빅데이터 기반 관광지 추천 시스템 구현†

– 한국관광공사 LOD를 중심으로 –

안진현* · 김응희** · 김홍기***

〈요 약〉

기존 전시회 정보 제공 서비스는 전시회가 열리는 장소 주변의 관광지를 추천한다. 이러한 위치기반 추천의 경우 전시회의 내용과 관련이 없는 관광지를 추천할 수 있다는 한계점이 있다. 전시회 내용과 관련된 관광지를 관람객에게 추천함으로써 전시회에서 획득한 지식을 관광지에서 경험하는 데에 도움을 줄 필요가 있다. 전시회 큐레이터들이 전시회 내용과 관련된 관광지를 일일이 찾아 추천하는 방법이 있지만, 수작업이다 보니 큐레이터가 가지고 있는 배경지식의 범위 내에서만 추천이 가능하다는 한계점이 있다. 수작업에 따른 오류가 있을 수도 있기 때문에 자동화된 방법이 필요하다. 본 연구에서는 언어자원 빅데이터를 활용하여 전시회 내용과 관련된 관광지를 자동으로 추천하는 방법을 제안한다. 언어자원으로는 한국관광공사 LOD(Linked Open Data), 위키피디아, 국립국어원 사전 등을 활용했다. 단일 컴퓨터로는 이러한 대용량 언어자원을 효율적으로 처리하기 어렵기 때문에, 클라우드 컴퓨팅 프레임워크인 아파치 스파크(Apache Spark)에 기반하여 구현했다. 사용자가 웹브라우저를 통해 전시회 정보를 열람하면 본 알고리즘에 의해 추천된 관광지들을 같이 보여주는 웹인터페이스도 구현했다(http://bike.snu.ac.kr/WARP). 주요 전시회에 대한 관광지 추천 정확도에 대해 전문가 평가를 진행했다. 기존 방법에 비해 본 논문에서 제안한 방법의 정확도가 더 높았다. 본 연구를 활용하면 전시회 큐레이터의 수작업을 줄여줄 수 있고 전시회 관람자들을 관광지로 자연스럽게 유도할 수 있기 때문에, 전시산업과 관광산업 모두에게 도움이 될 수 있다.

핵심주제어: 관광지 추천, 전시회 정보 관리 시스템, 링크드 오픈 데이터, 대용량 지식 처리

논문접수일: 2017년 10월 13일 수정일: 2017년 11월 08일 게재확정일: 2017년 11월 09일
† 이 논문은 2017년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (2013-0-00109, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)
* 서울대학교 치학연구소 선임연구원(제1저자), jhahncs@snu.ac.kr
** 서울대학교 치의학생명과학사업단 BK21 PLUS 조교수, eungheekim@snu.ac.kr
*** 서울대학교 치과대학 의료경영과정보학 교수(교신저자), hgkim@snu.ac.kr

Ⅰ. 서 론

전시회 정보 관리 시스템(EIS)은 전시회 관련 정보 처리 모듈들의 집합이다(Yang Haiying et al., 2010). EIS의 대표적인 기능들로는 관람객 등록 관리, 전시회 내용 및 관련 정보 제공, 관람객의 피드백 관리, 광고 관리 등이 있다(유성열・이강배, 2013; Yang Changhui and Meng Hongyan 2016). 본 논문에서는 전시회 관련 정보 제공 기능 중에서 특히 관광지 추천 기능에 초점을 맞춘다.

대표적인 EIS의 예로는 문화포털, 네이버 미술작품, 한국관광공사 웹서비스 등이 있다. 국내에서 열리는 전시회(또는 축제, 문화제, 공연 등)를 주제, 시기, 장소 등의 기준으로 검색하고 전시회에 대한 자세한 정보를 제공한다. 전시회가 열리는 장소 주변의 관광지에 대한 정보도 제공한다. 이러한 위치 기반 관광지 추천은 지도 서비스로부터도 얻을 수 있는 정보이기 때문에 EIS의 고유 기능이라고 할 수 없다. EIS에서는 전시회에서 획득한 지식을 실제 장소(관광지)에서 경험하는 데에 도움이 되는 정보를 제공할 필요가 있다. 변상우(2015)는 관광지의 정서적 이미지에 유의한 영향을 미치는 요인들 중 하나로 문화행사 관람을 제시했다. 즉, 전시회의 내용과 관련된 관광지를 추천하는 내용기반 관광지 추천 기능이 필요하다.

예를 들어, 교산허균 문화제에 대한 정보를 열람한다고 가정하자. <그림 1>은 한국관광공사 웹서비스에서 제공하는 교산허균 문화제에 대한 정보

<그림 1> 한국관광공사 웹서비스의 스크린샷 (위치기반 추천)

이다. 축제에 대한 설명글, 위치, 관람 가능 연령 등의 정보를 제공한다. 추가적으로 아래에 “주변 가볼만한곳”이라는 정보를 제공한다. 일종의 위치기반 관광지 추천으로, 문화제가 열리는 장소와 추천된 관광지 사이의 거리가 병기돼 있다. 추천된 강릉 이광노가옥, 초당두부마을, 경포호 등의 관광지 모두 교산허균 문화제의 내용과 관련이 없다. 교산허균 문화제 관람객들은 허균이 쓴 홍길동전의 주인공인 홍길동의 생가에 더 관심이 있을 것이다. 본 연구의 목표는 교산허균 문화제에 대해 허균과 관련이 있는 홍길동 생가, 이매창묘 등의 관광지를 자동으로 추천하는 것이다.

본 연구에 의한 내용기반 관광지 추천 기능이 추가된 EIS는 전시 산업과 관광 산업 모두에게 도움이 된다. 지도 서비스 등 다른 서비스로부터 얻을 수 없는 관광지 정보를 사용자에게 제공할 수 있기 때문에, EIS의 활용도가 높아진다. 전시회 관람객들에게 관광지를 홍보하는 효과도 있다. 전시회의 내용과 관련된 관광지이기 때문에 기존 위치기반 관광지 추천보다 흥미를 끌 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 EIS와 관광지 추천 관련 기존 연구들을 살펴본다. 3장에서는 전시회와 관광지 데이터를 가지고 있는 한국관광공사 LOD에 대해 설명한다. 4장에서는 내용기반 관광지 추천에 필요한 전시회-관광지 관련도 계산 방법에 대해 설명한다. 5장에서는 내용기반 관광지 추천 기능이 포함된 EIS 구현 방법에 대해 설명한다. 6장에서는 본 논문에서 제안하는 관광치 추천 방법의 효율성과 효과성에 대해 논의한다. 7장에서 본 논문을 정리하고 향후 연구에 대해 논의한다.

Ⅱ. 관련 연구

유러피아나 데이터 모델(Europeana Data Model, EDM)은 문화유산을 기술하기 위한 데이터모델이다(Martin Doerr et al., 2010; Cesare Concordia et al., 2010; Constantia Kakali et al., 2007). EDM의 데이터 표현 방법론은 객체 기반 기술 모델 방법론과 이벤트 기반 기술 모델 방법론으로 구성된다. 객체 기반 기술 모델은 객체(문화유산)를 중심으로 문화유산의 창작자, 주제, 제목, 형태 등을 기술하는 방법이다. 이벤트 기반 기술 모델은 이벤트(전시회)를 중심으로 전시회에서 전시되는 문화유산, 작가, 제목, 전시 기간 등을 기술하는 방법이다. EDM 데이터 기반의 웹서비스를 통해서 전시회 정보를 검색할 수 있다. 하지만, EDM에는 관광지에 관련된 데이터 또는 기술 방법이 없기 때문에 관광지 추천 기능이 없다.

Data.go.kr은 정부에 의해 유지관리가 되는 공개 데이터 저장소이다(Zhenbin Yang and Atreyi Kankanhalli, 2013; Christian Bizer et al., 2009; Liyang Yu, 2001). 전시회, 관광, 교육 등의 다양한 분야의 데이터가 공개돼 있다. 각 분야의 데이터가 별도의 파일로 독립적으로 존재하기 때문에 각각에 대한 검색 및 정보 열람은 가능하지만 관련 분야의 데이터를 연계해서 열람할 수 있는 기능이 없다. 즉, 전시회-관광지 연계 데이터가 존재하지 않기 때문에 관광지 추천 기능을 제공하지 않는다.

한국관광공사는 응용애플리케이션에서 활용할 수 있는 전시회 및 관광 관련 데이터와 일반 사용자를 위한 웹서비스를 제공한다. 데이터는 3절에서 자세히 설명한다. 웹서비스는 웹브라우저를 통해 사용할 수 있으며, 키워드 기반으로 전시회를 검색할 수 있는 기능을 제공한다. 전시회의 설명, 위치, 관람소요시간, 관람가능연령 등의 정보를 제공하고 추가적으로 근처 관광지 정보를 거리를 기준으로 제공한다. 근처 관광지 정보는 전시회가 열리는 장소의 위도/경도와 관광지의 위도/경도 정보를 활용하여 제공하는 것으로 위치기반 관광지 추천이다.

위치기반 관광지 추천의 경우 사용자의 관심사

를 고려하지 못하는 한계점이 있기 때문에, 내용기반 관광지 추천 연구가 진행됐다. 기존 대부분의 내용기반 관광지 추천 연구는 사용자의 상황정보를 분석하여 사용자가 관심이 있을만한 관광지를 추천하는 데에 초점을 맞추고 있다. 사용자의 상황정보는 사용자의 개인정보, 온라인에서의 활동정보, 오프라인에서의 활동정보 등이다. 박연진(2015)은 페이스북과 같은 소셜미디어서비스로부터 정보를 수집한 뒤 이를 온톨로지로 표현하여 개인의 선호도를 효과적으로 분석하는 방법을 제시했다. Kevin Meehan(2013)은 사용자의 위치, 특정 장소에 머무는 시간, 날씨, 소셜미디어서비스 등의 정보를 분석하여 관광지를 추천하는 방법을 제안했다. Joel P. Lucas(2013)은 사용자와 유사한 사용자들의 모임에서의 연관관계 분석기법을 통해 관광지를 추천하는 방법을 제안했다.

기존 내용기반 관광지 추천 연구와는 달리 본 연구에서는 사용자가 선택한 특정 객체(전시회)의 내용을 분석한다. 기존 연구가 일반적인 상황에서 또는 이동하는 상황에서 추천을 하는 시스템이라고 한다면 본 연구는 전시회 정보를 열람하는 특정한 상황에 초점을 맞춘다. 또한, 대부분의 기존 연구가 일반 사용자에 초점을 맞췄다면, 본 연구의 경우 전시회 큐레이터가 관람객에게 제공해야 할 정보를 구성함에 있어서 필요한 의사결정을 돕는 전시회 의사결정지원시스템(Decision Support System)으로도 응용할 수 있다.

Ⅲ. 한국관광공사 LOD

한국관광공사에서는 관광관련 데이터가 포함된 한국관광공사 LOD(KTO-LOD)를 제공하고 있다.[1] LOD(Linked Open Data)는 웹상에 공개하는 데이터의 표준으로, 전세계적으로 각광을 받고 있으며 우리나라 정부가 운영하는 공공데이터포털(http://www.data.go.kr)에서도 LOD 표준을 따르고 있다. LOD는 RDF(Resource Description Framework) 트리플의 집합이고 RDF 트리플은 (주어, 술어, 목적어)로 구성된 트리플로 정보를 구조화시켜서 표현하는 방법이다. 예를 들어, “교산허균 문화제는 강릉시에서 열리고 관람가능연령은 전연령이다”라는 정보는 (교산허균 문화제, 장소, 강릉시), (교산허균 문화제, 관람가능연령, 전연령)과 같이 2개의 RDF 트리플로 표현할 수 있다. 이와 같이 구조화돼 표현된 정보는 컴퓨터가 읽을 수 있는 장점이 있기 때문에 기존의 다양한 분야의 데이터가 RDF 트리플로 변환돼 공공데이터포털에 공개되고 있다. 응용애플리케이션 개발자는 LOD 데이터를 이용해 다양한 응용애플리케이션을 개발할 수 있다.

KTO-LOD에는 445,875개의 주어에 대한 1,640,977개의 RDF 트리플로 구성됐다. DBPedia(Sören Auer et al., 2007), Wikidata(Denny Vrandečić, 2014), YAGO(Farzaneh Mahdisoltani, 2014) 등 전세계적으로 활발히 활용되고 있는 데이터셋에도 연결돼 있어서 다양한 분야로 확장이 가능하다. 본 논문에서는 KTO-LOD에 존재하는 다양한 술어들 중에서 “설명글(dc:description)” 술어에 초점을 맞춘다. 예를 들어, <그림 1>에 교산허균 문화제에 대한 설명글이 나타나있다. 이는 (교산허균 문화제, 설명글, “강릉은 예로부터 면면이...”)와 같은 RDF 트리플 하나로 표현된다. 본 연구에서는 목적어 부분에 있는 텍스트가 교산허균 문화제의 내용을 표현한 설명글로 간주한다. 전시회와 관광지의 구분은 타입(rdf:type) 술어를 가진 RDF 트리플을 통해 구분할 수 있다. 예를 들어, (교산허균 문화제, 타입, Event)와 (허균・허난설헌 기념공원, 타입, Attraction)과 같이 주어의 종

1) http://data.visitkorea.or.kr/linked_open_data

류를 명시한 RDF 트리플이 존재한다. Event는 전시회로 Attraction은 관광지로 간주한다. 2017년 7월 버전의 한국관광공사 LOD에는 1,902개의 전시회 그리고 13,494개가 관광지가 존재한다.

Ⅳ. 전시회-관광지 관련도

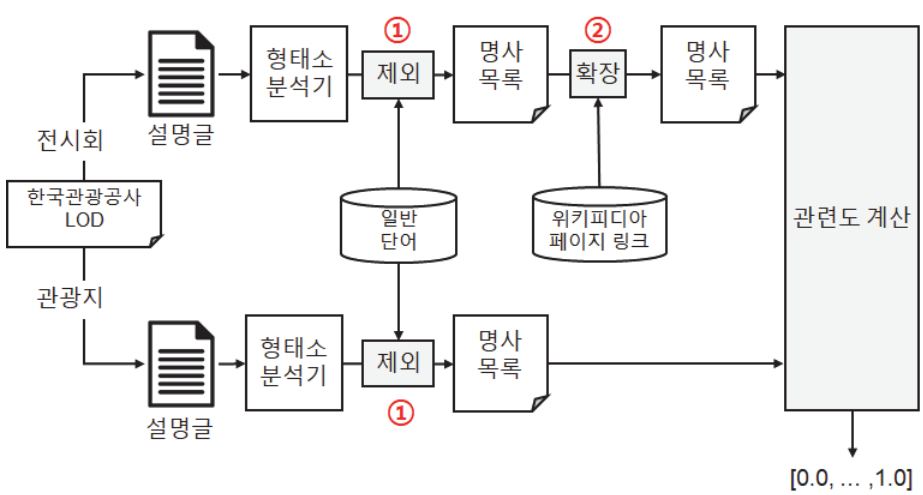
KTO-LOD 데이터셋에는 전시회와 관광지에 대한 정보가 존재한다. 하지만, 서로 독립적으로 존재하고 연결돼 있지 않다. 사용자에게 관광지를 추천하기 위해서는 내용이 유사한 전시회와 관광지가 연결돼 있어야 한다. 전시회와 관광지에 대한 사진, 위치, 입장시간, 주차 등에 대한 정보가 있는데, 내용을 가장 잘 대표하는 것은 설명글이다. 본 연구에서는 설명글로부터 단어를 추출하여 관련도를 계산하는 방법을 사용한다.

<그림 2>를 바탕으로 전시회와 관광지를 연결하는 과정에 대해 개념적으로 설명한다. KTO-LOD 데이터셋으로부터 모든 전시회와 관광지 설명글을 추출한다. 즉, 술어가 "설명글"인 RDF 트리플만 추출한다. 각 전시회 및 관광지에 대한 설명글만 모을 수 있다. 형태소 분석기(박상원, 2010)를 사용해서 설명글에서 명사만 추출한다. 형태소 분석기는 자연언어로 된 문장이 주어지면 단어를 구분하고 각 단어의 품사를 식별하는 소프트웨어이다. 다음의 2개의 단계를 거쳐서 명사 목록을 갱신한다.

① 지나치게 일반적이어서 전시회와 관광지의 관련도를 계산함에 있어서 관련이 없는 일반 단어들은 제외시킨다. 예를 들면, "박물관"이란 단어는 많은 전시회의 설명글에서 언급되는 단어이기 때문에 제외시킨다(1절에서 자세히 설명).

② 전시회 설명글로부터 획득한 명사에 대해서는 위키피디아 페이지 링크 데이터셋을 활용하여 확장을 한다(2절에서 자세히 설명). 예를 들면, 전시회 설명글에 "허균"이란 단어가 있고 위키피디아 페이지 링크 데이터셋에 (허균, 홍길동)이란 연관어 쌍이 있을 경우, "홍길동"이란 단어를 전시회의 명사 목록에 추가한다. 즉, 연관어 기반으로 명사를 확장한다.

마지막 단계에서는 ①과 ② 과정을 거쳐서 얻은 전시회와 관광지의 명사 목록을 비교하여 겹치는



<그림 2> 한국관광공사 LOD의 전시회와 관광지의 관련도를 계산하는 개념도

정도에 따라 관련도를 계산한다. 관련도는 0.0에서 1.0 사이의 실수로 산출된다.

전시회와 관광지의 관련도를 계산할 때 개별 단어의 중요도를 반영하기 위해 정보 검색 시스템에서 널리 활용되고 있는 TF-IDF값을 활용한다(Jiaul H. Paik, 2013; Kewen Chen et al., 2016). $tf(w,d)$는 문서 d에 단어 w가 언급된 횟수를 의미한다. 많이 언급될수록 중요한 단어이다. $idf(w,D)$는 D에 있는 문서의 개수를 w가 전체 문서들에서 출현한 횟수를 나눈 값이다. 적은 수의 문서들에서 언급된 단어가 많은 수의 문서들에서 언급된 단어보다 중요하다는 의미이다. TF-IDF는 다음과 같이 TF와 IDF를 곱한값이다.

$$tfidf(w,d,D) = tf(w,d) \times idf(w,D)$$

적은 수의 문서에 많이 언급된 단어에 더 가중치를 준다. 상기 설명한 TF-IDF값을 기반으로, 전시회와 관광지의 관련도를 계산하는 3가지 방법을 고안했다.

1. 공통 용어 기반 관련도(CT)

공통 용어 기반 관련도(CT)는 전시회 e와 관광지 t의 단어 목록을 비교하여 공통된 단어의 가중치 곱의 합으로 계산된다. 즉, 전시회와 관광지가 공유하는 단어가 많을수록 관련도가 높다.

$$\frac{1}{Z}\sum_{w \in C} tfidf(w,e,LOD) \times tfidf(w,t,LOD)$$

여기서, C는 e와 t의 단어 목록의 교집합, Z는 관련도를 0.0에서 1.0 사이의 값으로 만들기 위한 정규화 상수이고, LOD는 전시회와 관광지의 설명글(하나의 문서로 간주)로 구성된 전체 문서 집합이다.

예제 1) "어둠 속 대화"라는 전시회 e와 "가나 아트센터"라는 관광지 t가 있다고 가정하자. e의 단어 목록은 {어둠, 상황, 초대, 화면}이고 t의 단어 목록은 {갤러리, 초대, 주제, 화면}이다. 이 경우, C는 {초대,화면}이 된다. CT 관련도는 다음과 같이 계산된다.

$$\frac{1}{Z}(tfidf(\text{화면}, e, LOD) \times tfidf(\text{화면}, t, LOD) + tfidf(\text{초대}, e, LOD) \times tfidf(\text{초대}, t, LOD))$$

예제 1에서도 확인할 수 있듯이, CT는 공통 단어만 고려하면 되기 때문에 계산하기 쉬운 장점이 있지만, 매우 일반적인 단어가 있을 경우 관련도값에 왜곡이 있을 수 있는 단점이 있다. "화면"과 "초대"라는 단어는 구체성이 떨어지는 단어로, 다수의 전시회와 관광지에서 사용되는 단어이다. 실제로, "어둠 속 대화"라는 전시회와 "가나 아트 센터"는 특별한 관련이 있어 보이지 않는다.

이 문제를 해결하기 위해 구체성이 떨어지는 단어는 제외시키는 방법(CT-F)을 고안했다. <그림 2>의 과정 ①에 해당한다.

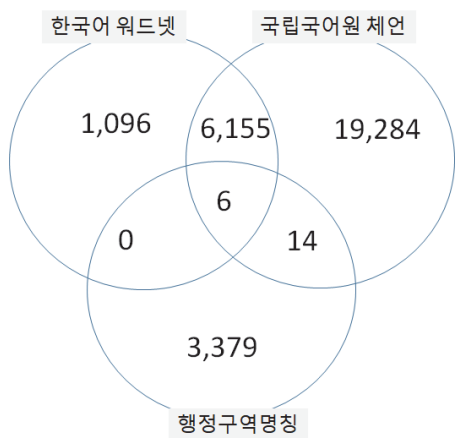
$$\frac{1}{Z}\sum_{w \in F(C)} tfidf(w,e,LOD) \times tfidf(w,t,LOD)$$

CT의 수식과의 차이점은 C 대신에 $F(C)$를 사용했다는 점이다. $F(C)$는 $C - G$ 이고 G는 일반 단어 집합이다. 즉, 공통된 단어들에서 일반 단어는 제외시킨다. 본 연구에서는 일반 단어 집합으로 다음의 언어자원들을 활용했다.

국립국어원 체언[2]은 일상생활에서 빈번하게 사용되는 명사, 동사, 형용사 등의 집합이다. 본 연구에서는 25,458개 명사를 사용했다. 한국어 워드넷[3]

은 영어 워드넷을 한국어로 번역한 의미단위(synset) 사전이다(Key-Sun Choi, 2014; 윤애선, 2009). 9,714개의 의미단위는 명사는 7,257개(74%)이고 동사는 2,457개(26%)로 구성됐다. 본 연구에서는 명사만 사용했다. 한국행정구역 용어집[4]은 한국의 읍, 면, 동, 구, 시 등의 3,400개의 행정구역 명칭으로 구성됐다.

<그림 3>은 위의 3개의 언어자원에서 공통으로 언급된 단어들의 개수이다. 공통 단어가 많지 않다. 어느 하나의 언어자원으로는 불충분하기 때문에, 본 연구에서는 3개의 언어자원 모두를 사용한다.



<그림 3> 언어자원들에 공통으로 포함된 단어의 개수

예제 2) 예제 1과 같은 예제에서, 만약 G에 "화면", "초대" 등의 단어가 포함됐다면 $F(C)$는 공집합이 되고, CT에 의한 관련도는 0.0이 된다.

2. 위키피디아 기반 관련도(WP)

위키피디아 기반 관련도(WP)는 전시회와 관광지의 단어 목록에 공통된 단어가 없음에도 단어 확장을 통해 관련도를 계산할 수 있는 방법이다. <그림 2>의 과정 ②에 해당한다.

위키피디아는 누구나 자유롭게 수정이 가능한 온라인 백과사전이다. 2017-07-24일 현재 5,045,856개의 영어 위키피디아 페이지와 332,451개의 한국어 위키피디아 페이지가 있다. 본 연구에서는 페이지 링크 데이터셋을 활용했다. 페이지 링크 데이터셋은 (페이지 제목, 링크 단어)쌍의 목록으로 구성됐다. 페이지 제목과 링크 단어는 연관어 쌍으로 볼 수 있다. 예를 들어, "허균"을 제목으로 하는 위키피디아 페이지는 <그림 4>와 같다. 파란색으로 표시된 단어(조선, 서자, 홍길동전 등)는 다른 위키피디아 페이지에 대한 링크이다. 따라서, 다음과 같은 연관어 쌍의 목록을 얻을 수 있다. (허균, 조선), (허균, 서자), (허균, 홍길동전). 본 연구에서는

2) https://ithub.korean.go.kr/user/electronicDic/electronicDicManager.do
3) http://wordnet.kaist.ac.kr/
4) http://kssc.kostat.go.kr/ksscNew_web/kssc/common/CommonBoardList.do?gubun=1&strCategoryNameCode=019&strBbsId=kascrr&categoryMenu=014

<그림 4> 허균에 대한 한글 위키피디아 페이지의 스크린샷.

위키피디아 페이지 링크 데이터셋 20170601 버전을 활용했다.[5] 총 44,284,699개의 연관어 쌍을 얻을 수 있다.

위키피디아를 활용한 WP에 의한 관련도는 다음과 같이 계산된다.

$$\frac{1}{Z}\sum_{(w_a,w_b)\in H} tfidf(w_b, Wiki(w_a), WIKI) \times tfidf(w_b, t, LOD)$$

여기에서 $WIKI$는 위키피디아 전체 문서 집합이다. H는 다음과 같이 정의된다.

$$H = \left\{ (w_a, w_b) : w_b \in Wiki(w_a) \text{ and } \begin{matrix} w_a \, in \ e \text{ and } w_b \, in \, t \end{matrix} \right\}$$

여기에서 $Wiki(w_a)$는 w_a라는 위키피디아 페이지를 가리키며 해당 페이지에 있는 링크 단어들의 집합이다. 즉, 전시회의 단어들의 연관어가 위키피디아 페이지 링크 데이터셋에 존재하는 경우 그 연관어를 전시회 단어들에 포함시켜서 확장한 후 관광지의 단어들과 공통 단어를 찾아서 구한 단어 쌍이 H이다. 전시회와 관광지에 공통적으로 포함되지 않는 2개의 단어(w_a와 w_b)가 위키피디어 페이지 링크를 통해서 연결된 것이다. H 대신에 $F(H)$를 사용하는 방법, 즉, 일반 단어는 제외시키는 방법은 WP-F로 칭한다.

예제 3) "다산 축제"라는 전시회 e와 "황사영 기념비"라는 관광지 t가 있다고 가정하자. e의 단어 목록은 {다산, 정약용, 실학}이고 t의 단어 목록은 {가톨릭, 신유박해}이다. $Wiki$(정약용)는 {신유박해}라 하자. 이 경우, H는 {(정약용, 신유박해)}이다. 따라서, WP에 의한 관련도는 다음과 같이 계산된다.

$$\frac{1}{Z}(tfidf(\text{신유박해}, Wiki(\text{정약용}), WIKI) \times tfidf(\text{신유박해}, t, LOD))$$

3. 하이브리드 관련도(CTWP)

WP에 의해 CT의 단점이 해결됐지만 그렇다고

5) https://dumps.wikimedia.org/kowiki/20170601/

WP만 할 경우 공통되는 단어에 대한 중요도가 전혀 고려가 되지 않기 때문에 둘 다 고려하는 방법이 필요하다. 하이브리드 관련도(CTWP)는 CT와 WP를 모두 고려하는 방법이다. CTWP는 다음과 같이 계산된다.

.

$$\frac{1}{Z}(\alpha \times CT + \beta \times WP)$$

CTWP-F는 CT 대신에 CT-F 그리고 WP 대신에 WP-F를 사용하여 계산된다. α와 β의 값은 응용에 따라 달리 정한다. CT가 더 중요할 경우 α을 크게 설정하고 WP가 더 중요할 경우 β를 크게 설정한다.

Ⅴ. 관광지 추천 시스템 구현

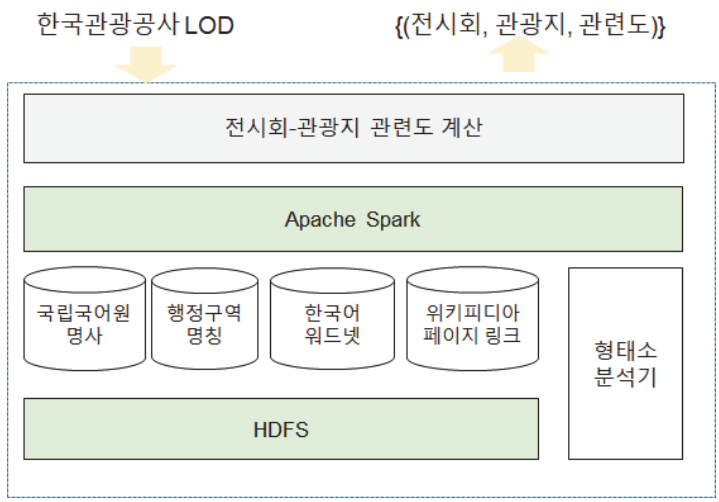
지금까지 기술한 전시회-관광지 관련도 계산 방법을 어떻게 구현했는지를 우선 설명하고, 이를 기반으로 내용기반 관광지 추천 기능을 제공하는 전시회 정보 관리 시스템을 어떻게 구현했는지를 설명한다.

전시회-관광지 관련도 계산 시스템을 우선 단일 컴퓨터 기반으로 구현했다. CT와 CT-F 방법의 경우 6분이 소요됐다. 방법의 이름에서 F는 일반 단어 제외 기능이 추가된 방법을 의미한다. WP, WP-F, CTWP, CTWP-F의 경우 모두 메모리 부족 문제로 작동하지 않았다. 이 방법들은 CT와 CT-F와 달리 위키피디아 페이지 링크 데이터를 사용한다. 단일 컴퓨터의 8 GB 메모리로는 위키피디아 페이지 링크 데이터를 처리할 수 없기 때문에 실패했다.

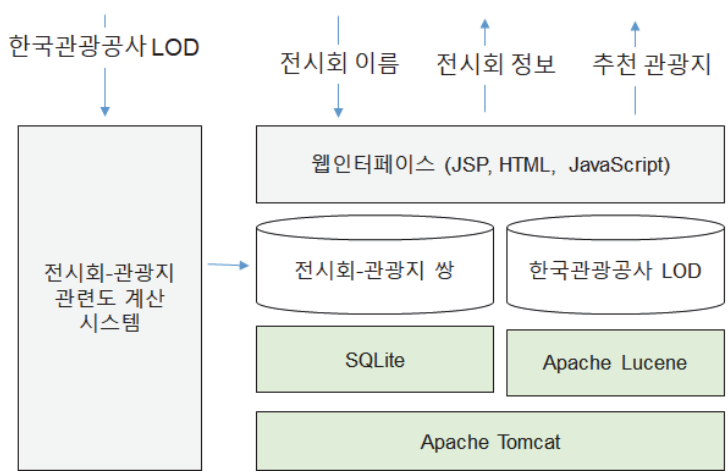
이 문제를 해결하기 위해 여러 대의 컴퓨터로 구성된 클러스터에 작동하는 시스템으로 다시 구현했다. 단일 컴퓨터 기반 시스템과 달리 클러스터에서 작동하는 시스템의 경우 데이터를 여러 대의 컴퓨터에 분산을 시키는 방법이 필요하다. 본 연구에서는 정보 검색 분야에서 널리 사용되는 단어 기반 역-인덱스 분할 기법(term-based inverted index partitioning)을 사용했다(Jeyavaishnavi Muralikumar et al., 2017; Abdullah Gani et al., 2016; B. Barla Cambazoglu et al., 2013; Karen Spärck Jones, 1972). 수많은 연관어 쌍으로 구성된 위키피디아 페이지 링크 데이터에 있는 단어들을 알파벳을 기준으로 분할하고 컴퓨터들에 분배를 함으로써, 하나의 컴퓨터가 전체 데이터를 처리할 필요가 없게 하는 방법이다. 단어 기반 역-인덱스 분할 기법을 구현하기 위해 <그림 5>와 같이 아파치 스파크(Apache Spark)에 기반하여 구현했다. 아파치 스파크는 최근에 활발히 사용되고 있는 클라우드 컴퓨팅 프레임워크로서 여러 대의 컴퓨터로 구성된 클라우드 환경에서 작동하는 프로그램을 쉽게 개발할 수 있는 개발환경을 제공한다(Matei Zaharia, 2012). 위키피디아 페이지 링크를 포함한 언어자원들은 아파치 스파크에서 사용하는 HDFS (Hadoop Distributed File System)에 저장한다. HDFS는 분산파일시스템의 일종으로 사용자에게는 마치 단일 컴퓨터의 파일을 저장하는 것처럼 보이지만 내부적으로는 클라우드에 자동으로 분산 저장을 해주는 시스템이다. 따라서, 대용량 파일도 클라우드에 쉽게 저장할 수 있다. 한편, <그림 5>의 전시회-관광지 관련도 계산 모듈의 내부 과정은 4장 및 <그림 2>를 참조한다.

구현된 전시회-관광지 관련도 계산 시스템을 기반으로 웹기반 전시회 정보 관리 시스템을 구현했다. <그림 6>에서 왼쪽의 전시회-관광지 관련도 계산 시스템은 <그림 5>에 해당된다. 전시회-관광지 관련도 계산 시스템의 출력인 전시회-관광지 쌍이 데이터베이스에 저장된다. 사용자로부터 전시회 정보 요청이 들어오면 이 데이터베이스로부터 추천 관광지를 가져와서 보여준다. 이와 같은

데이터 흐름은 MVC (Model-View-Controller) 모델을 따른다(J. Wojciechowski, 2004). MVC 모델은 특히 웹기반 시스템에서 널리 사용되는 시스템 디자인 패턴의 하나이다. Model은 시스템에서 다루는 데이터 구조를 가리키며, 본 시스템에서는 한국관광공사 LOD, 언어자원, 전시회-관광지 쌍 등에 해당된다. Controller는 데이터 가공처리 모듈을 가리키며, 본 시스템에서는 전시회-관광지 관련도 계산 시스템에 해당한다. 즉, 핵심 계산 모듈은 독립적인 시스템으로 구현하는 것이다. View는 사용자와 상호작용하는 수단을 가리키며, 본 시스템에서는 JSP, HTML 등으로 구현된 웹인터페이스를 가리킨다.



<그림 5> 클러스터 기반의 전시회-관광지 관련도 계산 시스템의 구조도



<그림 6> 웹기반 전시회 정보 관리 시스템의 구조도

전시회-관광지 관련도 계산 시스템의 출력인 전시회-관광지 쌍은 SQLite 데이터베이스(Michael Owens, 2010)에 적재된다. 다양한 데이터베이스가 존재하는데 본 연구에서는 복잡한 질의 처리가 필요하지 않기 때문에 쉽게 설치 및 사용이 가능한 SQLite를 선택했다. 아파치 루씬(Apache Lucene)은 문서검색엔진(Andrzej Białecki, 2012)으로, 사용자가 키워드를 입력했을 때, 해당 키워드를 포함하는 전시회 목록을 검색하는 역할을 한다. 아파치 톰캣(Apache Tomcat)은 사용자 웹브라우저와의 메시지 교환을 처리하는 엔진이다.

<그림 7>은 본 연구에서 구현한 웹기반 전시회 정보 관리 시스템의 스크린샷이다. 교산허균 문화제에 대한 정보가 화면 왼쪽에 보여지고, 추천된 관광지가 화면 오른쪽에 보여진다. 왼쪽의 화면은 한국관광공사 LOD에서 제공하는 전시회 정보를 보여준다. 교산허균 문화제에 대해서 추천된 관광지로 이매창묘, 매창공원, 허난설헌 솔숲 등이 추천된 것을 볼 수 있다. 4장에서 제안된 관련도 계산 방법들 중 CTWP-F의 결과이다. 관광지 이름의 오른쪽의 숫자는 교산허균 문화제와 관광지의 관련도를 의미한다. 관광지 아래의 정보는 어떤 단어들에 의해서 관련이 맺어졌는지에 대한 근거를 보여준다. 관련 근거 정보는 2개의 행으로 구성됐다. 첫 번째 행은 CT-F의 근거 해당하고 두 번째 행은 WP-F에 해당한다. 예를 들어, 허난설헌 솔숲의 첫 번째 행은 교산허균 문화제의 설명글과 허난설헌 솔숲의 설명글에서 허균, 강릉, 문화제 등의 단어가 공통적으로 나타났다는 의미이다. 그리고, 허난설헌 솔숲의 두 번째 행에서 (홍길동, 허균)의 의미는 홍길동이란 단어가 교산허균 문화제에 나타났고 동시에 위키피디아 페이지의 제목이고, 허균은 해당 페이지의 링크 단어이면서 허난설헌 솔숲의 설명글에도 나타났다라는 의미이다.

<그림 7> 웹기반 전시회 정보 관리 시스템의 스크린샷

VI. 논 의

본 연구에서 제안한 시스템의 성능을 효율성과 효과성을 평가한 실험 결과를 기술하고 관광지 추천 예제에 대해 논의한다.

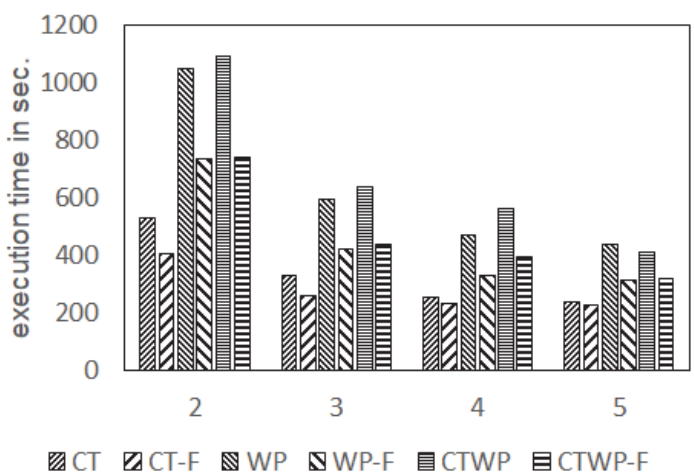
1. 효율성 평가

효율성은 모든 전시회-관광지 쌍에 대한 관련도를 계산하는 데에 걸리는 시간을 가리킨다. 실험은 5개의 컴퓨터로 구성된 클러스터에서 수행

했다. 각 컴퓨터의 하드웨어 사양은 3.10 GHz CPU와 8 GB 메모리이다. 여기에서 기술하는 모든 실험은 10번의 동일한 실험에 대한 평균이다.

각 관련도 계산 방법이 소비한 시간은 다음과 같다. CT(230초), CT-F(215초), WP(390초), WP-F(285초), CTWP(473초), CTWP-F(300초). CT-F가 가장 빨랐고 CTWP가 가장 느렸다. CTWP의 경우 위키피디아 페이지 링크 데이터는 사용하지만 CTWP-F와 달리 일반 단어를 제외시키지 않기 때문에 더 많은 단어들을 처리해야 한다. 본 실험 결과가 의미하는 바는, 고려해야하는 단어의 개수가 많을수록 관련도 계산은 오래 걸린다는 것이다.

<그림 8>은 클러스터에 참여하는 컴퓨터의 개수에 따른 소요시간을 나타낸다. X축은 컴퓨터의 개수를 의미한다. 예를 들어, X축 값에서 2의 경우 2개의 컴퓨터로 구성된 클러스터에서 실험했다는 의미이다. 많은 컴퓨터를 사용할수록 시간이 줄어듦을 알 수 있다. 하지만, 컴퓨터 2개와 3개에 의한 차이보다 컴퓨터 4개와 5개에 의한 차이는 작다. 이러한 현상은 단어의 개수와 관계가 있다. 본 연구에서 사용한 언어자원의 규모의 경우 5개 이상의 컴퓨터를 사용할 정도의 규모는 아니라는 의미이다. 물론, 추가적인 언어자원을 사용하여 규모가 커질 경우 컴퓨터 추가에 따른 성능 향상을 기대할 수 있다.



<그림 8> 컴퓨터의 개수에 따른 효율성

2. 효과성 평가

효과성은 특정 전시회에 대해 추천된 관광지들이 얼마나 유의미한지를 가리킨다. 전시회-관광지 연계 데이터가 존재하지 않기 때문에 본 시스템에서 추천한 관광지를 관광분야 전문가가 평가하는 방법으로 실험을 수행했다. 전문가가 수작업으로 평가하는 방식이기 때문에 한국관광공사 LOD의 모든 전시회 및 관광지에 대해 평가할 수 없다. 대신, 관광지식정보시스템에서 제공하는 "2016년 문화관광축제"[6]에 있는 전국의 43개의 축제를 실험 대상으로 삼았다. 43개의 축제 중 한국관광공사

6) https://know.tour.go.kr/stat/cultureTourFestivalReportDis.do

LOD에 존재하는 축제는 완주와일드푸드축제, 논산강경젓갈축제, 울산옹기축제 등 41개이다.

평가 대상 추천 방법은, 한국관광공사 LOD 웹 서비스에서 제공하는 위치기반 추천(KTO-LOD)과 본 논문에서 제안한 내용기반 추천 방법인 CT-F, WP-F, CTWP-F이다. 일반 단어를 제외하지 않는 방법들(CT, WP, CTWP)은 무의미한 추천이 많기 실험 대상에서 제외하였다. KTO-LOD 방법의 경우 축제가 열리는 장소를 제외한 거리가 가까운 순으로 상위 3개의 관광지를 선택하였다. 내용기반 추천 방법의 경우 관련도가 높은 순으로 상위 3개를 선택하였다.

전문가에게 익명화된 각 추천 방법에 대해 41개의 전시회 당 3개의 관광지를 제시하고 관련이 있으면 O, 관련이 없으면 X 그리고 모호하면 ?를 표시하게 하였다. <표 1>은 41개의 전시회에 대한 총 123개의 관광지 추천에 대한 전문가 평가결과이다. CTWP-F의 경우 관련 있음의 개수가 가장 많았고 KTO-LOD의 경우가 가장 적었다. KTO-LOD가 위치기반 방법임에도 불구하고 CT-F보다 관련있음의 개수가 많은 것은, 전시회가 내용이 관련이 있는 장소 근처에서 열리는 경우가 많기 때문이다. 하지만 그렇지 않은 경우 전혀 관련이 없기 때문에, KTO-LOD의 경우 관련없음이 가장 많았다. 또한, KTO-LOD의 경우 위치기반이기 때문에 모든 전시회에 대한 추천 결과가 있었다. 반면 CT-F의 경우 추천 결과 없음이 가장 많았다. 추천 결과 없음이 15라는 의미는, 전시회당 3개의 관광지를 추천하게 돼 있기 때문에 약 5개의 전시회에 대해 추천 결과가 3개 미만이라는 의미이다. 설명글이 짧거나 추출되는 명사가 적어서 공통단어를 찾을 수 없기 때문이다. WP-F와 CTWP-F의 경우 위키피디아 연관어를 고려하기 때문에 CT-F에 비해서 추천 결과없음이 적다. WP-F와 CTWP-F에서 관련없음이 나온 것은 위키피디아 페이지 링크로부터 추출한 연관어의 경우 간접적인 연관어도 존재하기 때문이다. 예를 들어, “낙동강”이라는 위키피디아 페이지를 보면 링크단어에 굴포천, 안양천 등의 한강의 지류도 포함돼 있다. 낙동강과 굴포천은 강 또는 하천이라는 공통점이 있지만 직접적으로 관련성은 없다. 실제 관광지 추천 예제는 3절에서 살펴본다.

<표 1> 관광지 추천의 효과성 평가 결과

	KTO-LOD	CT-F	WP-F	CTWP-F
관련있음	38	26	47	**50**
관련없음	**75**	73	52	49
모름	10	9	10	**11**
추천 결과 없음	0	**15**	14	13

한편, 전시회-관광지의 관련이 맺어질 때, 단 하나의 공통 단어(또는 위키피디아 페이지 링크에 의한 공통 단어)에 의해 맺어질 수도 있고 많은 수의 공통 단어에 의해 맺어질 수도 있다. <표 2>는 한국관광공사 LOD에 존재하는 전시회 및 관광지에 대한 통계치이다. 2,000여 개의 전시회의 설명글에 있는 명사의 평균 개수는 36.5개이다라는 의미이다. 어떤 전시회의 경우 설명글이 매우 길어서 251개의 명사로 구성됐다. 13,000여 개의 관광지가 존재했고 설명글의 평균 명사의 개수는 54.0개이다.

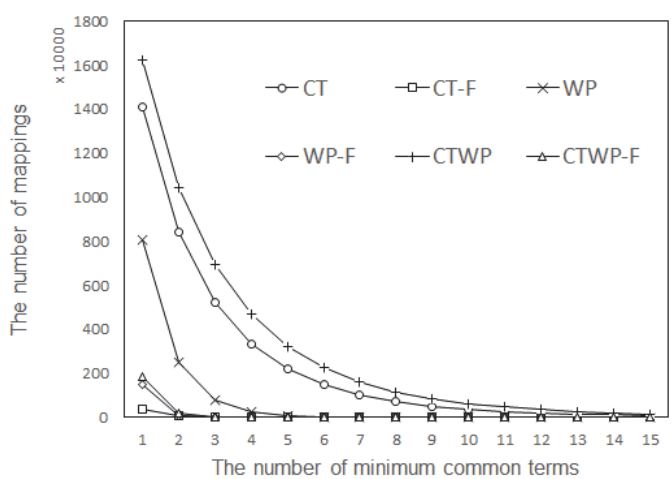
<그림 9>는 공통 단어 개수에 따라 얼마나 많은 전시회-관광지가 관련이 맺어졌는지를 나타낸

다. 특정 응용을 고려하지 않고 일반적인 상황을 고려하기 위해 CT와 WP의 중요도가 동일하다고 가정하고 CTWP와 CTWP-F에서의 가중치를 $\alpha = \beta = 0.5$로 설정했다. X축은 공통 단어의 개수이다. 예를 들어, X축 값 1의 경우 WP 방법에 의해 약 800개의 전시회-관광지 쌍의 관련도가 0.0보다 크다는 것을 의미한다. 15개의 공통 단어를 가진 전시회-관광지 쌍도 존재함을 알 수 있다. CTWP보다 CTWP-F의 개수가 작은 것은 일반 단어를 제외시킴으로써 공통 단어가 줄어들었고 결국 관련도값이 감소했기 때문이다.

최소 공통 단어 개수를 조정함에 따라 관련된 관광지가 많이 추천될 수도 있고 적게 추천될 수도 있다. 보다 관련이 있는 관광지를 추천하고 싶은 경우 최소 공통 단어 개수를 크게 설정하면 되고, 덜 관련이 있는 관광지도 추천하고 싶은 경우 최소 공통 단어 개수를 작게 설정하면 된다. 서비스 대상에 따라 적절한 값을 정해야 한다.

<표 2> 한국관광공사 LOD의 전시회-관광지 통계치

	개수	설명글의 명사 개수		
		평균	최대	최소
전시회	1,902	36.5	251	3
관광지	13,494	54.0	846	1



<그림 9> 공통 단어 개수에 따른 전시회-관광지 쌍의 수

3. 관광지 추천 예제

<표 3>은 “금련산 은하축제”에 대해 각 방법에 의해 추천된 관광지 목록이다. KTO-LOD는 한국관광공사 LOD 서비스에 의해 추천된 결과로서, 축제 장소와 거리 정보도 제공한다. 나머지는 본 연구에서 제안한 방법이다.

KTO-LOD의 경우 부산광역시 금련청소년수련원, 금련산, 광안리해수욕장 등을 추천했는데, 금련산 은하축제가 천문에 관련된 축제라는 것을 감안

<표 3> 금련산 은하축제에 대해 추천된 관광지

추천 방법	추천된 관광지
KTO-LOD	부산광역시 금련청소년수련원 (120m)
	금련산 (966m)
	광안리해수욕장(1533m)
CT-F	부산광역시 금련청소년수련원 [금련산], [금련산청소년수련원], [천문대]
WP-F	여주 영릉 (천문대, 측우기, 천문대, 관천대)
	국립고궁박물관 (천문대, 앙부일구)
	송현근린공원 (금련산, 송림산)
CTWP-F	창경궁 [천문대] (천문대, 측우기), (천문대.간의), (천문대.첨성대), (천문대.창경궁)
	서울 관상감 천문대 [천문대] (천문대, 첨성대), (천문대, 관천대), (천문대. 창경궁)
	김해천문대 [천문대] (천문대,김해천문대), (천문대,첨성대)

하면 관련이 없는 관광지라 할 수 있다. 축제 장소 주변이라는 것 이외에는 연관성이 없다. 이러한 관광지는 해당 장소에서 지도 서비스를 사용해서 알 수 있는 관광지들이다.

CT-F의 경우 역시 부산광역시 금련청소년수련원를 추천했다. 그 근거는 (금련산, 금련산청소년수련원, 천문대) 단어가 공통으로 포함됐기 때문이다. 축제 설명글에 장소에 대한 단어가 있기 때문에 공통 단어에 의해 추천됐다.

WP-F의 경우 여주 영릉, 국립고궁박물관, 송현근린공원 등을 추천했다. 여주 영릉의 경우 측우기를 개발한 세종의 능이고 측우기는 천문대와 관련이 있다. 천문대 위키피디아 페이지에 측우기라는 링크 단어가 있었기 때문에 가능한 추천이다. 은하축제 관람객들에게 여주 영릉을 추천함으로써 우리나라 천문학의 역사를 상기시킬 수 있는 추천이다. 마찬가지로, 국립고궁박물관의 경우 앙부일구를 소장하고 있기 때문에 은하축제 관람객에게는 중요한 관광지이다.

CTWP-F의 경우 창경궁, 서울 관상감 천문대, 김해천문대 등을 추천했다. 창경궁에는 관천대라는 천문관측대가 있다. WP-F와의 차이는 천문대라는 단어이다. 즉, CTWP-F의 경우 CT도 고려하는 것이기 때문에, 금련산 은하축제가 가지는 천문대라는 단어를 역시 가지는 관광지들을 WP-F보다 더 가중치를 높게 고려한다. WP-F에 의한 추천보다 더 직접적으로 관련이 있는 추천이라 볼 수 있다.

또 다른 예로서 <표 4>는 “보성다향축제”에 대해 각 방법에 의해 추천된 관광지 목록이다.

KTO-LOD의 경우 보성천문과학관, 보성녹차밭 대한다원, 봇재다원 등을 추천했다. 보성녹차밭 대

한다원의 경우 축제의 내용과 관련이 있다. 축제가 관련된 장소에서 개최된 것이기 때문에 KTO-LOD에 의한 의도적인 관광지 추천이라고 보기 어렵다.

CTWP-F의 경우만 살펴보면, 경남 하동 악양(슬로시티)가 추천됐다. 관광지 이름만 보고는 관련성을 알기 어렵다. 관련성은 우리나라 역사를 통해 알 수 있다. 경남 하동 지방은 신라 시대의 지역이고 그 지방에서 재배되는 녹차는 흥덕왕 때 수입이 됐다. 구체적인 역사가 경남 하동 악양(슬로시티)의 설명글과 위키피디아에 기술이 돼 있었기 때문이 이러한 추천이 가능하다.

<표 4> 보성다향축제에 대해 추천된 관광지

추천 방법	추천된 관광지
KTO-LOD	보성천문과학관 (220 m)
	보성녹차밭 대한다원 (549 m)
	봇재다원 (838 m)
CT-F	보성녹차밭 대한다원 [차산업],[다전],[다신제]
	국립 낙안민속자연휴양림 [다향제]
WP-F	파사석탑 (차나무,삼국유사),(차나무,아유타국)
	서천 마량리 동백나무 숲 (차나무,차나무과)
	해은사(김해) (차나무,김해),(차나무,아유타국)
CTWP-F	경남 하동 악양(슬로시티) [차나무] (차나무,대렴),(차나무,흥덕왕)
	쌍계사차나무시배지 [차밭] (차나무,대렴),(차나무,흥덕왕),(차나무,당나라)

Ⅶ. 결 론

본 연구에서는 기존 전시회 정보 관리 시스템의 위치기반 관광지 추천이 전시회의 내용과 관련이 없는 관광지도 추천할 수 있다 점에 착안하여, 전시회의 내용과 관련이 있는 내용기반 관광지 추천 방법을 제안했다. 전시회 및 관광지의 설명글을 바탕으로 관련된 전시회-관광지 쌍을 찾기 위해 국립국어원, 행정구역명칭, 위키피디아 등의 언어자원을 활용했다. 대용량 언어자원을 효율적으로 처리하기 위해 클라우드 컴퓨팅 프레임워크를 활용하여 구현했다. 웹 기반의 전시회 정보 관리 시스템을 구현하여 실제 서비스가 가능함을 보였다.

본 연구의 한계점은 다음과 같다. 1) 전시회와 관광지의 설명글을 기반으로 추천을 하기 때문에, 설명글이 존재하는 경우로 추천의 범위가 제한됐다. 본 연구에서는 한국관광공사 LOD에 존재하는 설명글을 활용하였는데, 향후 연구에서는

웹에 공개된 문서들로부터 설명글을 자동으로 수집하는 모듈을 추가하여 추천 범위를 넓히는 방안을 고려하고 있다. 2) 설명글로부터 명사만 추출하여 비교하였는데, 동음이의어를 처리하지 못한다는 문제점이 있다. 이를 해결하기 위해 자연어처리 분야에서 활발히 연구되고 있는 word sense disambiguation 기법(Ignacio Iacobacci, 2016)을 활용하는 방안을 고려하고 있다. 3) 본 연구에서는 전시회의 내용과의 관련도에 의해서만 관광지를 추천하였다. 관광의 경우 이동하는 거리도 중요하기 때문에 전시회가 열리는 장소로부터의 거리도 고려할 필요가 있다. 향후 연구에서는 위치와 내용을 모두 고려하는 관광치 추천 방법을 고안할 계획이다.

본 연구에서 제안하는 내용기반 관광지 추천은 전시회와 관광 산업 간의 시너지 효과를 발생시킴으로써 두 산업이 동시에 발전하는 데에 기여할 수 있다. 관람객 및 관광객 입장에서는 전시회에서 배운 내용을 관광지에서 상기할 수 있기 때문에 학습의 효과가 있다. 전시회 개최측 입장에서는 본 연구에 의해 추천된 관광지에 전시회를 개최함으로서 전시회를 더욱 의미있게 구성할 수도 있다(오창호 등, 2011; 오창호 등 2012).

참고문헌

1. 박상원・최동현・김은경・최기선(2010), “플러그인 컴포넌트 기반의 한국어 형태소 분석기,” *제22회 한글 및 한국어 정보처리 학술대회 논문집*, pp.197-201.
2. 박연진・송경아・황재원・창병모(2015), “온톨로지 기반의 개인화된 여행 추천 시스템의 구현,” *한국콘텐츠학회논문지*, 15(9), pp.1-10.
3. 변상우(2015), “관광지 선택 동기가 관광지 이미지, 재방문의도에 미치는 영향에 관한 연구 — 감천문화마을을 중심으로 —,” *경영과 정보연구*, 34(3), pp.197-213.
4. 오창호・남경화・공기열(2011), “Kano모형을 이용한 컨벤션서비스의 요인별 평가와 서비스 회복에 관한 연구,” *경영과 정보연구*, 30(2), pp.57-79.
5. 오창호・육풍림・황재위・강선구(2012), “Means-End Chain과 Laddering을 이용한 컨벤션도시의 브랜드가치 개발에 관한 연구,” *경영과 정보연구*, 31(2), pp.253-272.
6. 유성열・이강배(2013), “유비쿼터스 기반의 컨벤션 서비스 모델,” *경영과 정보연구*, 32(5), pp.89-100.
7. 윤애선・황순희・이은령・권혁철(2009), “한국어 어휘의미망 KorLex 1.5의 구축,” *정보과학회논문지: 소프트웨어 및 응용*, 36(1), pp.92-108.
8. Abdullah Gani, Aisha Siddiqa, Shahaboddin Shamshirband, and Fariza Hanum(2016), “A survey on indexing techniques for big data: taxonomy and performance evaluation,” *Knowledge and Information Systems*, 46(2), pp.241-284.
9. Andrzej Białecki, Robert Muir, and Grant Ingersoll(2012), “Apache lucene 4,” *SIGIR 2012 workshop on open source information retrieval*, pp.17-24.
10. B. Barla Cambazoglu, Enver Kayaaslan, Simon Jonassen, and Cevdet Aykanat(2013), “A term-based inverted index partitioning model for efficient distributed query processing,” *ACM Trans. Web*, 7(3), Article 15, pp.1-23.
11. Cesare Concordia, Stefan Gradmann, and

Sjoerd Siebinga(2010), “Not just another portal, not just another digital library: A portrait of Europeana as an application program interface,” *International Federation of Library Associations and Institutions*, 36(1), pp.61 - 69.

12. Christian Bizer, Tom Heath, and Tim Berners-Lee(2009), “Linked data-the story so far,” *International Journal on Semantic Web and Information Systems*, 5(3), pp. 1-22.
13. Constantia Kakali, Irene Lourdi, Thomais Stasinopoulou, Lina Bountouri, Christos Papatheodorou, Martin Doerr, and Manolis Gergatsoulis(2007), “Integrating Dublin Core Metadata for Cultural Heritage Collections Using Ontologies,” *Proceedings of the International Conference on Dublin Core and metadata Applications*, pp.128-139.
14. Denny Vrandečić, Markus Krötzsch(2014), “Wikidata: A Free Collaborative Knowledge Base,” *Communications of the ACM*, 57, pp.78-85.
15. Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek(2014), “Yago3: A knowledge base from multilingual wikipedias,” *Proceedings of 7th Biennial Conference on Innovative Data Systems Research*, pp.697-713.
16. Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli(2016), “Embeddings for Word Sense Disambiguation: An Evaluation Study,” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp.897 - 907.
17. J. Wojciechowski, B. Sakowicz, K. Dura, and A. Napieralski(2004), “MVC model, struts framework and file upload issues in web applications based on J2EE platform,” *Proceedings of the International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science*, pp.342-345.
18. Jeyavaishnavi Muralikumar, Sri Ananda Seelan, Narendranath Vijayakumar and Vidhya Balasubramanian(2017), “A statistical approach for modeling inter-document semantic relationships in digital libraries,” *Journal of Intelligent Information Systems*, 48(3), pp.1-22.
19. Jiaul H. Paik(2013), “A novel TF-IDF weighting scheme for effective ranking,” *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 343-352.
20. Joel P. Lucas, Nuno Luz, María N. Moreno, Ricardo Anacleto, Ana Almeida Figueiredo, Constantino Martins(2013), “A hybrid recommendation approach for a tourism system,” *Expert Systems with Applications*, 40(9), pp.3532-3550.
21. Karen Spärck Jones(1972), “A Statistical Interpretation of Term Specificity and Its Application in Retrieval,” *Journal of Documentation*, 28(1), pp.11 - 21.
22. Kevin Meehan, Tom Lunney, Kevin Curran, Aiden McCaughey(2013), “Context-Aware Intelligent Recommendation System for Tourism”, *IEEE International Conference on Pervasive Computing and Communications*

Workshops, pp.328-331.

23. Kewen Chen, Zuping Zhang, Jun Long and Hao Zhang(2016), "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Systems with Applications*, 66(C), pp.245-260.
24. Key-Sun Choi and Hee-Sook Bae(2014), "Korean-Chinese-Japanese Multilingual WordNet with Shared Semantic Hierarchy," *Proceedings of the International conference on language resource and evaluation*, pp.1131-1134.
25. Liyang Y(2011), "Linked open data," *A Developer's Guide to the Semantic Web*, Springer Berlin Heidelberg, pp.409-466.
26. Martin Doerr, Stefan Gradmann, Steffen Hennicke, Antoine Isaac, Carlo Meghini, and Herbert van de Sompel(2010), "The Europeana Data Model (EDM)" *World Library and Information Congress: 76th IFLA General Conference and Assembly*, pp.1-12.
27. Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, and Justin Ma(2012), "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association,* pp.2-2.
28. Michael Owens, and Grant Allen(2010), "SQLite," *Apress LP.*
29. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives(2007), "DBpedia: A nucleus for a web of open data," *Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, pp.722-735.
30. Yang Changhui and Meng Hongyan(2016), "Research on Constructing Application System of Exhibition Integrated Information Service Platform in Airport Economic Zone," *International Journal of u-and e-Service, Science and Technology*, 9(4), pp.165-174.
31. Yang Haiying(2010), "Exhibition Management Information System Design and Implementation," *Proceedings of the International Conference on Computer and Automation Engineering*, pp.633-636.
32. Zhenbin Yang and Atreyi Kankanhalli (2013), "Innovation in government services: The case of open data," *Proceedings of the International Working Conference on Transfer and Diffusion of IT*, pp.644-651.

Abstract

Big Data based Tourist Attractions Recommendation†

– Focus on Korean Tourism Organization Linked Open Data –

Ahn, Jinhyun* · Kim, Eung-Hee** · Kim, Hong-Gee***

Conventional exhibition management information systems recommend tourist attractions that are close to the place in which an exhibition is held. Some recommended attractions by the location-based recommendation could be meaningless when nothing is related to the exhibition's topic. Our goal is to recommend attractions that are related to the content presented in the exhibition, which can be coined as content-based recommendation. Even though human exhibition curators can do this, the quality is limited to their manual task and knowledge. We propose an automatic way of discovering attractions relevant to an exhibition of interests. Language resources are incorporated to discover attractions that are more meaningful. Because a typical single machine is unable to deal with such large-scale language resources efficiently, we implemented the algorithm on top of Apache Spark, which is a well-known distributed computing framework. As a user interface prototype, a web-based system is implemented that provides users with a list of relevant attractions when users are browsing exhibition information, available at http://bike.snu.ac.kr/WARP. We carried out a case study based on Korean Tourism Organization Linked Open Data with Korean Wikipedia as a language resource. Experimental results are demonstrated to show the efficiency and effectiveness of the proposed system. The effectiveness was evaluated against well-known exhibitions. It is expected that the proposed approach will contribute to the development of both exhibition and tourist industries by motivating exhibition visitors to become active tourists.

Key Words: Tourist Attraction Recommendation, Exhibition Management Information System, Linked Open Data, Large-Scale Knowledge Processing

† This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.2013-0-00109, WiseKB: Big data based self-evolving knowledge base and reasoning platform)

* Senior Researcher, Dental Research Institute, Seoul National University(First Author), jhahncs@snu.ac.kr

** Assistant Professor, BK21 Plus Dental Life Science, Seoul National University, eungheekim@snu.ac.kr

*** Professor, Healthcare Management and Informatics, Department of Dentistry, Seoul National University(Corresponding Author), hgkim@snu.ac.kr