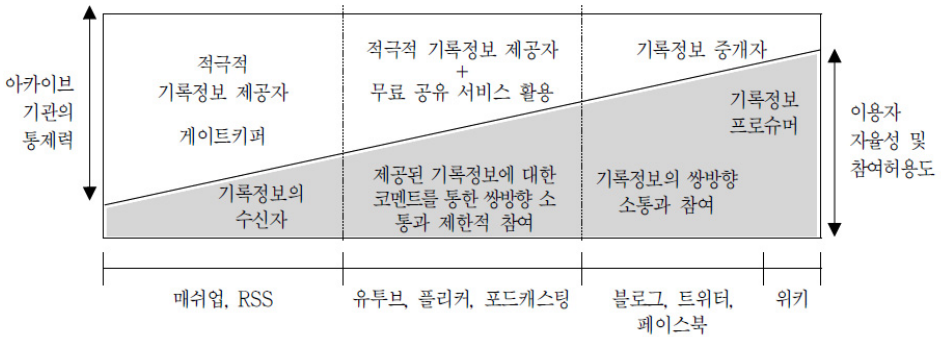
이와 연관된 후속연구로는 2008년 국내 언론사에서 구축운영하는 인물데이터베이스와 네이버에서 구축한 인물데이터베이스의 신규인물 선정과 업데이트체계에 대하여 분석하고 검색

방식을 비교한 연구가 있다(한상길 2008). 이 연구에서도 대상인물을 특정기관에서 선정하다보니 대상인물의 다양성이 부족한 것이 문제점으로 지적되었고 정보 업데이트의 한계에 대하여 지적하였다. 또한 인물을 기술하는 항목이 고정되어 있어 인물의 특성을 다양하게 기술할 수 없는 것이 문제점으로 지적되었다. 도서관에서 활용을 위해 인물데이터베이스를 평가한 최근 해외 연구에서도 검색방법과 최신성, 신뢰도에 대한 문제점이 있는 것을 지적한 바가 있다(Soules 2012).

이와 같은 인물선정의 다양성 이슈와 업데이트 문제를 해결할 수 있는 방안으로 웹 2.0을 적용하여 기록서비스를 확장하는 목적으로 다양한 연구가 수행되었는데(설문원 2010; 남재우, 김성희 2009) 이러한 연구들을 정리하여 아카이브와 웹 2.0을 결합한 아카이브 2.0의 개념을 정의하고 이에 대한 서비스 방식에 대한 논의를 한 사례가 있다(김유승 2010). 이 연구에는 아카이브 2.0의 핵심 개념을 이용자 중심성과 개방성으로 언급하였고 아카이브 2.0 적용기술 유형을 '매쉬업과 RSS', '유투브, 플리커, 포드캐스팅', '블로그, 트위터, 페이스북', '위키'로 유형을 나누고 이용자의 자율성과 참여 허용도 연관성을 〈그림 1〉과 같이 제시하였다.

〈그림 1〉에서 제시한 분석 결과에 따르면 위키방식의 기록정보서비스는 제한된 인적 자원과 지식의 한계를 이용자를 통해 극복하기 위한 방안(설문원 2010, 30)이라고 언급하면서 아카이브 2.0 핵심 요소인 이용자 중심성과 개방성에 가장 부합하는 서비스는 위키라고 제시하였다. 이러한 위키의 특성은 인물 선정의 다양성 문제가 있고 업데이트의 한계가 있는 인물기록 서비스에서는 대안이 될 수 있는데 이용자가 직접 인물을 선정하고 이용자가 생각하고 있는 인물의 의미를 서술할 수 있고 이를 다른 이용자가 수정, 보완할 수 있는 형태이기 때문이다.

이와 같은 맥락에서 해외에서는 위키피디아의 인물기록에 대한 연구가 수행되기 시작했는데 위키피디아의 인물 기록 서술에 대한 문제점을 분석한 연구(Ofek and Rokach 2015)와 인물기록 서술에 대한 신뢰도에 대한 문제점을



〈그림 1〉 기술 및 어플리케이션에 따른 기록정보 서비스와 이용자 참여 유형 및 허용도의 연관성 (김유승 2010, 41)

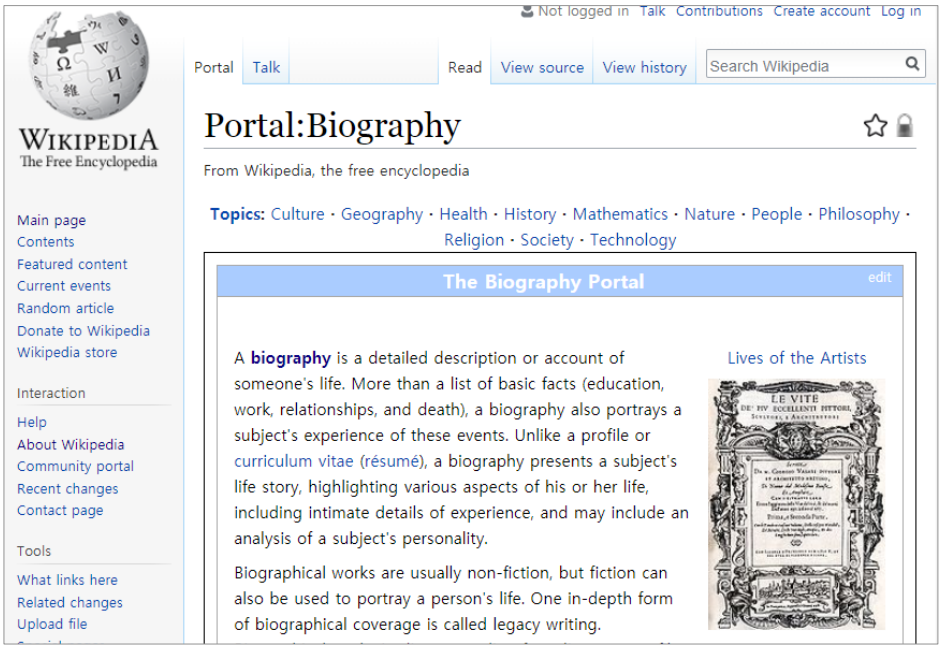
분석한 연구(Callahan and Herring 2011)가 있다. 특히 Ofek의 연구에서는 위키피디아에서 의미있는 인물을 선정하여 서술하라고 가이드는 주고 있지만 실제 인물기록을 작성하는 이용자들이 쉽게 참조할 수 있는 수준이 아니며 실제 서술한 인물기록이 승인되거나 승인되지 못하는 기준도 모호한 것으로 나타났다. 또한 승인을 하는 에디터들도 인물의 특성을 반영하여 승인할 수 있는 방안이 없는 것으로 나타났다.

이러한 문제점을 고려하여 이 연구에서는 위키피디아에서 인물기록을 대상으로 이용자가 참여하는 디지털 아카이브 서비스를 운영하는데 인물기록을 서술하는 이용자와 이를 승인하는 에디터가 인물의 유형을 반영하여 어떠한 기록들이 서술되어야 하는지 참조할 수 있는 방안을 제시하고자 한다.

2.2 위키피디아 바이오그래피 포털(Wikipedia Biography Portal)

위키피디아는 2001년부터 이용자들이 직접 참여하여 지식을 공동생산하는 대표적 참여형 인터넷 정보원으로서 다양한 분야의 지식이 축적되고 있다. Wikipedia 통계 사이트에 의하면 영문 위키피디아 기준(2018년 1월) 대략 5,560,000건의 콘텐츠 중에 약 26%의 콘텐츠가 인물에 대한 것으로 인물기록에 대한 비중이 상당히 높은 것으로 나타났다. 이에 위키피디아는 바이오그래피 포털(Biography Portal)을 별도로 운영하고 있는데 〈그림 2〉와 같이 위키피디아의 인물에 대한 항목들을 별도의 카테고리로 서비스하고 있다.

위키피디아 바이오그래피에서 이용자에게 서



〈그림 2〉 Wikipedia Biography Portal

술을 요청하는 인물군은 'a person', 'an actor', 'an animator', 'an artist', 'a bishop', 'a business person', 'a diplomat', 'an educator', 'an engineer', 'a lawyer', 'a military person', 'a musician', 'a photographer', 'a saint', 'a politician', 'a scientist', 'a sports person', 'a writer' 등이다. 배우나 외교관처럼 특정 직군에 해당하는 인물군이 대부분이지만 'a person'과 같이 이용자들에게 화제가 되거나 이용자가 기록할만한 인물을 모두 포함시킬 수 있도록 허용하여 실제 모든 인물을 대상으로 하고 있다.

제공되고 있는 인물기록의 카테고리를 살펴보면 대부분의 카테고리 하위분류로 국가와 민족분류가 적용되어 있어 국제적으로 다양한 인물들을 대상으로 기록이 서술되고 있는 것을 알 수 있다. 위키피디아 바이오그래피 포털에서 제공하고 있는 인물기록 카테고리의 대분류는 'By association', 'By behavior', 'By ethnicity', 'By gender', 'By language', 'By nationality', 'By occupation', 'By place', 'By political orientation', 'By religion', 'By status', 'By time' 등이다. 대분류는 인물이 가지고 있는 속성으로 패싯분류에 해당하는 기준이며 각 카테고리에 속한 인물들은 중복되어 하위분류되어 있다. 각 패싯에 해당하는 대분류는 컨테이너 카테고리라고 하며 실제 인물이 분류되지는 않고 하위분류로 찾아갈 수 있는 접근점 역할만을 하고 있다. 대분류 아래 하위분류는 기본적으로 다시 국가나 민족, 성별 등 다른 패싯개념이 중분류로 나타나고 있으며 해당 대분류 특성에 맞는 주제 카테고리도 포함하고 있다(〈그림 3〉 참조).

〈그림 3〉 위키피디아 바이오그래피 카테고리

바이오그래피 포털과 연관하여 위키피디아에서 운영하는 연관 프로젝트는 'Arts and entertainment', 'Actors and Filmmakers', 'Composers, Living persons', 'Philosophers', 'Saints', 'Romance authors', 'Military', 'Peers and Baronets', 'Politics and government', 'U.S. Members of Congress', 'U.S. presidents', 'Royalty and nobility', 'Science and academia', 'Aerospace', 'Sports and games', 'Baseball players' 등이 있어 점차 인물기록에 대한 내용이 다양하게 발전되고 있으나 각 인물 유형별로 특화된 가이드나 검색방식을 제공하고 있지는 못하고 있다.

3. 연구데이터 및 연구방법

이 연구에서 수집한 인물기록은 위키피디아의 바이오그래피 포털에 축적되어 있는 서술형 인물기록으로서 각 유형별로 기술된 기록의 특성을 살펴보기 위하여 범죄자(Criminals), 교수(Faculty), 영화감독(Directors) 등 3개의 카테고리에서 랜덤으로 각각 500인의 인물기록을 수집하여 총 1,500명의 인물기록을 분석하였다. 선정한 3개의 카테고리는 상대적으로 인물의 특성상 공통된 요소가 적은 인물집단으로 500건의 데이터를 랜덤으로 추출할 수 있을 정도로 축적된 인물 건수가 많은 유형이다. 500건 이상의 수집된 서술형 인물기록은 평균 6,240자였으며 유형별로 수집된 인물기록의 평균, 최대, 최소 문자수는 〈표 1〉과 같다.

각 유형별로 특성을 살펴보면 평균적으로는 범죄자 유형에 서술된 인물기록이 다른 유형에 비해 2배에 가까운 문자수를 나타내고 있어 평균적으로 범죄자 유형의 인물기록이 가장 양이 많은 것으로 조사되었다. 그러나 최대 문자수를 보면 영화감독이 가장 많고 최소 문자수에서도 가장 적은 것으로 조사되어 영화감독이 편차가 가장 큰 것으로 나타났다.

수집된 데이터는 Porter Stemming 알고리즘으로 용어를 추출하였고 빈도수 100개 이상인 용어 중 of, the 등 특정 의미를 담고 있지 않는 불용어, 인명, 지명 등을 제외하고 용어를 추출하였다. 추출된 용어는 유형별로 고빈도순으로 분석하여 상위 50개의 고빈도 용어를 산출하였다. 상위 50개 고빈도어에서는 해당 유형의 주제적 특성을 나타내는 용어와 일반 개념을 나타나는 명사로 분류를 하였는데 대학교수 1명, 대학원생 2명, 학부생 2명 등 총 5명이 각각 주제성 평가를 하여 분류를 하였다. 5명 모두 일치하거나 1명만 다른 의견을 나타나면 다수의 의견으로 주제성의 유무를 지정하고 의견이 다른 인원이 2명 이상으로 나타나면 주제성 유무가 모호(Y/N)인 것으로 판단하였다.

〈표 1〉 유형별 수집데이터 문자 수

	평균 문자수	최대 문자수	최소 문자수
범죄자	9392	69747	339
교수	4398	51760	335
영화감독	4929	72632	279

고빈도분석 외에 다른 유형에서보다 특정 유형에서 상대적으로 빈도수가 높은 용어를 분석하였다. 이러한 용어를 차별어라고 하는데 차별어는 특정 군집의 주제를 파악하기 위해 다른 군집에서보다 특정 군집에서 많이 출현하는 키워드를 분석하였는데 차별어가 특정 군집의 주제를 파악하는데 효과적인 것으로 나타났다(이재윤, 김수정 2016). 이 연구에서도 차별어의 개념을 적용하여 특정 유형별 인물기록에서 많이 나타나는 용어가 있는지를 조사하여 차별어가 특정 유형의 인물기록 특성을 표현하는지 분석하였다. 차별지수는 유형별 인물기록에서 출현한 용어의 상대적 빈도 차이를 산출한 것으로 다이버전스(divergence) 개념을 적용하여 측정하였다. 다이버전스 개념을 적용하여 차별어를 도출한 연구는 2013년 수행된 '정치와 언어의 관계에 대한 양적 분석 시론' 연구에서 사용한 방법을 차용하였다(김하수 외 2013). 이 연구는 대통령 선거에서 특정 후보가 다른 후보들에 비해 더 많이 사용한 어휘(차별어)를 추출하여 분석하여 특정 후보의 특성을 파악하고자 한 연구이다. 이 연구에서 적용한 다이버전스 개념은 확률 분포 간의 차이를 측정하는 방법으로서 Kullback-Leibler 다이버전스 개념을 응용한 것이며 차별지수를 산출한 공식은 이재윤이 제시한 다이버전스 공식을 사용하였다(이재윤 2007).

$$D_A(q_A||r) = q_A(W) \times \log \frac{q_A(W)}{r(W)}$$

위 차별지수 공식은 KL-Divergence를 사용하여, 세 유형의 텍스트에서 출현한 용어들을 대상으로 각각의 용어가 각 유형의 인물기록에서 출현할 확률들의 평균값과 특정 유형의 인물기록에서 출현할 확률값을 산출하여 이 두 값의 차이를 계산한 것이다. 용어 W가 유형 A의 인물기록에서 출현할 확률 $q_A(W)$와 세 유형을 합한 전체 인물기록에서 출현할 확률의 평균 $r(W)$을 적용하여 차별지수 값인 $D_A(q_A||r)$이 계산되는 것이다(이재윤 2007).

서술형 인물기록을 분석하면서 각 유형별 특성도 조사하였지만 인물기록라는 보편적 속성을 표현할 수 있는 내용도 서술되고 있는지를 파악하기 위하여 각 유형별로 나타난 용어 중 3개 유형에 공통적으로 나타난 용어 중 빈도 10 이상 용어 중 상위 100개의 용어를 추출하여 공통어로 제시하였다.

각 유형별로 추출된 차별어와 세 유형에 모두 나타난 공통어는 워드 클라우드로 표현하여 서술형 인물기록을 유형별로 내용상 특성을 파악하는 도구로 제시하였는데 워크 클라우드는 인터넷상에서 무료로 제공되는 워드 클라우드 작성도구인 Worditout(https://worditout.com/word-cloud/create)을 사용하였다.

4. 유형별 인물기록 계량적 내용분석

4.1 인물기록 유형별 고빈도어

4.1.1 유형 1 - 범죄자 고빈도어

범죄자 카테고리에서 수집된 500인의 서술형 인물기록에서 추출된 총 용어수는 31,357개이며 이 중 불용어를 제외하고 빈도수 100개 이상으로 추출한 키워드는 총 338개이다. 추출된

338개의 용어 중 상위 50개의 용어를 살펴보면 〈표 2〉와 같다. 상위 10위까지 랭크된 고빈도어를 살펴보면 'murder', 'death', 'kill', 'convict', 'sentence', 'police', 'trial' 등 범죄와 재판에 관련된 용어가 포함되어 있다. 특히 'murder'는 500인의 인물기록에서 총 2,640번이나 나타나 평균 1인 기록에서 약 5번 이상 나타난 것으로 조사되었다. 반면 'state', 'time', 'family', 'born', 'name', 'age'와 같은 모든 인물에게서 보편적으로 나타날 수 있는 용어들도 많이 출현하고 있는 것으로 나타났다.

범죄자 유형의 인물기록에서 출현한 고빈도어 상위 50위 용어의 주제성을 평가한 결과 주제성이 있는 용어는 50개 중 21개였고 일반 개념을 표현한 용어는 27개로 일반 개념을 표현한 용어가 더 많이 포함되어 있는 것으로 조사되었다. 또한 'death'나 'body' 같은 용어는 범죄자 유형에서 주제성에 해당하는지 용어만으로는 개념이 모호하여 주제성을 평가한 5인의 의견이 일치하지 못하여 Y/N로 주제성 불일치어로 분류하였다.

〈표 2〉 범죄자 인물기록 고빈도어 상위 50위

순위	용어	빈도	주제성	순위	용어	빈도	주제성
1	murder	2640	Y	26	county	805	N
2	death	1782	Y/N	27	crime	795	Y
3	state	1716	N	28	age	792	N
4	time	1635	N	29	gang	765	Y
5	kill	1496	Y	30	house	763	N
6	life	1462	N	31	american	749	N
7	convict	1201	Y	32	report	706	N
8	sentence	1192	Y	33	york	696	N
9	police	1152	Y	34	son	677	N
10	trial	1147	Y	35	attempt	642	Y
11	execute	1146	Y	36	children	641	N
12	prison	1140	Y	37	father	635	N
13	arrest	1099	Y	38	release	618	Y
14	family	1061	N	39	men	611	N
15	name	1041	N	40	november	602	N
16	born	1032	N	41	wife	596	N
17	shot/shoot	912	Y	42	july	596	N
18	claim	902	Y	43	june	586	N
19	court	898	Y	44	black	581	N
20	victim	864	Y	45	october	580	N
21	criminal	860	Y	46	month	577	N
22	charge	857	Y	47	office	575	N
23	work	855	N	48	hang	575	Y
24	body	833	Y/N	49	april	569	N
25	brother	823	N	50	law	562	Y

4.1.2 유형 2 - 교수 고빈도어

교수 카테고리에서 추출된 용어는 총 27,873개이며 이 중 불용어를 제외하고 빈도수 100개 이상으로 추출한 키워드는 총 241개이다. 추출된 241개의 용어 중 상위 50개의 용어를 살펴보면 〈표 3〉과 같다. 상위 10위까지 랭크된 고빈도어를 살펴보면 'university', 'research', 'science', 'study', 'professor' 등 대학과 연구에 관련된 용어가 포함되어 있다. 그러나 'work', 'nation', 'born', 'born'과 같이 교수집단에 특화되지 않는 용어들도 고빈도어로 나타나고 있다.

제일 많이 출현한 'university'와 두 번째로 많이 출현한 용어인 'research'와의 빈도차이는 약 2.6배로 'university'가 교수 유형의 인물기록에서 압도적으로 많이 출현하고 있는 것으로 조사되었다.

교수 유형의 인물기록에서 출현한 고빈도어 상위 50개에서는 주제성이 있는 용어가 50개 중 25개였고 일반 개념을 표현한 용어는 19개로, 범죄자 유형과 달리 교수 유형의 고빈도어에 주제성이 있는 용어들이 더 많이 포함된 것으로 나타났다. 또한 'book'이나 'education'처럼 용어만으로는 주제성을 평가하기 힘든 용어도 6개로 범죄자 유형보다는 많이 나타났다.

〈표 3〉 교수 인물기록 고빈도어 상위 50위

순위	용어	빈도	주제성	순위	용어	빈도	주제성
1	university	3495	Y	26	isbn	475	Y
2	research	1327	Y	27	school	473	Y
3	work	1326	Y/N	28	history	461	N
4	award	1300	Y	29	member	453	Y/N
5	science	1112	Y	30	alberta	450	N
6	study	948	Y	31	art	434	Y
7	professor	931	Y	32	develope	426	Y
8	nation	830	N	33	journal	401	Y
9	book	821	Y/N	34	politics	395	Y
10	born	741	N	35	social	393	N
11	institute	739	Y	36	department	392	Y
12	american	734	N	37	canadian	392	N
13	california	614	N	38	computer	388	Y/N
14	career	601	N	39	college	386	Y
15	public	593	N	40	mexico	380	N
16	state	574	N	41	intern	379	N
17	year	565	N	42	economy	373	Y
18	life	560	N	43	field	362	Y
19	publish	546	Y	44	theory	358	Y
20	press	537	Y	45	world	354	N
21	canada	527	N	46	york	351	N
22	society	514	Y	47	prize	351	Y
23	berkeley	513	Y	48	scientific	343	Y
24	association	494	Y	49	content	342	N
25	education	491	Y/N	50	author	331	Y/N

4.1.3 유형 3 - 영화감독 고빈도어

영화감독 카테고리에서 추출된 용어는 총 18,162개로 세 유형 중 가장 적었다. 교수 유형보다 평균 문자수는 많았으나 실제 추출된 용어는 교수유형보다 적게 나타났다. 이는 영화감독 유형의 인물기록에서는 다른 유형에 비해 상대적으로 인물에 대하여 서술하는데 공통된 용어를 많이 사용한 것으로 해석될 수 있다. 즉, 인물기록을 기술하는데 사용한 용어의 종수가 상대적으로 다른 집단에 비해 적게 나타난 것이다.

불용어를 제외하고 빈도수 100개 이상으로 추출한 용어는 총 203개이다. 추출된 203개의 용어 중 빈도기준 상위 50개의 용어를 살펴보면 〈표 4〉와 같다. 상위 10위까지 랭크된 고빈도어를 살펴보면 'film', 'award', 'produce', 'direct', 'television' 등 영화제작과 수상에 관련된 용어가 포함되어 있다.

제일 많이 출현한 'film'은 다른 유형 범죄자에서는 최상위 고빈도어 'muurder'가 2,640번 출현하고 교수유형에서 'university'가 3,495번 출현한 것과 비교하였을 때 거의 2-3배에 해당하는 7,245번으로 빈도차이가 크게 나타났다. 이는 영화감독 유형의 인물기록이 다른 유형에 비하여 평균 문자수는 크게 차이가 나지 않았지만 용어

〈표 4〉 영화감독 인물기록 고빈도어 상위 50위

순위	용어	빈도	주제성	순위	용어	빈도	주제성
1	film	7425	Y	26	story	529	Y
2	award	2296	Y	27	role	507	Y
3	produce	2073	Y	28	picture	500	Y
4	direct	1844	Y	29	video	496	Y
5	best	1785	Y	30	screenplay	488	Y
6	television	1575	Y	31	universal	476	Y
7	work	1336	Y/N	32	screen writer	441	Y
8	nomination	1109	Y	33	school	439	N
9	feature	1008	Y	34	show	435	Y
10	born	979	N	35	appear	403	N
11	episode	931	Y	36	academy	401	Y
12	festival	928	Y	37	love	382	N
13	writer	860	Y	38	comedy	373	Y
14	life	842	N	39	name	370	N
15	movie	823	Y	40	film maker	363	Y
16	series	803	Y	41	angel	361	N
17	career	799	N	42	animation	359	Y
18	actor	780	Y	43	project	359	Y
19	product	647	Y	44	occupation	350	N
20	filmography	604	Y	45	title	328	Y/N
21	release	592	Y	46	studio	325	Y
22	music	587	Y	47	company	316	N
23	critic	566	Y	48	drama	309	Y
24	documentary	545	Y	49	origin	309	N
25	art	531	Y	50	book	305	Y/N

종수에서는 차이가 나타났던 것과 연관된 현상이다. 즉, 영화감독에 대하여 서술할 경우 특정 용어를 집중적으로 사용하면서 기술하는 특성이 있는 것이다.

영화감독 유형의 고빈도어 주제성평가에서는 주제성이 있는 용어가 50개 중 35개였고 일반 개념을 표현한 용어는 10개로 세 유형 중 주제성이 있는 용어들이 가장 많이 포함된 것으로 나타났다. 이는 영화감독 유형의 인물기록에서 산출한 용어 종수가 적었던 현상과도 연관이 있는 것으로 이 유형의 인물을 서술할 때는 주제성이 높은 특정 용어를 집중적으로 사용하는 성향이 있는 것으로 해석된다.

4.2 인물기록 유형별 차별어

4.2.1 유형 1 - 범죄자 차별어

범죄자 유형의 차별어를 도출해내기 위하여 총 빈도수 10회 이상 7,621개 용어를 추출한 후 차별지수의 값을 산출하였다. 차별지수 값이 높은 순으로 정렬을 한 후 불용어를 제거 한 후 차별지수가 높은 상위 50개의 용어를 대상으로 〈표 5〉와 같이 주제성을 평가하였다.

고빈도어 상위 50위와 비교하였을 경우 범죄자 유형의 인물기록 주제적 특성을 나타내는 용어들에 변화가 나타났다. 대표적으로 'robbery',

〈표 5〉 범죄자 차별어 상위 50위

순위	키워드	차별지수	주제성	순위	키워드	차별지수	주제성
1	murder	0.00143304	Y	26	robbery	0.00026468	Y
2	kill	0.00073872	Y	27	guilty	0.00025768	Y
3	death	0.00068557	Y/N	28	family	0.00025172	N
4	sentence	0.00068059	Y	29	month	0.00024395	N
5	prison	0.0006179	Y	30	time	0.00023497	N
6	trial	0.00061176	Y	31	brother	0.00022508	N
7	convict	0.00060266	Y	32	escape	0.00022428	Y
8	police	0.00059966	Y	33	house	0.00022004	N
9	arrest	0.00059844	Y	34	evidence	0.00021213	Y
10	victim	0.00047136	Y	35	imprison	0.00021196	Y
11	criminal	0.00045863	Y	36	suspect	0.00020859	Y
12	charge	0.00042192	Y	37	penalty	0.00020772	Y
13	gang	0.00040614	Y	38	confess	0.00020299	Y
14	county	0.00040257	N	39	accuse	0.00019623	Y
15	court	0.00040081	Y	40	capture	0.00019578	Y
16	claim	0.00038728	Y	41	dure	0.0001948	Y
17	crime	0.00038418	Y	42	men	0.00019086	N
18	body	0.00038264	Y/N	43	commit	0.00018928	Y
19	hang	0.00031811	Y	44	trial	0.00018856	Y
20	execute	0.00030516	Y	45	investigate	0.00018739	Y
21	case	0.00030217	Y/N	46	slave	0.00017099	Y
22	may	0.0002948	N	47	wound	0.00017058	Y
23	attempt	0.00028248	Y	48	accord	0.00016666	N
24	report	0.0002755	N	49	testify	0.00016496	Y
25	state	0.00027202	N	50	son	0.00016332	N

'guilty', 'evidence', 'imprison', 'suspect', 'penalty', 'confess', 'accuse' 같은 용어들은 범죄자 유형의 인물기록 특성을 표현하는 주제성이 있는 용어임에도 불구하고 고빈도순으로는 상위 50위 안에 포함되지 못한 용어들인데 차별지수 순에서는 상위로 랭크되어 차별지수가 특정 유형을 특화하여 표현하는 용어들을 선별해내는 역할을 하고 있는 것으로 나타났다.

범죄자 차별어 상위 50위 용어의 주제성을 평가한 결과 주제성이 있는 용어가 35개, 일반 용어가 12개, 주제성을 평가하는데 평가자간에 불일치한 용어가 3개로 나타났다. 이를 고빈도 상위 50개 용어의 주제성 평가결과와 비교해보면 고빈도 상위 50개 용어에서는 21개로 나타났던 주제어가 차별어 상위 용어 50개에서는 35개로 늘어나 차별어가 주제성이 높은 용어를 선별해내는 효과가 있는 것으로 분석되었다.

4.2.2 유형 2 - 교수 차별어

교수 유형의 인물기록을 표현하는데 적합한 차별어를 도출해내기 위하여 총 빈도수 10회 이상 6,536개 용어에 대하여 차별지수의 값을 산출하였다. 불용어를 제거 한 후 차별지수가 높은 상위 50개의 용어를 산출하여 〈표 6〉과 같이

〈표 6〉 교수 차별어 상위 50위

순위	용어	차별지수	주제성	순위	용어	차별지수	주제성
1	university	0.004051	Y	26	department	0.000387	Y
2	research	0.001692	Y	27	politics	0.000366	Y
3	science	0.001322	Y	28	physics	0.000365	Y
4	professor	0.001161	Y	29	develope	0.000364	Y
5	study	0.001063	Y	30	association	0.000364	Y
6	institute	0.000828	Y	31	mathematics	0.000361	Y
7	nation	0.000691	Y/N	32	german	0.00034	N
8	book	0.00066	Y/N	33	fellow	0.000325	Y
9	press	0.000582	Y	34	prize	0.00032	Y
10	isbn	0.000536	Y	35	college	0.000314	Y
11	work	0.000534	Y/N	36	engine	0.000314	Y
12	public	0.000491	N	37	doctor	0.000308	Y
13	society	0.000487	Y	38	faculty	0.000292	Y
14	publish	0.000481	Y	39	technology	0.000285	Y
15	economy	0.000464	Y	40	chemistry	0.000284	Y
16	journal	0.000463	Y	41	ph.d	0.000279	Y
17	education	0.000455	Y/N	42	human	0.000272	N
18	computer	0.000451	Y/N	43	oxford	0.000269	Y
19	scientific	0.000451	Y	44	award	0.000267	Y
20	social	0.00043	N	45	member	0.000259	Y/N
21	theory	0.00043	Y	46	contribute	0.000256	N
22	history	0.000422	N	47	teach	0.000255	Y
23	field	0.000414	Y	48	degree	0.000243	Y
24	academy	0.000409	Y	49	poetry	0.00023	Y
25	philosophy	0.000392	Y	50	policy	0.000216	Y

주제성을 평가하였다.

교수유형의 인물기록에서 나타난 고빈도어 상위 50위와 비교하였을 경우 범죄자 유형과 유사하게 기존의 50위에 포함되지 못하였던 주제적 특성을 나타내는 용어들이 차별어 상위 50위에 나타났다. ‘academy’, ‘physics’, ‘mathematics’, ‘chemistry’, ‘ph.d’ 같은 용어들은 교수유형의 인물기록 특성을 용어들이 차별지수 순에서는 상위로 랭크되어 교수유형에서도 차별지수가 특정 유형을 특화하여 표현하는 용어들을 선별해내는 역할을 하고 있는 것으로 나타났다. 그러나 상대적으로 범죄자 유형의 비교 결과보다는 고빈도어와 차별어의 순위변동과 변경사항이 크게 나타나지 않았다.

교수 차별어 상위 50위 용어의 주제성을 평가한 결과 주제성이 있는 용어가 38개, 일반 용어가 6개, 주제성을 평가하는데 평가자간에 불일치한 용어가 6개로 나타났다. 범죄자 유형과 유사하게 차별어 상위에 주제성이 있는 용어가 25에서 38개로 차별어가 고빈도어보다 주제성이 높은 용어를 선별해내는 효과가 있는 것으로 분석되었다.

4.2.3 유형 3 - 영화감독 차별어

영화감독 유형을 표현하는 차별어를 도출해내기 위하여 총 빈도수 10회 이상 6,907개 용어를 추출한 후 차별지수의 값을 산출하였다. 차별지수 값이 높은 순으로 상위 50개의 용어를 살펴보면 〈표 7〉과 같다.

〈표 7〉 영화감독 차별어 상위 50위

순위	용어	차별지수	주제성	순위	용어	차별지수	주제성
1	film	0.008107	Y	26	role	0.000398	Y
2	director	0.00295	Y	27	film maker	0.000396	Y
3	produce	0.002194	Y	28	critic	0.000388	Y
4	direct	0.001921	Y	29	release	0.00038	Y
5	best	0.001828	Y/N	30	edit	0.000366	Y
6	television	0.001474	Y	31	theater	0.000365	Y
7	award	0.001371	Y	32	studio	0.000337	Y
8	nomination	0.001208	Y	33	love	0.000323	Y/N
9	festival	0.001009	Y	34	career	0.000321	N
10	star	0.001008	Y	35	drama	0.000317	Y
11	feature	0.001006	Y	36	play	0.000317	Y
12	episode	0.000914	Y	37	year	0.000307	N
13	movie	0.000855	Y	38	animation	0.000301	Y
14	actor	0.000824	Y	39	hollywood	0.000298	Y
15	writer	0.000797	Y	40	story	0.000288	Y
16	filmography	0.000678	Y	41	angel	0.000258	N
17	series	0.000652	Y	42	work	0.000253	Y/N
18	product	0.000602	Y	43	credit	0.000248	Y
19	screenplay	0.000549	Y	44	artist	0.000241	Y
20	documentary	0.000545	Y	45	actress	0.000239	Y
21	video	0.000499	Y	46	show	0.000238	Y
22	picture	0.000498	Y	47	cinema	0.000229	Y
23	screen writer	0.000495	Y	48	script	0.000227	Y
24	music	0.000486	Y	49	title	0.000226	Y/N
25	comedy	0.000409	Y	50	appear	0.000224	N

영화감독 인물기록에서 나타난 고빈도어 상위 50위 용어와 비교하였을 경우 다른 유형의 차별어 분석결과와 유사하게 이 유형에서도 인물기록 주제적 특성을 나타내는 용어들이 차별지수 상위권에 나타났다. 'theater', 'hollywood', 'credit', 'actress', 'cinema', 'script' 등 영화감독 유형의 인물기록 특성을 표현하는 주제성이 있는 용어들이 차별지수 순에서는 상위로 랭크되어 이 유형에서도 차별지수가 특정 유형을 특화하여 표현하는 용어들을 선별해내는 역할을 하고 있는 것으로 나타났다. 영화감독 차별어 상위 50위 용어의 주제성을 평가하여 고빈도어 상위 50위 용어와 비교한 결과 다른 유형과 동일하게 차별어 상위 50개 용어에서 주제성이 있는 용어 수가 증가한 것으로 나타났다. 고빈도어 상위 50개중 35개였던 주제어가 차별어에서는 42개로 나타났으며 일반 용어가 3개, 주제성을 평가하는데 평가자간에 불일치한 용어가 5개로 조사되었다. 전체 50개 용어 중 대부분이 영화감독과 연관된 용어인 것으로 분석되어 영화감독 유형의 특성을 표현하는데 차별어가 효과적인 것으로 나타났다.

5. 공통어 및 차별어 워드클라드 적용 서비스 개선방안

Wikipedia의 Biography portal에 축적되는 인물기록은 2장에서 살펴보았듯이 이용자들이 참여하여 자유롭게 서술하는 형태의 인물기록으로 실제 인물에 대하여 서술하는데 있어 어떤 측면으로 서술을 해야 하는지 구체적인 가이드가 주어지지 않고 있다. 이용자가 인물기록을 서술하는데 일반적인 가이드는 아래와 같이 일부 주어지고 있지만 인물별로 어떤 요소들을 고려해서 기록정보를 추가해야 하는지에 대해서는 가이드를 주고 있지 못하다. 예를 들어 인물기록 페이지에서 제공하고 있는 가이드를 살펴보면 〈예 1〉에서 제시한 가이드는 서술한 근거를 제시하라는 일반적인 내용이며 〈예 2〉에서 제시한 가이드는 서술한 해당 인물기록이 문제가 있다는 지적을 하고 있지만 구체적으로 어떤 내용을 보완하라는 설명을 해주고 있지 못하다.

〈예 1〉
This article includes a list of references, related reading or external links, but its sources remain unclear because it lacks inline citations. Please help to improve this article by introducing more precise citations. (January 2016) (Learn how and when to remove this template message)

〈예 2〉
This article has multiple issues. Please help improve it or discuss these issues on the talk page.
This article is an orphan, as no other articles link to it. Please introduce links to this page from related articles; try the Find link tool for suggestions. (April 2014)
This article relies largely or entirely on a single source. (April 2017)

이러한 맥락에서 이 연구에서는 3개 유형을 실험적으로 선정하여 각 유형별로 인물기록을 서술하는 이용자들이 선택한 용어로 특정 유형의 인물 특성을 나타낼 수 있는 방안을 도출하여 해당유형의 인물기록을 서술하거나 검색할 때 참조할 수 있도록 제안하고자 하였다. 그 방안으로 차별지수로 산출한 차별어가 각 유형별

로 특성을 나타낼 수 있는 주요어 역할을 할 수 있는지 조사하였는데 〈표 8〉과 같다. 〈표 8〉에 의하면 고빈도어 상위 50개에 포함되어 있는 주제어의 비율이 차별어 상위 50개에서 모두 크게 향상된 것을 알 수 있다.

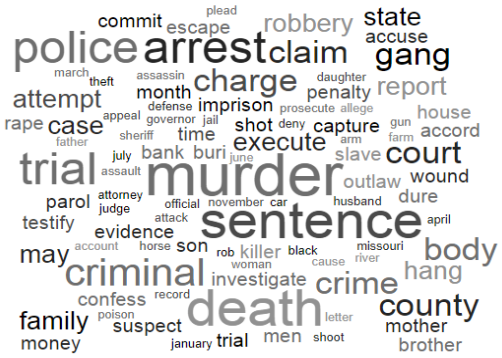
특히 범죄자 유형에서는 차별어에서 주제어의 비율이 고빈도에 미하여 거의 30% 향상된 것으로 나타났고 교수 유형과 영화감독 유형에서도 모두 10-20% 이상 향상된 것으로 조사되었다. 따라서 3개 유형 모두 차별어가 특정 유형의 인물에 대하여 서술하거나 검색하는데 참조할 수 있는 주제성이 있는 용어를 제시해주는데 고빈도어보다 효과적인 것으로 나타났다. 이 연구에서는 차별어의 특성을 활용하여 차별지수로 산출된 용어 100개를 추출한 후 차별지수값을 반영하여 〈그림 4〉와 같이 워드클라우드로 생성하여 해당 유형의 인물기록을 서술하거나 열람하는 이용자에게 제시하고자 하였다.

각 유형별 워드 클라우드를 비교해보면 각 유형별로 이용자가 해당 유형의 인물을 서술하는데 사용하는 용어 간에 차별점이 있는 것을 쉽게 알아볼 수 있다. 범죄자 차별어 클라우드에서는 'murder', 'sentence', 'arrest', 'police' 같은 용어들이 주요어로 쉽게 파악될 수 있었고 교수 유형에서는 'university', 'professor', 'study', 'journal', 'book' 같은 용어들이 주요어로 표현되었다. 또한 영화감독 유형의 차별어 클라우드에서는 'film', 'award', 'episode', 'festival'과 같은 용어들이 다른 집단과 차별되어 주요하게 표현되어 있어 이용자들이 해당 유형의 인물에 대하여 서술할 때 어떤 측면을 주요하게 서술하는지 알 수 있게 해준다. 따라서 〈그림 4〉와 같이 각 유형별로 차별지수를 중심으로 주요 개념들을 표현해준다면 해당 유형의 인물기록을 서술하는 이용자나 검색하는 이용자에게 가이드 역할을 할 수 있을 것이다. 또한 이용자가 서술한 인물기록을 승인하거나 보완하라고 의견을 첨부하는 에디터에게도 인물의 특성을 반영한 기록이 적절하게 서술되고 있는지 판단할 수 있는 기준으로 활용될 수 있을 것이다.

인물기록을 서술할 경우 각 유형별로 특화되는 정보도 서술하지만 인물이라는 서술대상이 공통적으로 가지고 있는 요소에 대해서도 서술을 하게 된다. 이 연구에서는 인물기록을 분석하면서 유형별 차별성도 파악하였지만 이용자가 인물에 대하여 서술할 때 공통적으로 나타나는 보편적인 개념도 분석 도출하여 이용자에게 가이드로 제공하고자 하였다. 범죄자, 교수, 영화감독 세 유형에서 공통적으로 나타나는 용

〈표 8〉 고빈도어 및 차별어 주제성 비교

인물유형	용어유형	주제어	주제성불일치어	일반용어
범죄자	고빈도어 상위 50	42%	4%	54%
	차별어 상위 50	70%	6%	24%
교수	고빈도어 상위 50	50%	12%	38%
	차별어 상위 50	76%	12%	12%
영화감독	고빈도어 상위 50	70%	10%	20%
	차별어 상위 50	84%	10%	6%



범죄자 - 차별어

교수 - 차별어

영화감독 - 차별어

〈그림 4〉 유형별 인물기록 차별어 워드 클라우드

공통어 - 워드클라우드

〈그림 5〉 인물기록 공통어 워드 클라우드

어 중 빈도 10개 이상의 총 11,841개 용어 중 불용어를 제거하고 총 빈도순으로 상위 100위까지 용어를 주요 공통어로 추출하였다. 추출된 공통어를 대상으로 생성한 워드 클라우드는 〈그림 5〉와 같다. 공통어 워드 클라우드는 이용자가 참여하여 서술한 인물기록에서 공통적으로 언급하는 내용을 표현하고자 하는 목적으로 생성하였는데 생성된 워드 클라우드를 살펴보면 'work', 'career', 'marriage', 'occupation', 'school', 'education' 등 모든 유형의 인물에 적용하여 서술할 수 있는 내용들을 표현하고 있다. 따라서 특정 인물을 대상으로 서술할 때도 보편적으로 인물에 대하여 서술하는 기본 사항을 가이드 해주는 용도로 공통어 워드 클라우드를 사용할 수 있으며 인물에 대하여 검색하는 용어를 참조하는 용도로도 활용될 수 있을 것이다.

6. 결 론

이 연구에서는 위키피디아 바이오그래피 포털의 범죄자, 교수, 영화감독 카테고리에서 인물 유형별로 500건의 데이터를 각각 수집하여 각 유형별로 특화된 요소들이 있는지 분석하여 이를 이용자와 에디터에게 제공함으로써 인물기록 서비스를 개선할 수 있는 도구를 제시하고자 하였다. 분석 결과는 다음과 같다.

첫째, 고빈도어로 상위 50개 용어를 분석한 결과 상위 50개의 용어에 특정 유형의 인물기록 특성을 표현할 수 있는 주제성이 있는 용어가 포함된 비율이 유형별로 차이를 나타내었다. 주제성 있는 용어의 비율은 범죄자의 경우 42%로 가장 낮았으며 교수는 50%로 중간수준이었으며 영화감독은 70%로 가장 높은 비율을 나타냈다. 이는 영화감독과 같은 유형의 인물기록을 서술하는데 있어서 이용자가 주제를 표현할 수 있는 용어를 집중적으로 활용하는 성향이 있는 것으로 해석할 수 있으며 이는 영화감독 인물기록에서 추출된 용어의 종수가 가장 적게 나타난 것과도 연관이 있다. 즉, 이용자가 이 집단의 인물기록을 서술하는데 일반적인 용어를 다양하게 많이 사용하지 않는 것이다.

둘째, 차별지수로 상위 50개 용어를 분석한 결과 주제성 용어가 포함된 비율의 편차가 고빈도어보다 적게 나타났다. 범죄자 70%, 교수 76%, 영화감독 84%로 비교적 상위 수준으로 고르게 나타났다. 차별지수로 주제성이 있는 용어를 선별할 경우 유형별 편차가 크게 나타나지 않아 고빈도어보다 안정성있게 주제성이 있는 용어를 추출해내는데 효과적인 것으로 나타났다.

셋째, 각 유형의 인물기록이 가지고 있는 특성을 살펴보고자 유형별 인물기록에서 많이 출현하는 고빈도어와 특정 유형에서 집중적으로 출현하는 용어를 차별지수로 산출하여 비교한 결과 차별지수로 산출한 용어가 특정 유형의 인물기록을 표현하는 주제성이 더 높은 것으로 나타났다. 고빈도어 상위 50개에 포함되어 있는 주제어의 비율이 차별지수 순 상위 50개에서 모두 크게 향상된 것을 알 수 있다. 가장 큰 차이를 보인 범죄자 유형에서는 차별어에서 주제어의 비율이 고빈도에 미하여 거의 30% 향상되었고 교수 유형과 영화감독 유형에서도 모두 10-20% 이상 향상되었다.

넷째, 범죄자, 교수, 영화감독 등 세 유형에서

차별지수로 추출한 용어 100개와 세 유형에서 공통어로 추출한 용어 100개를 대상으로 워드클라우드를 생성하였더니 각 유형별 차별어 클라우드와 공통어 클라우드 용어간이 차별성이 나타났다. 범죄자 유형에서는 'murder', 'police'와 같은 범죄와 관련된 용어들이 중요하게 나타났고 교수 유형에서는 'university', 'research', 영화감독 유형에서는 'film', 'episode' 같은 용어들이 표현되었다. 반면 공통어에서는 'work', 'marriage' 같은 인물기록 전체에 적용될 수 있는 용어들이 중요한 용어로 표현되어 인물 공통요소를 가이드해주는 효과가 있는 것으로 나타났다.

이 연구에서는 각 유형별 인물기록에서 나타난 용어를 계량적으로 분석하여 주제성을 나타내는 용어와 공통성을 나타내는 용어를 추출하여 활용하는 방안을 제시하였는데 다음과 같은 제한 사항이 도출되어 추가 연구로 발전시킬 필요성이 있는 것을 파악하였다. 첫째, 용어의 의미가 유형별 인물기록에서 다르게 나타났다. 예를 들어 범죄자 유형의 인물기록에서 'death'는 범죄사실이기도 했으며 범죄자의 일생에 관련된 기록이기도 했다. 따라서 주변어와의 의미적 관계를 분석해야만 실질적인 의미가 해석될 수 있는 한계점이 나타났다. 둘째, 'public'이나 'computer'와 같이 일반적인 용어가 유형에 따라서는 실제적으로는 복합명사의 개념으로 주제성을 나타내는 경우가 있었다. 즉, 교수 유형에서 public은 public health와 같은 학문 분야를 표현한 용어여서 주제성이 있는 용어가 될 수 있는 가능성이 있었다. 따라서 복합명사나 구를 처리하는 방안도 추가적으로 고려해야 할 것이다.

이와 같이 용어의 실질적 의미를 파악하기 위해서는 이 연구에서 제시한 차별지수를 적용하여 주요어를 추출하고 추가적으로 네트워크 분석을 하여 용어간의 관계를 표현하는 것도 고려하는 것이 필요하며 추후 연구로 제안하는 바이다.

참 고 문 헌

[1] 김유승. 2010. 아카이브 2.0 구축을 위한 이론적 고찰.『한국기록관리학회지』, 10(2): 31-52.
[2] 김하수 외. 2013. 정치와 언어의 관계에 대한 양적 분석 시론.『담화와 인지』, 20(1): 79-111.
[3] 김흥식, 김진형. 2011.『경기도 역사인물의 문화콘텐츠화를 위한 OSMU 적용방안』. 수원: 경기개발연구원.
[4] 남재우, 김성희. 2009. 기록정보서비스를 위한 Web2.0 적용에 관한 연구.『한국문헌정보학회지』, 43(2): 123-146.
[5] 설문원. 2010. 기록 검색도구의 발전과 전망.『기록학연구』, 23: 3-43.
[6] 송정숙. 2012. 부산의 기억과 로컬리티.『한국도서관·정보학회지』, 43(2): 343-364.
[7] 이재윤. 2007. 분포 유사도를 이용한 문헌클러스터링의 성능향상에 대한 연구.『정보관리학회지』,

24(4): 267-283.

[8] 이재윤, 김수정. 2016. 국내 재난 관련 연구 동향에 대한 계량정보학적 분석.『정보관리학회지』, 33(3): 103-124.

[9] 장혜란. 2001. 우리나라 온라인 인물데이터베이스의 비교 평가 연구.『한국도서관·정보학회지』, 32(4): 283-302.

[10] 한상길. 2008. 국내 인물데이터베이스의 구축과 서비스에 관한 비교 분석.『한국도서관·정보학회지』, 39(4): 331-352.

[11] 현문수, 김동철. 2013. 식별된 저자 지역사 인물 콘텐츠 개발을 위한 연구: 박기종 사례를 중심으로.『기록학연구』, 36: 195-231.

[12] Callahan, E. S. and Herring, S. C. 2011. "Cultural Bias in Wikipedia Content on Famous Persons." *Journal of the Association for Information Science and Technology*, 62(10): 1899-1915.

[13] Connor, P. 2014. "Quantifying Immigrant Diversity in Europe." *Ethnic and Racial Studies*, 37(11): 2055-2070.

[14] Keith, G. F. 2017. "Population Movements in a Warwickshire Village 1841-1891: Bidford-on-Avon." *Local Population Studies*, 98(1): 74-86.

[15] Leskinen P. et al. 2017. Modeling and Using an Actor Ontology of Second World War Military Units and Personnel. In: d'Amato C. et al. (eds) *The Semantic Web - ISWC 2017. ISWC 2017, Vienna: Lecture Notes in Computer Science*, vol. 10588: 280-296.

[16] Ofek, N. and Rokach, L. 2015. "A Classifier to Determine which Wikipedia Biographies Will be Accepted." *Journal of the Association for Information Science and Technology*, 66(1): 213-218.

[17] Thomson, A. 2016. "Digital Aural History: An Australian Case Study." *The Oral History Review*, 43(2): 292-314.

[18] Soules, A. 2012. "Where's the Bio? Databases, Wikipedia, and the Web." *New Library World*, 113(1/2): 77-89.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Kim, You-Seung. 2010. "A Theoretical Study on Establishing Archive 2.0." *Journal of Korean Society of Archives and Records Management*, 10(2): 31-52.

[2] Kim, Ha-Soo et al. 2013. "A Quantitative Approach to the Relation between Politics and

Language." *Discourse and Cognition*, 20(1): 79-111.

[3] Kim, Heung-Sik and Kim, Jin Heung. 2011. *Application of OSMU to the Cultural Contents of Historical Characters in Gyeonggi-Do*. Suwon: Gyeonggi Research Institutue.

[4] Nam, Jae-Woo and Kim, Seong-Hee. 2009. "A Study on the Application of Web 2.0 for Archival Information Services." Journal of the *Korean Society for Library and Information Science*, 43(2): 123-146.

[5] Seol, Moon-Won. 2010. "A Study on Development and Prospects of Archival Finding Aids." *The Korean Journal of Archival Studies*, 23: 3-43.

[6] Song, Jung-Sook. 2012. "Memories and the Locality of Pusan - Focusing on Historical Figures of Busan and Cultural Properties of Busan -." *Journal of Korean Library and Information Science Society*, 43(2): 343-364.

[7] Lee, Jae Yun. 2007. "Improving the Performance of Document Clustering with Distributional Similarities." *Journal of the Korean Society for Information Management*, 24(4): 267-283.

[8] Lee, Jae Yun and Kim, Soojung. 2016. "A Bibliometric Analysis of Research Trends on Disaster in Korea." *Journal of the Korean Society for Information Management*, 33(3): 103-124.

[9] Chang, Hye Rhan. 2001. "Evaluating Online Biographical Databases in Korea: A Comparative Study." *Journal of Korean Library and Information Science Society*, 32(4): 283-302.

[10] Han, Sang-Kil. 2008. "A Comparative Study about Construction and the Service of the Domestic Biographical Database." *Journal of Korean Library and Information Science Society*, 39(4): 331-352.

[11] Hyun, Moon Soo and Kim, Dong Chul. 2013. "A Study on Developing Archival Contents for Documenting Local Historical Characters." *The Korean Journal of Archival Studies*, 36: 195-231.