Sequential Information Bottleneck을 이용한 음식점 추천 웹사이트를 위한 메뉴 기반 클러스터링

윤두밈 ○ 도해용 김경중
세종대학교 컴퓨터공학과
krad@hanmir.com, ceo@myfolder.net, kimkj@sejong.ac.kr

Menu-based Clustering for Restaurant Recommendation Websites using Sequential Information Bottleneck

Du-Mim Yoon ○ Hae-Yong Do, Kyung-Joong Kim
Dept of Computer Engineering, Sejong University

요 약

최근에는 사회구조가 복잡해 짐에 따라 각 분야의 전문성이 두드러지고 있다. 이러한 흐름은 음식점에도 영향을 주어 기존의 형식에서 벗어난 자신들만의 독특하고 차별성 있는 요리 메뉴의 개발을 가속화시켰다. 그로 인해 한식, 중식, 양식으로 구분하는 전통적인 음식점 분류 방식이 한계를 보였고, 기존의 분류를 포함하면서도 새로이 등장하는 음식점들을 다룰 수 있는 방식이 필요하다. 본 논문에서는 최근 폭발적으로 늘어나는 음식점 추천 웹 사이트의 데이터를 토대로 자동적으로 음식점 분류를 수행하는 방법을 제안한다. 본 연구에서는 각 음식점이 갖고 있는 특징을 메뉴 정보를 통해 파악하려 하였다. 음식점 사이트에서 수집한 2 천개의 음식점, 6 만개의 메뉴 정보를 미리 정의된 필터로 정제한 후 Sequential Information Bottleneck Clustering 알고리즘을 적용하여 구분해 보았다. 실험결과 제안한 방법이 다른 Clustering 방법에 비해 높은 성능을 보였으며 음식점주가 수동적으로 음식점 분류를 입력하는 수고를 줄일 수 있는 가능성을 보였다.

1. 서론

음식점의 분류는 과거에는 단순히 중식, 양식, 한식 등으로 구분되어 큰 어려움이 없었는데에 반해 현대에 와서는 음식점이 늘어나고 서로간에 경쟁관계가 형성되면서, 경쟁우위를 확보하기 위해 차별성 전략을 사용하게 되고 그로인해 음식점들의 세분화/전문화가 가속화되어 분류의 경계가 불확실하게 되었다.

최근 폭발적으로 늘어나는 음식점 관련 정보는 외식산업에 새로운 가능성을 제시하고 있다. 인터넷을 통해 음식점에 관한 각종 정보가 제공되고 있고, 이를 활용하여 추천 서비스 및 정보 분석 등이 시행되고 있다. 예를 들어, Austin 등은 지역내의 Fast Food 음식점 위치와 학교의 위치 사이의 관계를 분석하여 대부분 학교와 매우 근접한 거리에 있다는 것을 밝혀내었다[1].

외식산업의 시장규모는 1982 년 7 조 4300 억원에서 1997 년 30 조, 2007 년 57 조원으로 급성장을 하고 있지만, 우리나라의 표준산업분류에서 외식산업은 메뉴를 기준으로 되어있지만 이 메뉴가 소비자의 욕구의 변화에 따라 다양화되고 퓨전화 되는등 빠르게 변화하여 기존 분류 체계로는 분류가 제대로 되지 않는 문제점이 생기고 있다는 연구들이 많이 나오고 있다[2].

본 논문은 폭발적으로 늘어나는 인터넷 상의 음식점 관련 데이터를 활용하여 새로운 음식점 개설 동향을 손쉽게 반영하는 자동화된 분류방법을 제안하고자 한다. 인터넷 상의 음식점 추천 사이트로부터 음식점들의 메뉴에 대한 정보를 수집하고 이를 토대로 군집화를 수행한다. 군집화 결과를 토대로 새롭게 개설되는 음식점이 손쉽게 분류되도록 한다.

음식점 사이트에서 총 2 천개의 음식점, 6 만개의 메뉴 정보를 얻어낸 뒤, 메뉴에서 핵심 단어만 뽑아내는 필터링 작업을 거쳐 각 음식점들의 특징들을 이끌어내었다. 그 후 Sequential Information Bottleneck (sIB) 군집화 기법을 적용하였다[3][4]. 본 방법은 문서와 단어 사이의 Co-Occurrence를 이용하여

군집화를 수행하는 방법이다. 기존 연구에서 sIB는 전통적인 문서분류 기법인 naïve Bayesian Classifier에 근접한 성능을 보인다고 보고되었다[4].

2. 음식점 분류 체계

한국의 경우 한식점, 중국음식점, 일본음식점, 서양음식점업, 기관구내식당업, 기타일반음식점업으로 나누며, 기존 분류에 들어가지 않는 것은 기타 항목을 두는 방식으로 기타 음식점업에는 피자,햄버거 및 치킨전문점, 분식 및 김밥전문점, 이동음식점업, 그 외 기타 음식점업으로 분류한다.

미국의 경우 NRA(National Restaurant Association)는 외식시장을 상업성 여부에 따라 분류하여 상업적 외식시장, 비상업적 외식시장, 군대(Military Restaurant Service)로 나누고 서비스 수준, 메뉴의 종류, 계약 등의 구분에 따라 세분류하고 있다[5].

일본의 경우는 우리와 유사하여 업종에 따라 식당,레스토랑(일반식당, 일본요리점, 서양요리점, 중화요리점, 기타식당 레스토랑)로 나누며 더불어 메뉴 중심의 소바/우동점, 초밥집, 찻집, 기타 일반 음식점으로 분류하고 있다.

이렇게 각 나라별로 외식업체를 분류하는 기준들은 다르나 외식업체의 빠른 변화를 제대로 수용하지 못하는 문제점들은 모두 가지고 있다. 이에 외식업의 분류 기준으로 서비스 형태와 수준등 메뉴외의 요소들을 반영한 새로운 분류체계를 제시하는 연구들이 되고 있으며[2] 업종별 분류방식에서 업태별 분류방식으로 조정되어야 하며, 업태 분류 기준으로는 메뉴(식사 또는 음료구분), 서비스 수준, 경영활동의 목적 등이 선택되었다[6].

그러나 분류 기준 중 서비스수준, 경영활동등은 객관적인 수치가 아닌 조사자의 주관적 판단이어서 분류 기준을 정하기 어렵고, 도입 후에도 지속적인 기준의 업데이트가 필요하여 도입이 쉽지 않으며 메뉴에 따른 분류는 대상 업체수와 메뉴의 숫자, 그리고 조사비용의 문제로 지속적인 업데이트가 어렵다는 것이 현재까지의 문제점이다.

다행히 최근 인터넷의 활성화와 스마트폰의 보급으로 외식업체에 대한 데이터베이스 구축이 빠른 속도로 이루어 지고 있다. 그동안 수집하기 어려웠던 음식점들의 객관적인 데이터 즉, 메뉴,가격,시설등이 데이터베이스화 되고 있기에 이를 분석한다면 기존의 분류방식의 문제점이었던 조사시간과 비용부분을 일정부분 해결해 낼 수 있을 것이다.

3. 데이터 전처리

음식점 추천 웹 사이트로부터 얻은 메뉴 데이터는 입력 자체가 수작업으로 이루어지고 있기 때문에 잘못된 표기, 오탈자, 사용자의 실수로 인한 무의미한 데이터 유입등에 의해 데이터의 질이 고르지 않기 때문에 무의미한 데이터를 제거하고 오탈자를 교정한다.

6 만개의 음식점 메뉴는 군집화에 이용하기에 너무 많기 때문에, 요리사의 도움을 받아 대표적인 메뉴들로 줄인다. 요리명 온톨로지의 매칭을 통해 검색을 하게 되는데, 여기에서 사용되는 요리명 온톨로지의 선발 기준은 외식 전공자에 의해 다음과 같이 정의했다.

- 한식 메뉴명은 '대표 한식 102 종 표준 영문표기안' (2008,농림수산식품부)을 기준으로 정리하였으며 양식과 일식, 중식은 해당 분야의 조리 전문가로 부터 조언을 구하여 정리하였다.

- 첫번째로 조리법, 명칭 등이 유사하나 소비자입장에서는 확연히 다른 메뉴는 별개로 구분하여, 갈비의 경우 돼지갈비와 소갈비의 차이가 크기에 두개로 구분하였다. 반대로 메뉴명이 세세하게 나뉘어져 있는 경우는 하나로 통일하였는데, 예를 들면 뇨끼, 링귀니, 라자니아, 퀘사딜라, 라비올리 등은 파스타로 통일하였다.

- 두번째로 메뉴명이 브랜드명인 경우 하나로 통일하였는데, 예를 들면 버드와이져, 카프리, 하이트, 카스 는 브랜드 명이기에 맥주로 통일하였으며 앙리페시까베르네 소비뇽, 고트두롬, 떼니멘띠안젤리니뚜또베네 등은 와인 브랜드 명이기에 와인으로 통일하였다.

- 세번째로 메뉴명으로 구분은 되어 있으나 유사한 것은 하나로 통일하였는데 예를 들면 식혜,콜라,사이다,에이드,주스,스프라이트,환타, 우유,소다 등은 음료로 녹차, 홍차, 허브티, 커피, 아이스티, 에스프레소, 아메리카노, 까페, 카푸치노, 캐러멜마끼야또, 티, 코코아, 라떼 등은 차로 코러마, 하라바라, 마카니 등은 커리로 통일하였다.

- 마지막으로 독립적인 메뉴명으로 보기 어려운 보통명사들은 연구자가 제거하였다. 구이, 국, 떡, 면, 빵, 공기밥, 전, 탕, 사리, 볶음 등의 경우 독립적인 메뉴명이 아닌 포괄적인 메뉴명을 의미하기에 제거하였으며, 주안상, 세트, 모듬, 반상, 안주, 요리, 정식, 코스 등의 경우 개별 메뉴가 아닌 세트 메뉴 형태의 종합적인 의미의 단어이기에 제거하였으며 닭, 두부, 문어, 라이스, 치즈, 해산물의 경우 메뉴명이 아닌 재료에 해당되는 단어이기에 제거하였다.

먼저 메뉴내에 등장하는 모든 공백을 제거하여 단순명료한 매칭이 가능하게 한다. 다음으로 매칭을 시킬 때 메뉴명에 나오는 특징을 사용하기로 한다. 그 다음으로 메뉴명에 걸맞지 않은 데이터는 정확한 결과 도출을 저해하므로 미리 배제한다. 마지막으로 길이가 긴 명칭을 우선하여 매칭시킨다.

4. 군집화

대부분의 클러스터링 알고리즘은 거리기준을 어떻게 정하느냐에 따라 큰 영향을 받는다. 이 거리 기준의 선택에 대한 확실한 정답은 없으며, 임의로 선택하는 수밖에 없다. 문서분류의 경우 해당 문서에 존재하는 단어의 조건부 확률을 이용하여 군집화를 수행한다. 즉 두 문서가 존재할 때 해당 문서에 속한 단어들의 조건부 확률 분포가 얼마나 유사한지를 통해 군집화를 수행한다.

Tishby 등은 조건부 확률분포 사이의 유사성을 측정하기 위한 원칙을 제안했으며, 이러한 방법은 거리측정 기준의 선택이라는 문제를 피하도록 하였다[7]. 이 방법에서는 p(X,Y)의 결합 확률에 대해, 중요 변수 Y에 대해 가장 많은 정보를 유지하는 변수 X의 축약 표현을 찾는다. 변수 X가 변수 Y에 대해 정보를 포함하는 정도는 mutual Information을 이용하였다.

$$I(X,Y) = \sum_{x \in X, y \in Y} p(x)p(y \mid x) \log \frac{p(y \mid x)}{p(y)}$$

sIB는 Information Bottleneck 기법을 이용한 군집화 방법으로 일종의 agglomerative 접근법이다 (그림 1).

```
입력:
|X| 객체가 군집화
파라미터: K, n, maxL, ε

출력:
X를 K개의 군집으로 분류하는 분할 T

Loop:
   For i=1, …, n
      T_i ← X의 Random 분할
      c ← 0, C ← 0, done = FALSE
      while (not done)
         For j=1, …, |X|
            t(x_j) 에서 x_j를 추출
            t^new(x_j) = arg min_t' d_F({x_j}, t')
            If t^new(x_j) ≠ t(x_j) then c ← c+1
            Merge x_j into t^new(x_j)
         C ← C+1
         If C >= maxL or c <= ε|X| then
            done → TRUE
   T ← argmax_Ti F(T_i)
```

그림 1. sIB의 의사코드 (F는 score function이다)

5. 결과 및 분석

음식점 웹 사이트에 음식점주들이 직접 입력한 음식점 분류는 아래와 같으며 총 10 개로 구성되어 있다.

표 1. 음식점주들이 직접 입력한 군집 분류

카테고리	개수	카테고리	개수
한식	1218	양식	205
일식	168	중식	119
카페	58	세계음식	55
뷔페	46	주점	43
퓨전	8	기타	2

전처리를 거친 데이터들은 명료하게 한개의 주요 요리를 나타내는 메뉴명으로 정제되었다. 먼저 raw 데이터는 137 의 음식명, 1921 개의 음식점으로 전처리 되었다. 본 연구에선 군집의 개수를 5 로 하였으며, sIB이외에 EM, Farthest First, Simple K-Means, XMeans를 적용하여 비교하여 보았다.

표 2. 군집화 결과

EM			
Cluste	내용		
0	한식: 719	일식: 41	뷔페: 11
	양식: 11	카페: 10	중식: 9
	세계음식: 9	주점: 1	퓨전: 1
1	한식: 287	일식: 105	양식: 49
	중식: 30	뷔페: 18	세계음식: 15
	카페: 13	주점: 12	
2	중식: 74	한식: 21	세계음식: 12
	양식: 5	일식: 3	뷔페: 2
	카페: 2	기타: 1	주점: 1
	퓨전: 1		
3	양식: 139	카페: 33	주점: 24
	세계음식: 17	뷔페: 15	일식: 15
	중식: 6	퓨전: 3	한식: 2
	기타: 1		
4	한식: 189	주점: 5	일식: 4
	퓨전: 3	세계음식: 2	양식: 1

Farthest First (이하 FF)			
0	한식: 1217	양식: 205	일식: 168
	중식: 97	카페: 58	세계음식: 55
	뷔페: 46	주점: 43	퓨전: 8
	기타: 2		
1	중식: 4		
2	중식: 5		
3	한식: 1		
4	중식: 13		

sIB			
0	한식: 152	일식: 144	양식: 4
	주점: 4	뷔페: 3	중식: 2
	퓨전: 1		
1	한식: 414	뷔페: 10	중식: 7
	일식: 6	주점: 6	세계음식: 5
	양식: 5	퓨전: 1	
2	양식: 193	카페: 58	세계음식: 47
	주점: 33	뷔페: 32	한식: 21
	중식: 14	일식: 12	퓨전: 5
	기타: 2		
3	한식: 150	중식: 96	일식: 6
	세계음식: 3	퓨전: 1	뷔페: 1
4	한식: 481	양식: 3	

Simple K-Means (이하 SKM)			
0	한식: 864	일식: 160	양식: 89
	중식: 67	세계음식: 43	카페: 41
	뷔페: 38	주점: 22	퓨전: 3
	기타: 1		
1	한식: 190	중식: 2	양식: 1
2	양식: 114	주점: 19	카페: 16
	세계음식: 12	뷔페: 8	일식: 7
	퓨전: 3	중식: 2	기타: 1
3	중식: 48		
4	한식: 164	퓨전: 2	주점: 2
	양식: 1	일식: 1	카페: 1
XMeans			
0	양식: 113	주점: 21	카페: 15
	세계음식: 12	뷔페: 8	일식: 8
	퓨전: 3	중식: 2	기타: 1
1	중식: 47		
2	한식: 294	중식: 4	일식: 2
	양식: 1		
3	한식: 116	퓨전: 2	주점: 2
	카페: 1	일식: 1	양식: 1
4	한식: 758	일식: 157	양식: 90
	중식: 66	세계음식: 43	카페: 42

	뷔페: 38 주점: 20 퓨전: 3 기타: 1

본 연구에서는 음식점의 각 카테고리간의 비율을 이용해 순수도[8]를 구함으로써 판단을 하기로 한다. 하지만 입력값이 일반 데이터 수치가 아닌 카테고리와의 비율 수치이기 때문에 공식을 수정하였다.

$$purity(C_j) = \frac{1}{|C_j|} MAX_i(|C_j|_{class=i})^2$$

$$purity = \sum \frac{1}{|D|} purity(C_j)$$

|D| : 데이터 크기

$|C_j|$: C_j의 클러스터 크기

$|C_j|class=i$: C_j의 아이템 개수

그리고 이 수식을 이용하여 다음과 같은 결과를 얻었다. (p(n) = purity(C_n))

표 3. Purity 인덱스를 이용한 평가

클러스터명	p(0)	p(1)	p(2)	p(3)	p(4)	purity
EM	0.2	0.15	0.23	0.13	0.20	0.09
FF	0.1	0.03	0.04	0.00	0.11	0.03
sIB	**0.56**	**0.11**	**0.16**	**0.56**	**0.38**	**0.18**
SKM	0.16	0.14	0.12	0.40	0.14	0.10
XMeans	0.12	0.39	0.20	0.14	0.14	0.10

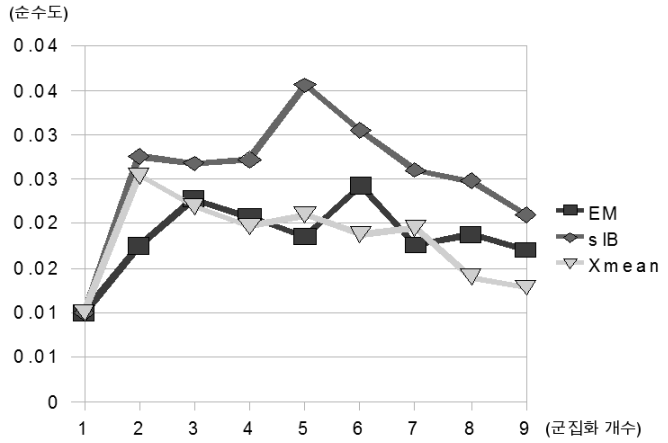
6. 추가분석

추가로 군집의 갯수가 순수도에 미치는 영향을 분석해 보았다. 분포도가 고르다고 여겨지는 EM,sIB,XMeans 알고리즘을 선택하였고, 분석방법은 이전과 마찬가지로 하되 1~9 개의 군집 개수에 대한 클러스터링 결과를 이용하였다.

표 4. EM, sIB, Xmean 순수도 분석 결과

클러스터명	EM	sIB	XMean
p(1)	0.01	0.01	0.01
p(2)	0.0175	0.0276	0.0254
p(3)	0.0228	0.0268	0.0219
p(4)	0.0208	0.0272	0.0197
p(5)	0.0186	0.0356	0.0210
p(6)	0.0243	0.0305	0.0188
p(7)	0.0176	0.0260	0.0195
p(8)	0.0188	0.0248	0.0144
p(9)	0.0170	0.0210	0.0128

군집 개수의 변화에 대한 영향을 알아보기 위해 기존의 순수도 공식 결과값을 군집 개수로 나눈 평균 순수도를 이용하였으며, 그래프로 표현하면 다음과 같다.



순수도 분석 결과 그래프

7. 결론

먼저 카테고리 분석에 있어서 한식의 수량이 너무나 차이가 심해 만족스러운 군집화 여부에 대한 우려에도 불구하고, sIB 알고리즘 사용시 사람에 의한 직접분류와 상당히 유사한 분류 결과가 나왔다.

이로 인해 메뉴의 이름만으로도 음식점의 카테고리 분류가 가능하다는 것을 알게 되었고, 이를 더욱 개선시켜 이용한다면, 단순하게 음식점의 메뉴 DB만을 이용하는 것만으로도 직접적인 추가 작업 없이 어떤 음식점이라도 분류가 가능하기 때문에 이를 이용한 음식점 추천 사이트는 사용자가 원하는 음식의 종류를 추천할 때에 단순히 알고 있는 음식점뿐만 아니라 그와 유사하면서도 그가 모르는 음식점을 같이 소개가 가능하기 때문에 사용자들의 본질적인 욕구에 맞는 서비스를 제공할 수 있다.

감사의 글

이 논문은 2010 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2010-0012876)

7. 참고문헌

[1] S. B. Austin et al. “Clustering of fast-food restaurants around schools: A novel application of spatial statistics to the study of food environments,” *American Journal of Public Health*, vol. 95, no. 9, pp. 1575-1581, 2005.

[2] 김광지, 박기용, “외식 범위 설정과 표준산업분류에 의한 외식산업 분류기준에 관한 연구,” 관광레저연구, vol. 19, no. 2, pp. 389-406, 2007.

[3] J. Peltonen, J. Sinkkonen, and S. Kaski, “Sequential information bottleneck for finite data,” *Proceedings of the 21st International Conference on Machine Learning*, pp. 82, 2004.

[4] N. Slonim, N. Friedman, and N. Tishby, “Unsupervised document classification using sequential information maximization,” *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 129-136, 2002.

[5] http://www.restaurant.org.

[6] 이계임.김민정, "외식통계의 현황과 개선방안," 한국농촌경제연구원, 연구보고 R513, 2005.

[7] N. Tishby, F. Pereira, and W. Bialek, “The information bottleneck method,” *Proceedings of the 37th Allerton Conference on Communication and Computation*, 1999.

[8] http://www.cse.iitm.ac.in/~cs672/purity.pdf