블로그 키워드 추출을 통한 음악 추천 기법

최홍구, 전상훈, 황인준
고려대학교 전기전자전파공학과
e-mail : {choihg,ysbhjun,ehwang04}@korea.ac.kr

Music Recommendation based on Blog Keyword Extraction

Hong-gu Choi, Sanghoon Jun, Eenjun Hwang
School of Electrial, Electronics Engineering, Korea University

요 약

본 논문에서는 블로그의 포스트로부터 주요 키워드를 추출하여 노래 가사 데이터와 유사도를 분석, 해당 블로그 포스트에 적합한 음악을 추천하는 기법을 제안한다. 또한, 블로거가 포스트마다 제시한 태그들도 주요한 키워드로서 활용한다. 이를 위해서, 첫째로 TF-IDF 기법을 사용하여 텍스트로 구성된 포스트의 중요 키워드를 추출한다. 둘째로 포스트의 태그와 추출된 키워드를 기반으로 유사한 노래 가사를 LSA 기법으로 검색하여 가장 높은 유사도를 갖는 음악을 선택, 적합한 음악으로써 추천한다. 사용자 만족도 평가 실험을 통해서 제안하는 기법이 실제 추천에 적합한지 검증한다.

1. 서론

최근, 온라인 공간에서 사용자의 관심사 및 일상에 따라 자유롭게 칼럼, 일기, 기사 등을 올리는 개인 웹 사이트인 블로그가 크게 확산되었다. 이러한 블로그 사용자들을 블로거(Blogger)라고 하며, 블로그의 게시물을 포스트(Post)라고 한다. 소셜 태그(Social tag)를 통해 포스트의 주제 및 키워드를 사용자가 명시함으로써 검색이 용이하도록 한다.

현재 블로그는 시각적 정보인 텍스트와 이미지를 주요 미디어 컨텐츠로 한다. 블로그에 배경음악(BGM)이 재생되는 경우도 있지만 전체 블로그에 동일한 음악이 재생되므로 포스트의 주제와 이질적인 경우가 대부분이다.

Katsuhiko Kaji 는 가사와 태그정보 사이의 유사도를 계산해 사용자에게 음악을 추천하는 연구를 진행하였다.[1] Kerstin Bischoff 는 역시 가사와 태그를 기반으로 음악을 추천하는 연구를 진행하였다.[2] Rui Cai 는 웹 문서를 분석하여 대표 키워드를 추출하고 그 키워드와 유사한 감정의 노래를 추천하는 연구를 진행하였다.[3]

본 논문에서는 포스트 주제에 맞는 음악을 자동으로 추천하는 기법을 제안하여 블로그의 시청각적인 측면을 모두 만족시킬 수 있도록 한다. 이를 위해, 첫째로 블로그 포스트의 텍스트정보로부터 노래 가사와의 비교에서 주요한 키워드가 될 수 있는 단어를 TF-IDF[4](Term Frequency - Inverse Document Frequency)알고리즘을 사용하여 추출한다. 추출된 키워드는 소셜 태그와 결합하여 키워드 집합을 생성한다. 둘째로 생성된 키워드 집합을 LSA[5](Latent Semantic Analysis)를 통해 유사한 노래가사를 검색한다. 키워드 셋과 가장 유사한 가사를 갖는 노래를 선택하여 추천한다. 제안한 기법을 검증하기 위해 사용자 만족도 평가 실험을 수행한다.

본 논문의 구성은 다음과 같다 2 장에서는 본 논문과 관련된 연구에 대해 기술한다. 3 장에서는 제안된 기법의 전체적인 과정에 대해 설명한다. 4 장에서는 사용자 만족도 평가를 통해 기법의 적합성을 검증한다. 5 장에서는 결론 및 향후 연구를 제시한다.

2. 관련연구

이 장에서는 본 논문과 관련되거나 사용된 기술들에 대해서 기술한다.

2.1. 키워드기반 음악 검색 및 추천

Katsuhiko Kaji 는 가사와 태그정보를 사용해서 음악의 종류와 사용자의 취향 사이에 유사도를 사용하여 재생목록을 만드는 방법을 제안했다[1]. 재생 목록은 세 단계를 거쳐 만들어진다. 첫 번째로 초기 재생목록은 데이터 베이스로부터 내용기반 추천에 의해 만들어진다. 두 번째로 트랜스코딩이 재생 목록을 사용자의 선호도와 상황에 따라 개선한다. 마지막으로 시스템과 사용자 간의 상호작용에 의해 재생 목록이 사용자에게 더 적합하게 된다. Kerstin Bischoff 는 음악을 추천하기 위해 가사와 태그를 기반으로 음악의 테마에 대한 태그를 추천하는 알고리즘을 제안했다[2]. 음악 테마에 관련된 태그는 음악 추천에 많은 도움이 되지만 사용자가 태그를 달기까지 시간이 필요하다. 따라서 이미 존재하는 태그와 가사 정보를 사용하여 테마에 관련된 태그를 추천하는 했다. Rui Cai 는 사용자가 웹 블러그와 같은 웹 문서를 읽을 때, 음악을 추천하는 MusicSense 를 제안했다. MusicSense 는 감정

측면에서 음악과 문서의 내용을 매치한다. 이를 위해, 단어 모음으로 감정의 혼합을 만드는 Generate 모델인 Emotional Allocation Modeling 을 제안했다. 음악은 메타 데이터와 관련된 웹 페이지로부터 텍스트 정보로 표현할 수 있다. 음악과 웹 문서는 Emotional Allocation Modeling 을 통해 감정의 혼합으로 표현된다. 주어진 웹 문서와 감성이 가장 일치하는 음악을 추천하게 된다.

웹 문서를 분석하여 대표 키워드를 추출하고 그 키워드와 음악에 태그 정보에 감정 라벨을 할당하고 유사한 감정을 가지는 음악을 추천하는 연구를 진행하였다[3].

2.2. TF-IDF

TF-IDF 가중치 모델은 정보검색을 위해서 문서 내부의 단어간 상대적 중요도를 평가하기 위해 문서의 표현방식으로 고안된 것이다. 이 TF-IDF 가중치로 표현된 문서는 정보검색엔진에서 주어진 질의어와 가장 유사한 문서들의 순위를 결정할 수 있게 할 뿐만 아니라, 유사 문서들의 그룹을 찾는 문서군집화를 용이하게 한다. TF-IDF 값이 큰 단어는 그것이 속한 문서의 주제 또는 의미를 결정지을 가능성이 크며, 따라서 이 측정치를 주요 키워드를 추출할 수 있는 척도로 활용할 수 있다.

tf 는 문서 내에서 해당 용어가 출현한 빈도수를 나타낸다. 본 논문에서 *tf* 가 계산되는 범위는 한 개의 블로그 포스트가 된다.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

n_{ij} 는 단어 t_i 가 문서 d_j 에서 출현한 회수이고, $\sum_k n_{k,j}$ 는 문서 d_j에서 모든 단어가 출현한 회수이다.

idf 는 전체 문서집합에서 해당 단어를 포함한 문서의 빈도수를 나타낸다. 본 연구에서는 노래 가사 데이터베이스가 전체 문서집합이 된다.

$$idf_i = \log \frac{|D|}{|\{d_i \mid t_i \in d_i\}|} \tag{2}$$

$|D|$는 문서집합에 포함되어 있는 문서의 수, $|\{d_i \mid t_i \in d_i\}|$ 는 단어 t_i 가 등장하는 문서의 수이다. *TFIDF* 값은 *tf* 값과 *idf* 값을 곱한 것이다.

$$TFIDF_{i,j} = tf_{i,j} \times idf_i \tag{3}$$

2.3. LSA(Latent Semantic Analysis)

LSA 는 개념적으로 동시출현(co-occurrence)정보를 이용하여 단어의 형태뿐만 아니라 의미를 이용하여 문서간의 유사도를 측정하는 방법이다. 여러 문서내의 공통 단어 출현에 대해 행렬을 생성하고, 해당 행렬의 분석을 통해 문서 간의 유사성을 판별한다. 이를 위해 LSA 는 선형대수학의 SVD(Singular Value Decomposition) 기술을 사용한다.

2.4 SVD(Singular Value Decomposition)

SVD[6]는 단어-문서 행렬을 단어 출현빈도수를 통해 생성하고 이를 3 개의 행렬로 분리함으로써 문서를 특정차원의 한 점으로 투영시키고, 문서간 유사도 측정이 용이하도록 하는 방법이다. 예를 들어 그림 1 과 같은 단어-문서 행렬 A 가 있다

A=

	문서1	문서2	문서3	문서4	문서5	문서6
여행	1	0	1	0	0	0
자전거	0	1	0	0	0	0
음식	1	1	0	0	0	0
자동차	1	0	0	1	1	0
휴가	0	0	0	1	0	1

(그림 1) 단어-문서 행렬

단어 × 문서 행렬은 다음과 같은 수식으로부터 분해가 가능하다.

$$A_{t \times d} = T_{t \times n} \times S_{n \times n} \times D_{d \times n}{}^{T} \tag{4}$$

t 는 단어 개수, *d* 는 문서 개수, *n* 은 *t* 와 *d* 중 작은 값을 취한다. 행렬 *A* 를 SVD 를 통해 분해하면 *T*, *S*, *D* 3 개의 행렬이 얻어진다. *T* 는 단어에, *D* 는 문서에 대

T

	문서1	문서2	문서3	문서4	문서5	문서6
여행	1	0	1	0	0	0
자전거	0	1	0	0	0	0
음식	1	1	0	0	0	0
자동차	1	0	0	1	1	0
휴가	0	0	0	1	0	1

X

S

2.16	0.00	0.00	0.00	0.00
0.00	1.59	0.00	0.00	0.00
0.00	0.00	1.28	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.39

X

D^T

	문서1	문서2	문서3	문서4	문서5	문서6
차원1	-0.75	-0.28	-0.20	-0.45	-0.33	-0.12
차원2	-0.29	-0.53	-0.19	0.63	0.22	0.41
차원3	0.28	-0.75	0.45	-0.20	0.12	-0.33
차원4	0.00	0.00	0.58	0.00	-0.58	0.58
차원5	-0.53	0.29	0.63	0.19	0.41	-0.29

(그림 2) $T \times S \times D^T$ 행렬

응되는 행렬이다.

그림 2 의 예제에서 각 단어와 문서는 5 차원 공간의 한 점으로 표현되고 있다. 실제 데이터를 적용한다면, 적용되는 차원의 수는 수백에서 수만이 될 수 있기 때문에 $S \times D^T$ 행렬의 차원감쇄는 반드시 필요하다. SVD 는 여기서 n 보다 작은 k 값을 설정해서 k 차원으로 감쇄를 진행한다. 행렬 스케일 행렬 S 의 (d,d) 에 해당하는 값은 d 가 1 에서 5 로 증가함에 따라 감소하는 경향을 보인다. 즉, 더 높은 차원으로 갈수록 스케일 값이 줄어드는데, 여기서 높은 스케일 값만 취함으로써 차원의 감쇄가 가능하다.

2.5 유사도 계산

2.4 SVD 의 예에서 구했던 3 개의 행렬을 $k = 2$ 로 차원 감쇄를 하면, 그림 3 과 같은 감쇄된 차원이 된다.

T

	차원1	차원2
여행	-0.44	-0.30
자전거	-0.13	-0.33
음식	-0.48	-0.51
자동차	-0.70	0.35
휴가	-0.26	0.65

S

2.16	0.00
0.00	1.59

D^T

	문서1	문서2	문서3	문서4	문서5	문서6
차원1	-0.75	-0.28	-0.20	-0.45	-0.33	-0.12
차원2	-0.29	-0.53	-0.19	0.63	0.22	0.41

(그림 3) $T \times S \times D^T$ 행렬

예제에서 $S \times D^T$ 행렬을 구해보면 그림 4 와 같다.

	문서1	문서2	문서3	문서4	문서5	문서6
차원1	-1.62	-0.60	-0.44	-0.97	-0.70	-0.26
차원2	-0.46	-0.84	-0.30	1.00	0.35	0.68

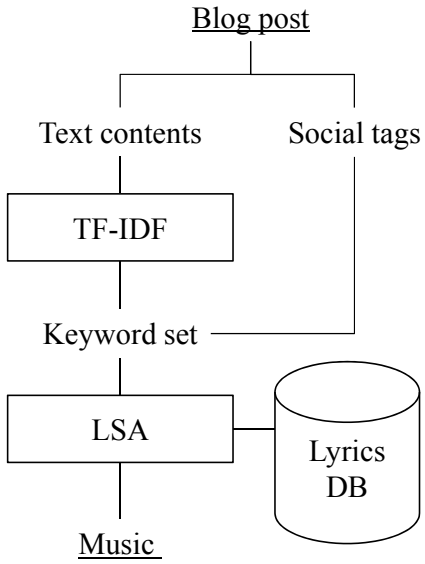
(그림 4) $S \times D^T$ 행렬

$S \times D^T$ 행렬은 D^T 행렬을 S 행렬에 의해 re-scaling 한 행렬로, 각 문서는 2 차원상의 한 점으로 투영될 수 있다. 예제에서 문서 1 과 문서 2 의 유사도는 코사인 거리(cosine distance)를 계산함으로써 도출할 수 있다.

3. 시스템 구조 및 구현

이 장에서는 본 논문과 관련된 기술들에 대해서 기술한다. 블로그 포스트에 적합한 음악 추천의 전체적인 과정은 그림 5 와 같다.

제안된 시스템의 입력은 텍스트형식의 블로그 포스트이고, 출력은 추천된 음악이다. 이를 위해, 음악을 추천하고자 하는 블로그 포스트의 TF 와 해당 사용자의 전체 포스트에 대한 IDF 를 계산하여 포스트의 주요한 키워드를 추출한다. 추출된 키워드는 포스트의 태그와 결합하여 키워드 집합이 된다. 생성된 키워드 집합은 하나의 문서로 간주되어, 노래가사 데이터베



(그림 5) 시스템 구조

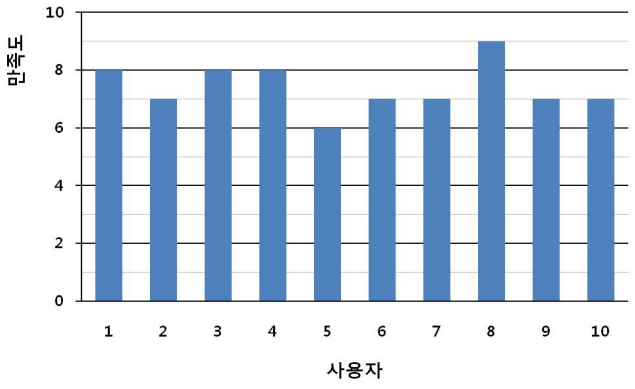
이스의 가사들과 LSA 를 수행한다. 이 결과로 나온 각 노래가사들 가운데 가장 높은 유사도 값을 갖는 음악이 추천음악으로 선택된다.

구현을 위해서 Mysql 기반의 노래가사 Database 와 PHP 기반의 Apachi 웹서버가 사용되었다. 노래가사 데이터는 400 개를 수집하였다.

4. 실험 및 결과

이 장에서는 본 논문에서 제안한 음악 추천 시스템에 대한 사용자 만족도 평가를 통해 얻은 결과를 분석한다.

실험 결과에 대한 평가를 위해 현재 블로그를 사용하고 있는 10 명의 피험자를 대상으로 각각 작성한 10 개의 포스트에 대한 음악 추천을 수행하였다. 피험자들은 추천을 통해 나온 음악에 대한 적합성을 평가하고, 0~10 사이의 만족도 점수로 평가한다. 그림 6 은 수행 후 만족도 조사에 대한 결과이다.



(그림 6) 피실험자 만족도 조사 결과

사용자 만족도 조사 결과 사용자 8 이 가장 높은 만족도를 보였다. 사용자 8 의 블로그 포스트에는 여행, 휴가, 바다 등과 같은 여행과 관련된 단어가 많이 사용되었다. 따라서 사용자의 만족도가 높은 음악을 추천할 수 있었다. 반면 사용자 5 의 경우 가장 낮은 만족도를 보였다. 사용자 5 의 블로그는 영화 리뷰를 주제로 하는 정보 전달을 위해 운영되고 있었고 음악 추천에 적합한 키워드를 추출할 수 없었다.

만족도의 평균은 7.4 로써, 대체로 만족스럽다는 평가로 볼 수 있다.

5. 결론

본 논문에서는 블로그의 포스트와 소셜 태그를 분석하여 포스트 내용에 적합한 음악을 추천하는 기법을 제안했다. 포스트의 텍스트정보 중에서 주요한 키워드를 추출하기 위해 TF-IDF 를 적용하였고, 생성된 키워드 셋을 노래 가사 DB 에 저장된 데이터들과 비교하여 적합한 음악을 선택하기 위해 LSA 를 적용하였다. 실험을 통해 사용자에게 비교적 만족스런 결과를 도출할 수 있었다. 향후에는 좀 더 다양한 키워드 추출 알고리즘과 유사도 측정 알고리즘을 사용하고 신호적 특징을 고려하여 보다 효과적인 기법 연구를 진행할 것이다.

Acknowledgements

“본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT 연구 센터 지원사업의 연구결과로 수행되었음” (NIPA – 2010 – C1090 – 1011 – 0008)

참고문헌

[1] Kaji, K., Hirata, K., Nagao, K., “A music recommendation system based on annotations about listeners’ preferences and situations,” Automated Production of Cross Media Content for Multi-Channel Distribution, (2005). AXMEDIS 2005. First International Conference on 30 Nov.-2 Dec. 2005 Page(s):4pp.

[2] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. (2009) “Deriving music theme annotations from user tags,” WWW ’09: Proceedings of the 18th international conference on World wide web, ACM, New York, USA, pp. 1193-1194

[3] R. Cai, C. Zhang, C. Wang, L. Zhang, and W.-Y. Ma. (2007) “MusicSense : Contextual music recommendation using emotion allocation modeling. In: Proc. ACM Multimedia, pp. 553-556

[4] Salton, Gerard and Buckley, C. (1988). "Term-weighting approaches in automatic text retrieval". Information Processing & Management 24 (5): 513–523

[5] The Latent Semantic Indexing home page http://lsa.colorado.edu/

[6] Wall, Michael E., Andreas Rechtsteiner, Luis M. Rocha (2003). "Singular value decomposition and principal component analysis". in A Practical Approach to Microarray Data Analysis. D.P. Berrar, W. Dubitzky, M. Granzow, eds. pp. 91–109, Kluwer: Norwell, MA