Journal of Korean Institute of Intelligent Systems, Vol. 24, No. 5, October 2014, pp. 470-476
http://dx.doi.org/10.5391/JKIIS.2014.24.5.470

관심 지점 명칭의 단어와 문맥 정보를 활용한 관심 지점의 분류

Categorization of POIs Using Word and Context information

최수정 · 박성배†
Su Jeong Choi, and Seong-Bae Park†

경북대학교 컴퓨터학부
† School of Computer Science and Engineering, Kyungpook National University

요 약

관심 지점이란 상점이나 공원, 음식점 등과 같이 사람들이 관심을 가지거나 유용하다고 생각하는 특정한 지리적 위치를 의미한다. 관심 지점은 명칭과 제공 서비스, 카테고리 등과 같은 여러 정보들로 구성되어 있다. 이와 같은 정보들은 위치기반 어플리케이션에서 필수적인 정보이고, 그 중에서도 카테고리 정보는 위치기반 서비스에서 가장 핵심적인 역할을 한다. 그러나 관심 지점의 카테고리 정보를 직접 모으는 것은 많은 비용과 노력이 들기 때문에 자동으로 수집되어야 한다. 본 논문에서는 카테고리를 자동으로 추정하기 위해서 관심 지점 명칭의 단어 정보와 제한적 주변 문맥 정보를 결합하여 사용하는 방법을 제안한다. 관심 지점 명칭의 단어에는 카테고리를 반영하는 단어들을 포함하고 있어 카테고리를 추정하는데 있어서 중요한 단서가 된다. 제한적 주변 문맥 정보는 관심 지점의 명칭이 언급된 문서에서 명칭이 언급된 주변의 문맥을 의미한다. 명칭이 언급된 주변의 문맥에는 관심 지점의 카테고리를 추정할 정보들을 포함하고 있어 카테고리를 추정하는 것에 있어서 가치있는 정보를 제공한다. 우리는 제안한 모델의 성능을 측정하기 위해 두 가지 데이터셋에서 성능을 평가한 결과, 각 정보를 따로 사용하여 카테고리를 추정한 성능보다 결합하여 사용한 모델의 성능이 더 높게 나타났다.

키워드 : 관심 지점, 관심 지점 카테고리 추정, 분류, SVM.

Abstract

A point of interest is a specific point location such as a cafe, a gallery, a shop, or a park. It consists of a name, a category, a location, and so on. Its information is necessary for location-based application, above all category is basic information. However, category information should be automatically gathered because it costs high to gather it manually. In this paper, we propose a novel method to estimate category of POIs automatically using an inner word and local context. An inner word is a word that contains POI's name. Their name sometimes expose category information. Thus, their name is used as inner word information in estimating category of POIs. Local context information means words around a POI's name in a document that mentioned the name. The context include information to estimate category. The evaluation of the proposed method is performed on two data sets. According to the experimental results, proposed model using combination inner word and local context show higher accuracy than that of model using each.

Key Words : Point of interest, POI category estimation, Classification, SVM.

1. 서 론

관심 지점(Point of interest)이란 상점이나 공원, 도서관, 음식점과 같이 사람들이 관심을 가지거나 유용하다고 생각하는 특정한 지리적 위치를 의미한다. 관심 지점은 명칭과 제공하는 서비스에 따른 카테고리, 주요 서비스, 가격대 등의 정보들로 이루어져 있다. 예를 들어 '두레'는 서울특별시 종로구 인사동에 위치하고 있으며, 3만원 이상의 한정식을 판매하는 곳이다. 이런 관심 지점의 정보들은 위치기반 어플리케이션에서 주로 사용한다. 위치기반 어플리케이션들은 주로 사용자에게 관심 지점의 검색이나 추천 등의 서비스를 제공하기 때문에, 위치기반 어플리케이션에서 관심 지점에 대한 정보들은 필수적인 정보이다. 그 중에서도 관심 지점의 카테고리 정보는 사용자들이 관심 지점 검색을 할 때, 대개 카테고리를 기준으로 검색하기 때문에 위치기반 어플리케이션에서 핵심적인 역할을 한다. 예를 들어 사용자가

접수일자: 2014년 2월 11일
심사(수정)일자: 2014년 5월 14일
게재확정일자 : 2014년 5월 23일
† Corresponding author
이 논문은 2013(2014)학년도 경북대학교 학술연구비에 의하여 연구되었음.

점심 식사를 하기 위해서 관심 지점을 검색한다고 하면, 대부분의 사용자들은 음식점 카테고리를 먼저 선택한 다음에 세부 제공 메뉴를 보고 점심 식사를 할 관심 지점을 선택한다. 이러한 사용자의 검색 패턴을 반영하여 대부분의 위치기반 어플리케이션은 사용자가 관심 지점을 검색하기에 용이하도록 카테고리 정보를 메인화면에서 제공하거나 검색 조건에 적용하기 편리하도록 되어 있다.

관심 지점에 관한 카테고리 정보를 수집하기 위한 방법 중 하나는 조사자가 직접 관심 지점을 방문하여 관심 지점에 대한 정보를 수집하는 것이다. 이 방법을 사용하여 관심 지점에 대한 정보들을 수집하면, 좋은 질의 정보들을 얻을 수 있다는 장점이 있다. 하지만 관심 지점들은 넓은 지역에 분포하고 있어, 많은 양의 정보를 수집하기 위해서는 많은 비용이 발생한다. 또한 음식점과 같은 일부 관심 지점들은 금방 생겨나고 없어지기 때문에, 양질의 정보를 위해서는 정보의 잦은 업데이트를 필요로 한다. 이런 이유들로 인해서 관심 지점에 대한 카테고리를 정보들은 자동으로 수집되어야 한다.

우리는 때때로 관심 지점의 명칭만으로 관심 지점의 카테고리를 추정할 수 있다. 예를 들어 '공주 떡볶이'와 같은 관심 지점의 명칭을 보면, '떡볶이'만으로 우리는 이 관심 지점이 떡볶이를 판매하는 음식점이라는 사실을 쉽게 알 수 있다. '갤러리 라메르'와 같은 관심 지점도 마찬가지로 '갤러리'를 통해서 예술 전시와 관련된 관심 지점이라는 사실을 쉽게 추정할 수 있다. 이런 예들을 보면 관심 지점의 명칭에는 그 관심 지점의 카테고리를 추정할 수 있는 정보가 내포되어 있음을 알 수 있다. 하지만, 모든 관심 지점의 명칭들이 카테고리를 추정할 수 있는 정보들을 내포하고 있는 것은 아니다. 관심 지점 '안다미로'는 명칭만으로 어느 카테고리인지 추정하기 쉽지 않다. 이런 관심 지점의 카테고리를 추정하기 위해서는 그 관심 지점에 대한 더 많은 정보가 필요하다.

> 국내 정상급 쉐프가 직접 구워내는 이탈리아 가정식 빵으로 시작하는 안다미로의 음식은 최상의 재료와 신선하고 다양한 재료의 맛을 세심히 살려 조화를 아우러내는 이태리 중부 토스카나 음식을 고수합니다. 최고의 맛과 최상의 서비스 아늑한 분위기로 고객 여러분께 즐거움을 선사해 드리겠습니다. 좋은 사람과 소중한 추억을 함께 하세요.

그림 1. 관심 지점 '안다미로'를 서술하는 문서
Fig. 1. A document to describe a POI 'Andamiro'

그림 1은 관심 지점 '안다미로'에 대해서 서술하고 있는 문서의 일부분이다. 관심 지점의 명칭인 '안다미로'만으로 쉽게 카테고리를 추정할 수 없지만, 그림 1의 '쉐프'나 '음식' 등의 단어로 '안다미로'가 음식점이라는 사실을 알 수 있다. 웹에는 그림 1과 같이 관심 지점을 서술하는 많은 문서들이 존재한다. 이런 문서는 대개 리뷰나 기사, 광고들이며, 관심 지점의 위치나 주요 서비스, 서비스에 대한 만족도와 같은 정보들이 포함되어 있다. 그렇기 때문에 우리는 앞서 언급한 명칭의 단어 정보와 문서의 주변 문맥 정보를 활용하면, 보다 적은 비용과 노력으로 관심 지점의 카테고리를 추정할 수 있다.

본 논문에서는 관심 지점에 대한 명칭의 단어 정보와 문서의 주변 문맥 정보를 사용하여 자동으로 관심 지점의 카테고리를 추정하는 방법을 제안한다. 관심 지점 명칭의 단어 정보는 관심 지점의 명칭에 쓰인 단어들을 사용한다. 주변 문맥 정보는 관심 지점의 명칭을 언급하고 있는 문서를 그 관심 지점에 대해서 서술하는 문서로 간주하고, 그 문서의 문맥 정보를 사용한다. 따라서 우리는 관심 지점의 명칭이 주어졌을 때, 이 두 가지 정보를 각각 고려하여 그 관심 지점의 카테고리를 추정한다.

먼저 관심 지점의 명칭이 주어지면, 그 명칭의 단어 정보를 벡터(Vector)로 표현한다. 또한 그 관심 지점을 언급하고 있는 문서들도 벡터로 표현한다. 관심 지점을 추정하기 위한 정보들을 벡터로 표현하고 나면, 관심 지점의 카테고리를 추정하는 문제는 다중 클래스 분류(Multiclass classification) 문제로 고려될 수 있다. 우리는 이 분류 문제를 관심 지점 명칭의 단어 정보를 사용한 모델과 주변 문맥 정보를 사용한 모델로 나누어 Support Vector Machine(SVM)을 사용하여 해결한다. 각 모델은 카테고리에 대해서 신뢰도 점수를 계산하고, 각 모델을 통해 나온 신뢰도 점수를 결합하여 최종적으로 카테고리를 추정한다. 실험에서는 본 논문에서 제안한 방법의 성능을 두 가지 데이터셋에서 평가한다. 실험 결과, 제안한 방법이 높은 성능을 보여주어 우수함을 증명하였다.

본 논문은 2장에서는 관심 지점에 대해서 서술하고 있는 웹의 문서들을 다루는 연구들을 살펴보고, 3장에서는 제안하는 방법인 카테고리를 추정하는 문제에 대해서 해결하는 방법을 제시하고 설명한다. 4장에서는 실험을 통해서 제안한 방법에 대한 성능을 평가하고 비교 실험을 보임으로서 우리가 제안한 방법이 효과적임을 증명한다. 마지막으로 5장에서는 결론 및 향후 계획으로 맺는다.

2. 관련 연구

GPS가 탑재된 스마트폰의 대중화로 인해 위치기반 서비스가 활발히 사용되면서 관심 지점에 대한 정보의 중요성이 부각되고 있다. 그에 따라 웹상에 존재하는 문서에 언급된 관심 지점을 다루는 연구들이 활발히 이루어 졌다. 하지만 본 논문에서와 같이 관심 지점의 정보 중 카테고리를 추정하는 방법에 대한 연구는 이루어지지 않았다. 본 장에서는 웹 문서에서 언급되는 관심 지점을 다루는 연구들을 살펴본다.

Rae et al.[1]는 사람의 개입 없이 문장에서 자동으로 관심 지점의 명칭이 나타난 위치를 찾는 문제를 해결하였다. 문장에서 나타난 관심 지점의 위치는 개체명 인식(Named Entity Recognition)문제[2, 3, 11]에서 자주 쓰이는 Conditional Random Fields (CRFs)[4]를 사용하여 찾았다. 실험에서는 Yahoo! Place Maker와 비교하여 평가하였다. Wikipedia 기사 데이터를 이용해 비교 실험한 결과, Rae et al.이 제안한 모델이 Yahoo! Place Maker보다 50% 높은 정확률과 재현률을 보여주었다. Yahoo! Place Maker는 지명 사전을 이용해 문장에서 나타난 관심 지점의 명칭과 비교하여 찾기 때문에, 지명 사전에 존재하지 않는 관심 지점의 명칭은 찾을 수 없어 제안한 모델보다 낮은 성능을 보여주었다. 하지만 Wikipedia 기

사 데이터는 크기가 작아 지리학적인 범위가 좁고 문장에서 나타나는 관심 지점의 종류가 한정되어 있다. 이러한 한계점을 극복하기 위해서 Rae et al.은 데이터를 Foursquare와 Gowalla를 이용하여 부트스트랩(Bootstrap)하여 실험을 하였다. 그 결과, 제안한 모델은 State-of-the-art보다 정확률은 52%, 재현률은 187% 높은 성능을 보여 효과적임을 증명하였다.

Web-a-Where[5]은 웹 문서에서 언급하고 있는 관심 지점에 대한 지리적 위치 정보를 태깅한 후, 그 웹 문서에서 중점이 되는 관심 지점을 추론하였다. 이 시스템은 3단계로 나누어져 수행한다. 1단계는 웹 문서에 나타난 모든 관심 지점들을 전 세계의 지명을 계층 구조로 가지고 있는 지명 사전을 이용해서 찾는다. 2단계는 찾은 관심 지점들에 대해서 계층적인 지리적 위치 정보를 지명 사전에서 찾아 태깅한다. 지명 사전에서 해당하는 관심 지점을 찾을 때, 관심 지점의 명칭에 대해서 geo/non-geo와 geo/geo 두 가지 종류의 모호한 문제가 나타난다. geo/non-geo는 관심 지점의 명칭에 지리학적이지 않고 생활에서 다른 의미로 사용되는 단어들이 관심 지점의 명칭에 나타나는 경우이고, geo/geo는 지리학적으로 서로 다른 위치이지만 명칭이 같은 경우에 발생하는 것이다. Web-a-Where은 이런 문제를 몇 가지 규칙을 만들어서 해결해 지리적 위치 정보를 태깅한다. 마지막 3단계에서는 알고리즘을 제안하여 그 웹 문서에서 중점이 되는 관심 지점을 추론한다. 제안한 알고리즘은 그 웹 문서에서 언급되지 않은 관심 지점도 추론해낼 수 있어 효과적이다. 실험 결과, 이 시스템은 지리적 위치 정보를 태깅하는 실험에서 81.7%의 정확률을 보여 주어 제안한 방법이 효과적임을 보여주었지만, 중점이 되는 관심 지점을 추론하는 실험에서는 38%의 정확률로 그다지 효과적이지 못 하였다. 그래서 Zong et al.[6]는 중점이 되는 관심 지점을 찾는 새로운 방법을 제안하였다. 제안한 방법은 Web-a-Where 시스템과 유사한 문제를 해결하였지만, 중점이 되는 관심 지점을 찾는 단위가 다르다. Web-a-Where은 웹 문서를 단위로 찾는 반면에, Zong et al.이 제안한 방법은 웹 문서와 세그먼트(Segment) 단위로 중점이 되는 관심 지점을 찾는다. 여기서 웹 페이지에서 HTML태그를 기준으로 노드들을 생성해 하위 노드 200개를 가지는 노드들을 세그먼트라 하였다. 중점이 되는 관심 지점을 찾기 위해서 몇 가지 규칙을 정해 그 규칙을 만족하는 것들에 대해서 차등 점수를 주어 추론하는 방법을 제안하였다. 웹 문서를 대상으로 중점이 되는 관심 지점을 추론한 실험에서는 Web-a-Where보다 48% 높은 정확률을 보여 주었고, 세그먼트를 대상으로 한 실험은 정확률 65.2%의 성능을 보여 제안한 방법이 Web-a-Where 보다 우수함을 증명하였다.

Wang et al.[7]은 기존의 확률적 그래픽컬 모델인 Latent Dirichlet Allocation(LDA)[8]를 확장해 단어와 장소간의 관계를 학습하는 토픽 모델 LATM(Location Aware Topic Model)을 제안하였다. LATM 모델은 모든 단어들은 많던 적던 간에 어떤 특정한 장소와 관계가 있다는 가정 하에, 이런 관계를 학습해 잠재적인 토픽들을 찾는다. 실험에서는 블로그나 뉴스기사 같은 문서의 각 단어들이 관계되어 있는 특정한 장소로 태깅된 데이터를 사용하였다. 그 결과 제안한 모델이 장소와 단어들 간의 관계를 잘 찾아내어 모델이 효과적임을 증명하였으나, 사용한 데이터 셋이 대부분 영어로만 되어 있어 많은 단어들이 미국과의 관계를 나타내는 문제점을 보여주었다.

Alves et al.[9]은 KUSCO 시스템을 개발하였다. KUSCO는 관심 지점에 대한 의미들을 찾아 색인으로 제공하는 시스템이다. 관심 지점이 주어 졌을 때, KUSCO는 그 관심 지점과 관련된 웹 페이지들을 검색한다. 그리고 그 관심 지점에 대한 의미들의 색인으로 만들기 위해서, 검색된 웹 페이지들에 대해서 정보를 추출한다. 관심 지점에 대한 의미의 색인은 일반적 의미와 특정한 의미로 이루어져 있다. 일반적 의미는 WordNet과 같이 단어에 대한 사전적 의미를 가진 온톨로지(Ontology)에서 찾게 되고, 특정한 의미는 보통 고유 명사들이다. Alves et al.은 그들의 시스템이 다양하고 풍부한 의미들을 제공한다는 것을 설명하기 위해, KUSCO를 Yahoo!Term Extraction API(Yahoo!TE)와 비교 하였다. 그 결과, KUSCO는 Yahoo!TE보다 약간 좋은 성능을 보여주었다.

3. 관심 지점의 분류

우리의 목적은 관심 지점의 명칭이 주어졌을 때, 그 관심 지점에 관한 두 가지 정보를 사용하여 카테고리를 추정하는 것이다. 두 가지 정보는 관심 지점의 명칭의 단어와 문맥 정보이고, 우리는 이 정보들을 Bag-of-words(BOW)로 표현하여 사용한다. 추정하기 위하여 사용할 정보들을 BOW로 나타내고 나면, 결국 카테고리를 추정하기 위한 문제는 다중 클래스 분류(Multiclass classification) 문제로 간주하여 해결할 수 있다.

먼저 두 가지 정보 중에서 관심 지점 명칭의 단어에 대하여 살펴본다. ‘Beauty Mark Salon’이라는 관심 지점의 명칭을 보면, 우리는 ‘Salon’라는 단어를 통해서 이 관심 지점이 미용 관련 카테고리에 속한다는 사실을 쉽게 알 수 있다. ‘Belmont Hotel’이나 ‘Bloor Animal Hospital’, ‘Leung′s Driving School’ 등도 마찬가지이다. 위의 예를 통해서 본 것과 같이, 관심 지점의 명칭에 나타난 특정한 단어들을 통해서 우리는 카테고리를 추정할 수 있다. 즉, 관심 지점의 명칭에 나타난 특정한 단어들은 관심 지점의 카테고리를 반영하고 있고 우리는 이런 특성을 이용하여 관심 지점의 카테고리를 추정할 수 있다.

$$nf = \langle p^1, p^2, p^3, .., p^k \rangle \quad (1)$$

본 논문에서는 관심 지점 명칭의 단어 정보를 사용해 카테고리를 추정하기 위해서, 이 정보를 위의 식 (1)과 같이 BOW로 나타내어 사용한다. BOW로 표현된 관심 지점 명칭의 단어 정보 nf는 길이가 k인 관심 지점 명칭의 단어 벡터이고, p^k는 명칭에서 k번째 단어의 TF·IDF값을 의미한다. 하지만 모든 관심 지점 명칭들이 카테고리를 추정할 수 있는 의미를 반영하는 특정한 단어들을 포함하고 있는 것은 아니다. 다음 표 1은 관심 지점의 명칭만으로 카테고리를 추정할 수 있는 관심 지점과 그렇지 않은 관심 지점의 명칭과 카테고리를 나열해 놓은 예이다.

표 1. 관심 지점의 명칭만으로 카테고리를 추정할 수 있는 예와 그렇지 않은 예
Table 1. Examples that POIs′ name expose category or not

	POIs′ name	Category
(a)	Austin Water Bikes	Active
	B Square Salon	Beauty
	Floss Dental	Health
	Mgnolia Hotel Dallas	Hotel
	Bubbly Paws Dog Wash	Pet
	Atlas Properties	Realestate
	CA Jewelers	Shop
(b)	Atomic Allure	Active
	Blink	Beauty
	All Out Effort	Health
	Bustonian	Hotel
	Anytime K9	Pet
	Anthony Falvo	Realestate
	Dandelion	Shop

표 1에서 (a)는 명칭만으로 카테고리를 추정할 수 있는 관심 지점들을 나열해 놓은 것으로 'Austine Water Bikes'의 'Bikes'를 보면 'Active'와 관련된 카테고리에 속하는 것을 알 수 있다. 나머지 예들도 'Salon'이나 'Hotel', 'Dog', 'Properties' 등을 통해서 카테고리를 유추할 수 있다. 표 1의 (b)에서는 (a)와는 달리 명칭만으로 카테고리를 추정할 수 없는 관심 지점들이 나열되어 있다. 관심 지점 'Atomic Allure'는 'Atomic'이나 'Allure'를 통해서 카테고리를 추정할 수 없다. 또한 'Blink'도 마찬가지이다. 이런 관심 지점들은 대부분 명칭에 카테고리를 반영하는 단어들을 포함하지 않고, 개인의 이름이나 관용어구 등을 포함하고 있다. 이와 같이 명칭만으로 카테고리를 추정하는 것에는 한계가 있다. 이러한 한계점을 극복하기 위해서 우리는 주변 문맥 정보를 결합하여 사용한다.

카테고리를 추정하기 위하여 사용할 정보 중 하나인 주변 문맥 정보는 관심 지점의 명칭을 언급하는 문서의 주변 문맥 정보이다. 웹에는 기사나 리뷰, 광고 등과 같은 관심 지점을 언급하며 관심 지점에 대해 서술하는 많은 문서들이 있다. 이러한 문서들은 관심 지점의 위치나 주요 서비스나 위치 정보, 서비스에 대한 만족도 등과 같은 정보들을 포함하고 있다. 예를 보면, 다음 그림 2는 관심 지점 'Gary Danko'에 대해서 서술하고 있는 문서이다. 이 문서를 통해서 우리는 'Gary Danko'가 샌 프란시스코에서 운영되고 있는 레스토랑이라는 것과 칵테일과 와인을 제공한다는 것을 알 수 있다. 'Gary Danko'의 명칭만으로 카테고리를 추정할 수 없지만, 이러한 문맥 정보들은 음식 관련 카테고리인 것을 추정할 수 있게 한다.

그림 2를 다시 살펴보면, 문서 전체가 'Gary Danko'에 대한 정보를 가지고 있는 것은 아니다. 첫 번째 문단에서는 'Gary Danko'가 샌 프란시스코에서 운영되는 레스토랑이고 칵테일과 와인을 제공한다는 사실을 알 수 있지만, 두 번째 문단에서는 'Gary Danko'의 카테고리를 추정할 수 있는 정보들을 포함하고 있지 않다. 이 두 문단은 'Gary Danko'라는 관심 지점의 명칭이 언급되어 있느냐 없느냐의 차이점을 가지고 있다. 'Gary Danko'의 카테고리를 추정할 수 있는 정보들은 관심 지점의 명칭이 언급된 주변 문맥에서 찾아볼 수 있다. 즉, 관심 지점을 서술하고 있는 문서에서 카테고리를 추정하는 것에 도움이 되는 정보들은 명칭이 언급된 주변 문맥에 주로 위치하고 있다. 우리는 이러한 사실을 이용하여 관심 지점의 카테고리를 추정한다.

Gary Danko is like Slanted Door, or Zuni Cafe, they don't need another review! What not to love about Gary Danko, the most staple fine dining destination in San Francisco. The good food is always consistent, the cocktails is awesome, the wine list is perfect, the service is excellent, and it's always happening here.

I don't care for the tightly tables, next to each other, it can get loud and you won't hear your date talking. Well, lucky of you can get a RSVP weekend at 7 PM, it books up a month. However, it's worth it. I usually take visitors here if their first few times in SF.

그림 2. 관심 지점 'Gary Danko'를 서술하는 문서
Fig. 2. A document to describe a POI, 'Gary Danko'

본 논문에서는 문서에서 관심 지점의 명칭이 언급된 주변 문맥만을 반영하기 위해 제한적 주변 문맥 정보를 사용하여 관심 지점의 카테고리를 추정한다. 우리는 제한적 주변 문맥 정보를 사용하기 위해서 제한적 주변 문맥 정보를 다음 식 (2)와 같이 표현한다.

$$cf = \{w^j | (a-l) \le j \le (a+l)\} \quad (2)$$

윈도우 크기 l은 관심 지점의 명칭이 언급된 주변의 문맥들만 사용하기 위한 변수이다. a가 문서에서 관심 지점 명칭이 언급된 위치라고 할 때 제한적인 주변 문맥 정보를 사용하기 위하여 윈도우 크기 l 안에 존재하는 단어들만 사용한다. 즉, 관심 지점에 대한 제한적 주변 문맥 정보 cf는 명칭이 언급된 위치 a를 중심으로 윈도우 크기 l 안에 존재하는 단어들의 BOW 벡터로 표현되고, 각 벡터의 값은 TF·IDF를 사용한다. 예를 들어 그림 2의 문서를 직접 벡터로 표현해보면, 다음 그림 3과 같이 표현될 수 있다.

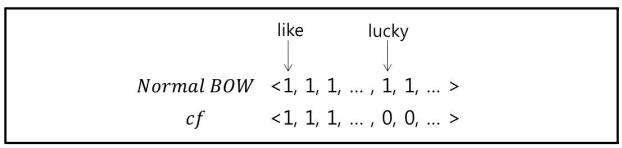

그림 3. 문서의 주변 문맥 벡터 표현
Fig. 3. An example of document representation using local context

그림 3의 $Normal\,BOW$는 일반적인 문서의 주변 문맥을 벡터로 표현한 것이고, cf는 본 논문에서 사용하는 문서의 제한적인 주변 문맥들만 벡터로 표현한 것이다. 일반적인 문서의 주변 문맥 표현은 문서의 전체를 모두 벡터화하지만, 본 논문에서는 관심 지점의 명칭이 언급된 제한적인 주

변 문맥에서만 관심 지점의 카테고리를 추정할 정보가 있다고 가정하고 윈도우 크기 l만큼만 벡터화 한다.

앞서 이야기한 명칭의 단어 정보 nf와 제한적 주변 문맥 정보 cf를 BOW로 표현하고 나면, 관심 지점의 카테고리를 추정하는 문제는 다중 클래스 문제로 고려할 수 있다. 기계학습적 관점에서 보면, 관심 지점의 카테고리를 분류하는 문제는 $f: X \rightarrow Y$ 함수를 추정하는 것이다. 관심 지점 데이터 D가 $D=\{(\overline{x_1}, y_1), (\overline{x_2}, y_2) ..., (\overline{x_n}, y_n)\}$와 같이 주어지면, $\overline{x} \in X$는 관심 지점의 카테고리를 추정하기 위해 사용할 정보를 벡터로 나타낸 것으로 $\overline{x_i}$는 i번째 관심 지점의 카테고리를 추정하기 위한 정보의 벡터로서, i번째 관심 지점 명칭의 단어 정보 nf_i와 제한적 주변 문맥 정보 cf_i로 구성되어 $\overline{x_i} = \langle nf_i, cf_i \rangle$로 정의된다. 우리가 하고자 하는 일은 i번째 관심 지점에 대한 $\overline{x_i}$가 주어졌을 때, 카테고리 y_i를 추정하는 것이다. 따라서 카테고리를 추정하는 함수 $f: X \rightarrow Y$ 는 다음 식 (3)의 $S(\overline{x_i}, c)$함수로 정의 된다. 식 (3)의 $S(\overline{x_i}, c)$함수는 $\overline{x_i}$가 주어졌을 때, 카테고리 c에 대한 신뢰도 점수를 계산한다.

$$S(\overline{x_i}, c) = \alpha I(nf_i, c) + (1-\alpha) O(cf_i, c) \quad (3)$$

식 (3)에서 함수 $I(nf_i, c)$는 관심 지점 명칭의 단어 정보 nf_i가 주어졌을 때, 카테고리 c에 대한 신뢰도 점수를 계산하는 함수이다. 함수 $O(cf_i, c)$는 관심 지점을 언급하는 문서의 제한적 주변 문맥 정보 cf_i가 주어졌을 때, 카테고리 c에 대한 신뢰도 점수를 계산한다. 가중치 α는 0과 1사이의 값으로 명칭의 단어 정보와 제한적 주변 문맥 정보를 반영하는 비율이다. 두 함수 $I(nf_i, c)$와 $O(cf_i, c)$는 Crammer et al.[10]이 제안한 다중 클래스 SVM(Multiclass SVM)을 통해 학습하여 나온 파라미터를 사용하여 신뢰도 점수를 계산한다.

먼저 다중 클래스 SVM은 여러 개의 클래스를 고려하기 위해 제안된 것으로, 클래스들 사이의 마진을 최대로 하는 가장 최적화된 초평면을 찾아 파라미터를 학습한다. 다음 식 (5)를 조건으로 하여 식 (4)를 통해 학습 데이터 D를 사용하여 클래스별 파라미터 w를 학습한다.

$$\min_{w,\xi} \frac{1}{2}\|w_k\|^2 + C\sum_{i=0}^{n}\xi_i \quad (4)$$

$$(w_{y_i} \bullet \phi(x_i, y_i)) - (w_k \bullet \phi(x_i, k)) \geq 1 - \xi_i, \quad \xi_i \geq 0, \forall i, \forall k \in K \setminus y_i \quad (5)$$

w는 학습데이터 D로부터 추정된 파라미터이며 ξ는 힌지 로스(Hinge Loss)이다. C는 사용자 파라미터이며, $\phi(x_i, y_i)$는 x_i와 y_i의 자질 표현이고, K는 클래스들의 집합을 의미한다. 식 (4)와 (5)를 통해서 파라미터 w를 학습하고 나면, 다음 식 (6), (7)을 사용하여 i번째 관심 지점의 카테고리별 신뢰도 점수를 계산한다.

$$I(nf_i, c) = w_c \bullet nf_i \quad (6)$$

$$O(cf_i, c) = w_c \bullet cf_i \quad (7)$$

식 (6)은 관심 지점 명칭의 단어 정보 nf_i가 주어졌을 때, 카테고리 c에 대한 신뢰도 점수를 구하는 식이다. 마찬가지로 식 (7)은 관심 지점의 제한적 주변 문맥 정보 cf_i가 주어졌을 때, 카테고리 c에 대한 신뢰도 점수를 계산한다. 식 (6)과 (7)은 다중 클래스 SVM을 통해 학습된 카테고리의 파라미터 w_c를 i번째 관심 지점 명칭의 단어 정보 nf_i와 cf_i를 내적하여 나온 각 값을 카테고리 c에 대한 신뢰도 점수로 취한다. 위의 두 함수가 계산한 신뢰도 점수를 선형 가중치 α를 적용하여 합한 것이 최종 신뢰도 점수가 된다.

$$c^* = \arg max_{c \in Y} S(\overline{x_i}, c) \quad (8)$$

마지막으로 식 (8)와 같이 최종 신뢰도 점수가 가장 높은 카테고리 c가 i번째 관심 지점의 추정된 카테고리가 된다.

4. 실험 및 평가

본 논문에서는 제안한 방법을 Yelp 데이터셋과 Web 데이터셋에서 성능을 측정하였다. Yelp 데이터셋은 리뷰 사이트인 Yelp(www.yelp.com)에서 Yelp API를 사용하여 모은 리뷰 데이터이고 Web 데이터셋은 웹에서 관심 지점을 언급하고 있는 웹 문서들을 모은 데이터이다. 다음 표 2는 두 데이터셋의 카테고리와 카테고리당 관심 지점의 개수를 보여주고 있다.

표 2. Yelp와 Web 데이터셋의 통계
Table 2. The statistics of Yelp and Web datasets

Category	Yelp	Web	Category	Yelp	Web
Active	272	20	Local services	269	20
Arts	272	20	Mass Media	134	20
Auto	263	20	Nightlife	291	20
Beauty & Spa	258	20	Pets	276	20
Education	217	20	Professional	202	20
Event services	252	20	Public services	207	20
Financial services	135	20	Real estate	182	20
Health	190	20	Religious orgs	118	20
Home services	273	20	Restaurants	281	20
Hotels & travel	237	20	Shopping	284	20

두 데이터셋 모두 20개의 카테고리로 이루어져 있으며, 관심 지점의 총 개수는 Yelp 데이터셋은 4,613개이고 Web 데이터셋은 400개 이다. 두 데이터셋에 대해서 관심 지점을 언급하고 있는 문서의 관심 지점의 명칭이 나타난 부분을 'POI'로 대체하고, 모든 문서들에 대해서 불용어(Stop word)를 제거한 후 사용하였다. 분류기로는 Multiclass

SVM를 사용하였고 커널은 선형커널을 선택하였다. 사용자 파라미터 C는 5.0으로 설정하였다.

매개 변수 l, α는 각 데이터셋에서 검증 데이터셋을 통하여 추정된 값을 사용하였다. 윈도우 크기 l은 Yelp 데이터셋과 Web 데이터셋 모두 2를 사용하였다. 윈도우 크기 l이 문장 단위로 값이 2이라는 것은 관심 지점이 언급된 한 문장을 기준으로 앞의 2문장과 뒤의 2문장, 그리고 언급된 문장까지 포함하여 총 5문장을 제한적 주변 문맥으로 사용한다는 것을 의미한다. 매개 변수 α는 명칭의 단어 정보와 제한적 주변 문맥 정보를 결합하기 위해 사용되는 변수로서, 마찬가지로 검증 데이터셋을 통해 추정된 값을 사용하였다. Yelp에서는 α값이 0.1을 Web 데이터셋에서는 0.4를 사용하였다.

표 3. 카테고리 추정 실험 결과
Table 3. Results of categories of POIs estimation

Accuracy	Yelp dataset	Web dataset
(1) Inner word information	45.84%	51.33%
(2) Local context information	70.13%	66.00%
(3) The proposed method	**70.67%**	**68.33%**

위의 표 3은 관심 지점의 카테고리를 추정한 실험 결과를 보여 준다. 표3의 (1)은 관심 지점 명칭의 단어 정보를 사용하여 카테고리를 추정한 결과로 Yelp 데이터셋에서 45.84%, Web 데이터셋에서는 51.33%의 정확률을 보여주었다. 카테고리를 올바르게 추정할 수 없는 경우는 'Hana Hou'같이 명칭의 단어에 카테고리를 추정할 단어를 가지지 않은 경우였다. 하지만 이런 경우는 사람 또한 카테고리를 추정할 수 없다. 표 3의 (2)는 제한적 주변 문맥 정보를 사용하여 관심 지점의 카테고리를 추정한 결과이다. Yelp 데이터셋에서는 70.13%를 Web 데이터셋에서는 66%를 보여주었다. Yelp 데이터셋은 리뷰 데이터이기 때문에 관심 지점에 대해서 서술하고 있어 Web 데이터셋보다는 조금 높은 성능을 보여주었다. 반면에 Web 데이터셋은 기사나 광고 같은 웹 문서들이기 때문에 Yelp 데이터보다는 상대적으로 성능이 낮지만 제한적 주변 문맥을 사용함으로서 나쁘지 않은 성능을 보여주었다. 주로 올바른 추정을 하지 못한 관심 지점들은 그 관심 지점을 언급하고 있는 문서의 길이가 짧은 경우이다. 문서의 길이가 짧은 것들은 대개 관심 지점에 대해서 단순한 만족감만 표현하고 있어 카테고리를 추정할 정보들이 부족하여 올바르게 추정하지 못하였다.

표 3의 (3)은 본 논문에서 제안한 방법으로 위의 두 가지 정보를 결합하여 관심 지점의 카테고리를 추정한 결과이다. Yelp 데이터셋과 Web 데이터셋은 각각 70.67%, 68.33%의 정확률을 보여주었다. 이는 명칭의 단어 정보만을 사용한 모델과 제한적 주변 문맥 정보를 사용한 모델을 비교해보았을 때, 제안한 결합 모델이 더 좋은 성능을 보여주었다. 앞서 예를 든 'Hana Hou'는 이름만으로 카테고리를 추정할 수 없었지만, 제한적 주변 문맥 정보를 사용하여 올바르게 추정할 수 있어, 제안한 결합 모델은 두 정보를 결합함으로서 성능을 올릴 수 있었다.

5. 결론 및 향후 연구

본 논문에서는 관심 지점의 명칭이 주어졌을 때, 명칭의 단어 정보와 관심 지점의 명칭이 언급된 문서의 제한적 주변 문맥 정보를 사용하여 자동으로 관심 지점의 카테고리를 추정하는 방법을 제안하였다. 관심 지점의 명칭에는 그 관심 지점의 카테고리를 반영하는 단어들을 가지고 있어, 이것을 사용하여 관심 지점의 카테고리를 추정하였다. 하지만 그렇지 않은 관심 지점의 명칭도 존재하기 때문에 명칭의 단어 정보만으로 카테고리를 추정하는 것에는 한계가 있었다. 우리는 이 한계점을 극복하기 위해 제한적 주변 문맥 정보와 결합하여 최종적으로 카테고리를 추정하였다. 제안한 결합 모델을 두 가지 종류의 데이터셋에 적용해본 결과, 정보를 각각 사용한 모델보다 제안한 모델이 더 높은 성능을 보여주어 제안한 모델이 더 우수함을 증명하였다.

본 논문에서는 윈도우 크기의 매개 변수를 추정할 때에 검증 데이터셋을 통해서 추정하여 사용하였지만, 윈도우 크기는 문서의 속성이나 카테고리에 따라서 달라질 수 있을 것이다. 따라서 문서에서 관심 지점을 서술하는 범주를 문서의 속성이나 카테고리에 따라 자동으로 추정할 수 있게 된다면 모델의 성능을 조금 더 향상 시킬 수 있을 것이다.

References

[1] A. Rae, V. Murdock, A. Popescu, and H. Bouchard, "Mining the Web for Points of Interest," In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.711-702, 2012.

[2] A. McCallum and W. Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons," In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL*, Vol. 4, pp. 188-191, 2003.

[3] M. Collins, "Ranking Algorithms Named-Entity Extraction: Boosting and the Voted Perceptron," In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*, pp. 489-496, 2002.

[4] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," In *Proceedings of the 18th International Conference on Machine Learning*, pp. 282-289, 2001.

[5] E. Amitay, N. Har'El, R. Sivan, and A.Soffer, "Web-a-Where: Geotagging Web Content," In *Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 273-280. 2004.

[6] W. Zong, D. Wu, A. Sun, E. Lim, and D. Goh, "On Assigning Place Names to Geography Related Web Pages," In *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries*, pp. 354-365, 2005.

[7] C. Wang, J. Wang, X. Xie, and W. Ma, "Ming Geographic Knowledge Using Location Aware Topic Model," In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, pp. 65-70, 2007.

[8] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, 2003.

[9] A. Alves, F. Pereira, A. Biderman, and C. Ratti, "Place Enrichment by Mining The Web," In *Proceedings of the European Conference on Ambient Intelligence*, pp. 66-77, 2009.

[10] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines," *The Journal of Machine Learning Research*, Vol. 2, pp. 265-293, 2002.

[11] S. Kang, "English-Korean Cross-lingual Link Discovery Using Link Probability and Named Entity Recognition," *Journal of The Korean Institute of Intelligent Systems*, pp. 191-195, 2013.

저 자 소 개

최수정(Su Jeong Choi)

2011년 : 안동대학교 컴퓨터멀티미디어학부 공학사
2014년 : 경북대학교 대학원 컴퓨터학부 석사
2014년～현재 : 경북대학교 대학원 컴퓨터학부 박사과정

관심분야 : Machine Learning, Natural Language Processing,
Phone : +82-53-940-8692
E-mail : sjchoi@sejong.knu.ac.kr

박성배(Seong-Bae Park)

1994년 : 한국과학기술원 컴퓨터학과 학사
1996년 : 서울대학교 컴퓨터공학과 석사
2002년 : 서울대학교 컴퓨터공학과 박사
2004년～현재 : 경북대학교 IT대학 컴퓨터학부 교수

관심분야 : Machine Learning, Natural Language Processing, Text Mining
Phone : +82-53-940-8692
E-mail : sbpark@sejong.knu.ac.kr