Journal of Korea Multimedia Society Vol. 21, No. 1, January 2018(pp. 45-50)
https://doi.org/10.9717/kmms.2018.21.1.045

TextRank 알고리즘을 이용한 음악 가사 요약 기법

손지영†, 신용태††

Music Lyrics Summarization Method using TextRank Algorithm

Jiyoung Son†, Yongtae Shin††

ABSTRACT

This research paper describes how to summarize music lyrics using the TextRank algorithm. This method can summarize music lyrics as important lyrics. Therefore, we recommend music more effectively than analyzing the number of words and recommending music.

Key words: Music Information Retrieval, Summary, TextRank, TF-IDF, Lyrics Analysis

1. 서 론

최근 음악이 대량으로 온라인상에 유통되면서 사용자가 원하는 음악을 찾기 위한 음악 정보 기술(Music Information Retrieval)의 중요성이 강조되고 있다. 음악을 추천해주는 서비스로는 스포티파이(Spotify)[1]의 'Discovery Weekly'와 멜론(Melon)[2]의 '음악 추천 서비스' 등이 있다. 스포티파이의 'Discovery Weekly'는 사용자의 음악 취향과 저장한 음악을 바탕으로 곡을 추천하는 서비스이고 멜론의 '음악 추천 서비스'는 사용자의 선호 아티스트, 유사 아티스트, 선호 장로 등을 기준으로 분석하여 개인 맞춤 추천 서비스를 제공하고 있다. 음악 추천 서비스에서는 음악의 멜로디나 음향적 특성과 같은 내용(Content)에 관한 연구가 주를 이루며 관련 연구에는 음악파일의 무드 분석을 통한 효과적인 음악분류 및 검색[1]이 있다. 사용자가 음악을 선택할 때, 자신이 공감한 멜로디를 선택하기도 하지만 자신이 공감한 노래 가사를 선택하여 듣기도 한다. 특정 집단에서 대중음악에 대한 특별한 선호가 없는 경우, 음악의 스타일보다 가사가 주는 영향력이 크다는 것으로 분석되었고 이는 음악 추천에 노래 가사가 중요한 의미를 지니고 있다는 것을 알 수 있다[2]. 노래 가사를 이용하여 음악을 분류하고 사용자의 음악적 성향을 추천 점수에 가중치로 반영하는 연구에서는 가사 분석과 함께 다른 메타데이터를 활용하는 방법을 결합한다면 추천 서비스 시스템이 개선될 것이라는 가능성을 제시하였다[3]. 음악 가사를 분석하는 방법으로는 가사에서 각 단어를 추출하는 방법이 있다. 기존의 연구에서는 가사를 하나의 텍스트로 간주하여 각 단어의 감정을 추출 후 벡터의 합을 계산하는 방법이 있다[4,5]. 하지만, 노래 가사는 주로 '시'의 형식 사용된다. 함축성과 다의성이 대표되는 시의 특성상 단어 자체만의 의미만으로 내용을 분석할 경우 의미가 왜곡될 가능성이 있다고 판단하였다[6]. 따라서 노래 가사 단어가 내포하고 있는 감성을 분석하는 방법이 아닌, 문장 자체가 의미하는 감성을 분석해야 한다.

문서를 요약하는 방법으로는 Google의 PageRank 알고리즘을 텍스트에 적용한 TextRank 알고리즘이

※ Corresponding Author : Yongtae Shin, Address: (06978) 369, Sangdo-ro, Dongjak-gu, Seoul, Republic of Korea, TEL : +82-2-820-0681, FAX : +82-2-828-7015, E-mail : shin@ssu.ac.kr
Receipt date : Oct. 24, 2017, Revision date : Dec. 8, 2017
Approval date : Dec. 27, 2017

† Dept. of Software, Graduate School of Software Soongsil University (E-mail : showhisper@gmail.com)
†† Dept. of Computer Science & Engineering, Soongsil University

1) Spotify, https://www.spotify.com/int/why-not-available/
2) Melon, http://www.melon.com/

있다[7]. TextRank는 Mihalcea 와 Tarau가 제안한 텍스트를 처리하기 위한 그래프 기반 알고리즘이다. 전체 텍스트에서 반복적으로 가져온 정보를 이용하기 때문에 다양한 항목 간의 연결을 식별할 수 있고 추천 단위의 중요도를 기반으로 재귀적으로 계산하기 때문에 전체 문장을 효과적으로 요약할 수 있다. TextRank 알고리즘을 이용한 기법으로는 정보과다 현상을 해결하기 위한 타임라인 요약기법이 있다. 이 기법은 타임라인을 정리하고 비슷한 단어들을 하나로 묶은 개념의 알고리즘이 적용되어, 빈도수 기반의 방법보다 향상된 성능을 보여준다[8].

본 논문에서는 TextRank 알고리즘과 TF-IDF (Term Frequency - Inverse Document Frequency)를 이용하여 노래 가사 요약 기법을 제시하고 구현한다. 멜론의 1964년부터 2016년 시대별 차트 TOP50의 가사 데이터를 이용하여 전체 가사를 주요 문장으로 요약하였다.

본 논문의 구성은 다음과 같다. 2장에서 본 논문의 기본이 되는 TextRank 알고리즘과 TF-IDF에 대하여 정의한 다음, 3장에서 제안하는 음악 가사 요약 기법을 설명한다. 여기에서는 음악 가사 요약 기법을 구현한 전체 프로세스 순서로 서술한다. 그리고 4장에서 제시한 가사 요약 기법을 구현한 결과를 분석하고 5장에서 결론을 맺는다.

2. 이 론

2.1 TextRank 알고리즘

TextRank 알고리즘은 텍스트에서 정점(Node)이 될 만한 단위(Text Unit)를 뽑아내고, 이 단위들의 연결(Edge)을 지정하여 PageRank 알고리즘을 적용한다[7]. TextRank 알고리즘은 다음과 같이 정의된다.

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \quad (1)$$

$WS(V_i)$는 문장 또는 단어 (V_i)에 대한 TextRank 값이고 d는 damping factor로 주어진 정점에서 그래프의 다른 임의의 정점으로 점프할 확률을 모델에 통합하는 역할을 한다. w_{ji}는 연관된 점수를 계산할 때, 문장 또는 단어 i와 j사이의 가중치를 의미하며 TextRank $WS(V_i)$을 계산하여 높은 순으로 정렬한다[7].

2.2 TF-IDF 모델

TF-IDF는 정보 검색에서 접하는 가중치를 구하는 알고리즘이다. TF(Term Frequency)는 문서 내 특정 단어의 빈도를 의미하고 IDF(Inverse Document Frequency)는 DF(Document Frequency)의 역수로 DF는 한 단어가 전체 문서에서 얼마나 공통으로 나타나는지를 의미한다. TF-IDF는 특정 문서에서 단어의 중요도를 평가하는 데 일반적으로 사용되고 있는 방법으로, 문서들 사이의 비슷한 정도를 구하는 용도로 사용된다[9].

3. 제안한 방법

3.1 제안한 음악 가사 요약 방법

본 논문에서는 음악 추천 서비스 시스템을 위한 음악 가사 요약 기법을 제시한다. Fig. 1에 TextRank 알고리즘을 이용한 가사문장 요약 프로세스 흐름도를 나타내었다.

먼저, 요약할 가사 데이터를 멜론 시대별 가요 차트 TOP50을 이용하여 수집한다. 수집된 가사 데이터를 가사와 상관없는 특수문자를 제거하여 데이터를 전처리한다. 다음으로 전처리된 데이터를 이용하여 Fig. 2와 같이 TF-IDF 모델을 만들고 이를 이용하여 Fig. 3과 같이 가중치 그래프를 생성한다. 그다음, TextRank를 적용하여 높은 순으로 가사를 정렬한 뒤 가사를 4문장 이하로 요약한다.

4. 구현 및 결과

4.1 가사 수집

멜론의 시대별 가요 차트(1964-2016) TOP 50의

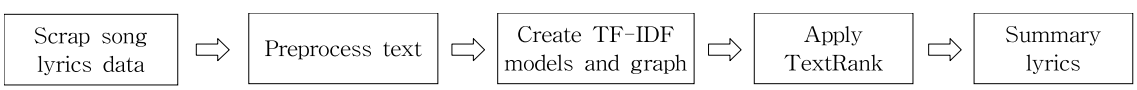

Fig. 1. Music lyric summary processing using TextRank algorithm.

	T_1	T_2	·	·	·	T_m
$Sentence_1$	w_{11}	w_{12}	·	·	·	w_{1m}
$Sentence_2$	w_{21}	w_{22}	·	·	·	w_{2m}
·	·	·	·			·
·	·	·		·		·
·	·	·			·	·
$Sentence_n$	w_{n1}	w_{n2}	·	·	·	w_{nm}

Fig. 2. Sentence-Term Matrix.

데이터를 BeautifulSoup[3]을 이용하여 가사 데이터를 수집한다.

4.2 Text 전처리

먼저 가사와 상관없는 특수 문자를 제거한 후 문장으로 데이터를 정제한다. Fig. 4에 수집된 가사를 줄 단위 기준으로 문장 처리한 결과를 나타내었다. 다음으로 줄 단위의 문장 길이가 너무 짧을 경우 가사의 종결 어미를 이용하여 종결 어미로 끝나는 문장과 앞의 문장을 합쳐 하나의 문장이 되도록 처리하여 Fig. 5에 나타내었다. 가사의 문장 종결 어미는 노래에 따라 달라진다. 그렇기 때문에 1995년부터 2015년도 멜론의 시대별 차트 가요 TOP 50의 가사를 분석하여 가사에 자주 사용된 종결 어미를 분석하여 Table 1로 나타내었다. 분석 결과가 100번 이상 사용된 종결 어미 '어, 요, 지, 해, 아, 야, 데, 다, 죠, 니, 나, 가, 봐, 까, 라, 줘, 걸, 마, 워, 네, 리, 돼, 여, 져, 께'를 이용하여 전처리한다.

Fig. 5와 같이 처리된 문장을 TF-IDF 모델로 만들기 위해 분리된 문장을 KoNLPy[4]의 Twitter를 이용하여 명사를 추출한다[10].

4.3 TF-IDF 모델 생성 및 그래프 생성

TF-IDF 모델을 Python의 머신러닝 패키지 Scikit-learn[5]을 이용하여 구현한 뒤 명사로 구성된 문장을 Vector 공간으로 나타낸 뒤 Fig. 6과 같이 문장과 명사로 이루어진 행렬을 받아 전치행렬을 곱하여 Fig. 6과 같은 상관행렬을 생성한다. TextRank 알고리즘

	$Sentence_1$	$Sentence_2$	·	·	·	$Sentence_n$
$Sentence_1$	1	w_{12}	·	·	·	w_{1n}
$Sentence_2$	r_{21}	1	·	·	·	w_{2n}
·	·	·	·			·
·	·	·		·		·
·	·	·			·	·
$Sentence_n$	r_{n1}	w_{n2}	·	·	·	1

Fig. 3. Correlation Matrix of Sentence-Term Matrix.

'₩r₩n₩t₩t₩t₩t내 곁에만 머물러요 떠나면 안돼요',
'그리움 두고 머나먼 길',
'그대 무지개를 찾아올 순 없어요',
'노을 진 창가에 앉아',
'멀리 떠가는 구름을 보면',
'찾고 싶은 옛 생각들 하늘에 그려요',
'음 불어오는 차가운 바람 속에',
'그대 외로워 울지만',
'나 항상 그대 곁에 머물겠어요',
'떠나지 않아요',
'노을 진 창가에 앉아',
'멀리 떠가는 구름을 보면',
'찾고 싶은 옛 생각들 하늘에 그려요',
'음 불어오는 차가운 바람 속에',
'그대 외로워 울지만',
'나 항상 그대 곁에 머물겠어요',
'떠나지 않아요',
'내 곁에만 머물러요 떠나면 안돼요',
'₩n'

Fig. 4. Some of the lyrics of Oh, Hyuk's "Girl" data.

3) BeautifulSoup, https://www.crummy.com/software/BeautifulSoup/bs4/doc/#
4) KoNLPy, http://konlpy-ko.readthedocs.io/ko/v0.4.3/
5) Scikit-learn, http://scikit-learn.org/stable/

Table 1. The analysis of the conjunctive endings table from 1995 to 2015

	어	요	지	해	아	야	데	다	죠	니	나	가	봐	까	라	줘	걸	마	워	네	리	돼	여	져	께	오	냐	렴	군
1995	124	24	102	52	62	85	22	9	7	19	44	28	7	22	12	7	24	26	6	28	16	1	6	1	2	1	1	7	0
1996	140	51	74	50	48	112	29	23	20	32	18	19	33	17	20	31	32	39	11	4	5	6	4	6	11	1	5	2	0
1997	131	123	65	60	48	76	38	37	20	29	20	27	17	12	9	36	24	9	1	16	15	9	9	1	6	10	6	1	6
1998	102	116	118	47	57	55	17	23	23	43	41	34	24	22	9	38	26	24	14	9	5	5	9	3	11	5	0	2	0
1999	171	54	91	80	53	93	32	28	32	50	30	33	23	17	35	40	56	29	13	6	19	17	1	1	7	0	0	0	0
2000	125	120	50	52	57	50	30	10	33	46	38	40	19	30	8	18	26	25	10	7	10	6	8	7	4	10	0	0	0
2001	79	134	94	35	29	30	38	15	56	29	24	30	13	39	18	12	31	6	3	13	9	6	13	6	2	2	3	0	0
2002	73	162	67	56	56	42	38	30	33	22	29	26	15	12	20	12	12	4	10	7	7	3	8	4	7	8	21	0	0
2003	100	88	72	44	87	45	53	23	29	22	25	22	36	32	15	28	27	31	14	3	5	6	7	2	8	0	2	7	1
2004	94	119	73	46	48	37	56	17	56	29	31	24	29	27	9	30	20	1	7	27	3	9	15	3	4	7	0	1	0
2005	91	115	76	53	40	74	47	40	51	39	40	23	37	34	39	18	19	15	20	6	7	13	4	4	8	10	2	0	0
2006	93	158	40	52	59	33	78	37	71	33	22	20	48	34	18	13	14	25	13	8	2	17	9	5	8	2	1	0	0
2007	78	161	59	64	81	39	52	55	99	45	38	25	31	26	25	46	13	11	14	9	8	6	5	2	10	0	2	0	0
2008	125	132	82	109	69	34	50	66	57	39	35	39	39	35	74	38	19	11	22	25	13	6	5	15	3	2	4	0	0
2009	137	55	69	118	56	53	74	59	16	44	37	34	87	39	53	36	14	19	21	15	11	14	3	12	6	2	2	0	0
2010	97	57	48	108	68	33	68	51	16	34	43	25	43	24	64	15	25	24	5	7	6	2	2	10	10	9	5	0	0
2011	181	49	63	89	82	49	31	20	1	20	17	41	33	25	30	14	28	51	35	27	6	24	5	10	2	10	3	0	0
2012	86	95	66	48	92	62	45	84	23	20	47	39	44	33	42	26	21	14	36	13	34	9	28	11	3	3	7	1	0
2013	149	134	86	96	76	64	57	54	25	32	31	66	18	46	30	13	29	17	33	22	15	21	13	12	7	9	9	0	0
2014	139	70	58	112	64	53	45	73	12	33	44	57	45	41	15	28	17	21	20	24	42	16	6	20	0	18	15	0	0
2015	147	140	78	105	102	72	38	41	34	37	41	34	42	36	32	24	11	26	20	24	26	23	9	5	1	5	2	0	0
SUM	2462	2157	1531	1476	1334	1191	938	795	714	697	695	686	683	603	577	523	488	428	328	300	264	219	169	140	120	114	90	21	7

'내 곁에만 머물러요 떠나면 안돼요',
'그리움 두고 머나먼 길 그대 무지개를 찾아올 순 없어요',
'노을 진 창가에 앉아',
'멀리 떠가는 구름을 보면 찾고 싶은 옛 생각들 하늘에 그려요',
'음 불어오는 차가운 바람 속에',
'그대 외로워 울지만 나 항상 그대 곁에 머물겠어요',
'떠나지 않아요',
'노을 진 창가에 앉아',
'멀리 떠가는 구름을 보면 찾고 싶은 옛 생각들 하늘에 그려요',
'음 불어오는 차가운 바람 속에',
'그대 외로워 울지만 나 항상 그대 곁에 머물겠어요',
'떠나지 않아요',
'내 곁에만 머물러요 떠나면 안돼요'

Fig. 5. Lyric data that processed the sentences in Figure 4.

	0	1	2	3	4	5	6	7	8	9	10	11
0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0	0
2	0	0.501804	0	0	0	0.864981	0	0	0	0	0	0
3	0	0	0	0.707107	0	0	0	0	0	0.707107	0	0
4	0.707107	0	0	0	0.707107	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0.707107	0	0	0.707107	0
6	0	0	0	0	0	0	1	0	0	0	0	0
7	0	0.577942	0	0	0	0	0	0	0.816078	0	0	0
8	0	0.577942	0	0	0	0	0	0	0	0	0	0.816078
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0.707107	0	0	0	0	0	0.707107	0	0
11	0.707107	0	0	0	0.707107	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0.707107	0	0	0.707107	0
13	0	0	0	0	0	0	1	0	0	0	0	0
14	0	0.577942	0	0	0	0	0	0	0.816078	0	0	0
15	0	0.577942	0	0	0	0	0	0	0	0	0	0.816078
16	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 6. Sentence-Term Matrix of Oh, Hyuk's "Girl" data.

	0	1	2	3	4	5	6	7	8	9	10	11
0	1	0	0	0	1	0	0	0	0	0	0	0
1	0	1	0	0	0	0.447213595	0	0	0.632455532	0	0	0.632455532
2	0	0	1	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	1	0	0
4	1	0	0	0	1	0	0	0	0	0	0	0
5	0	0.447213595	0	0	0	1	0	0	0	0	0	0
6	0	0	0	0	0	0	1	0	0	0	0	0
7	0	0	0	0	0	0	0	1	0	0	1	0
8	0	0.632455532	0	0	0	0	0	0	1	0	0	0
9	0	0	0	1	0	0	0	0	0	1	0	0
10	0	0	0	0	0	0	0	1	0	0	1	0
11	0	0.632455532	0	0	0	0	0	0	0	0	0	1

Fig. 7. Correlation Matrix of Oh, Hyuk's "Girl" data.

을 수행하기 위해 각 문장이 노드이고 두 문장 사이의 연결이 두 문장 사이의 유사성을 지정하는 가중치 그래프를 상관행렬을 이용하여 구성한다[11,12].

4.4 TextRank 알고리즘 적용

Fig. 7과 같이 생성된 단어 가중치 그래프를 사용하여 TextRank 알고리즘 식(1)을 통해 Rank 값이 높은 순으로 정렬한 결과를 Fig. 8에 나타내었다. 가사의 특성상 중복되는 문장이 많기 때문에 문장이 중복될 경우 중복된 문장을 배제해 결과적으로 4문장 이하의 주요 문장이 뽑히게 된다. 식(1)에 사용되는 damping factor는 0.85를 사용하였다.

Rank 값이 높은 순으로 정렬한 명사의 값 또한 확인할 수 있으며 TF와 IDF 값은 Table 2와 같다.

1. 나 항상 그대 곁에 머물겠어요
2. 그대 외로워 울지만

Fig. 8. Example of Oh, Hyuk's "Girl" summary sentences.

Table 2. TF and IDF of words in Oh, Hyuk's "Girl" data

Term	TF	IDF
그대	5	1.918918919
구름	2	1
노을	2	1
멀리	2	1
생각	2	1
하늘	2	1
외로워	2	0.752518191
항상	2	0.752518191
무지개	1	0.576044699
그리움	1	0.15
바람	2	0.15

5. 결 론

본 논문에서는 TextRank 알고리즘을 이용하여 음악 가사 내용을 요약하는 기법을 제시하고 구현하였다. 구현 결과, 전체 문장의 특정 단어 빈도수(TF)가 높다고 하여 가사의 특정 단어의 중요도(IDF)가 높은 것이 아니라는 결론이 나왔다. 이러한 결론은 가사의 특정 단어에 대한 빈도수가 높아도 그 단어가 노래를 대표하는 단어로 정의하기 어렵다는 것을 의미한다. 따라서 제안한 음악 가사 요약 기법을 이용하여 가사를 문장 그대로 요약하고 요약된 문장을 분석하여 음악 추천 서비스에 활용한다면 효과적인 음악 추천 서비스를 제공할 수 있다. 향후 정교한 전처리과정이 진행된다면 더 좋은 결과를 얻을 것으로 예상하며 본 논문에서는 TextRank 알고리즘을 이용한 노래 가사 요약 기법을 제시하고 구현해 보는 것에 의의를 둔다.

REFERENCE

[1] G. Park, S. Park, and S. Kang, "Effective Mood Classification Method Based on Music Segments," *Journal of Korea Multimedia Society*, Vol. 10, No. 3, pp. 391-400, 2007.

[2] S. Jin and B. Choi, "The Influences of Song Lyrics for the Perceived Preferences of Music Listening," *Korean Journal of Music Therapy*, Vol. 8, No. 1, pp. 74-82. 2006.

[3] C. Lee, H. Bang, and J. Lee, "Music Recommendation System Using Text Analysis of

Lyrics," *Proceeding of the Fall Conference of the Korea Institute of Intelligent Systems*, pp. 99-100, 2015.

[4] J. Lee, H. Lim, and H. Kim, "Similarity Evaluation of Popular Music Based on Emotion and Structure of Lyrics," *Korea Institute of Information Scientists and Engineers Transactions on Computing Practices,* Vol. 22, No. 10, pp. 479-487, 2016.

[5] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," *Computer Networks and Integrated Service Digital Network Systems*, Vol. 30, pp. 107-117, 1998.

[6] S. Koh, "A Study on Applying Storytelling for teaching Song-Lyrics: Focusing on Story-Making," *Teacher Education Research,* Vol. 54, No. 4, pp. 561-572, 2015.

[7] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts," *Proceeding of the Conference on Empirical Methods in Natural Language Processing,* pp. 404-411, 2004.

[8] I. An, H. Kim, and H. Kim, "A User Timeline Summarization Technique Using TextRank Algorithm," *Journal of Koreastudies Information Sservice System : Databases*, Vol. 39, No. 4, pp. 238-245, 2012.

[9] S. Lee and H. Kim, "Keyword Extraction from News Corpus Using Modified TF-IDF," *The Journal of Society for e-Business Studies,* Vol. 14, No. 4, pp. 59-73, 2009.

[10] S. Robertson, "Understanding Inverse Document Frequency: on Theoretical Arguments for IDF," *Journal of Documentation*, Vol. 60, No. 5, pp. 503-520, 2004.

[11] SpookyQubit/TextRank Github, https://github.com/spookyQubit/TextRank/ (accessed, Oct., 17, 2017).

[12] Summa NLP, https://github.com/summanlp/textrank (accessed, Oct., 17, 2017).

손 지 영

2015년 연세대학교 작곡과(학사)
현재 숭실대학교 SW특성화대학원 소프트웨어전공(석사)
관심분야 : 자연어처리, 기계학습, 데이터마이닝, IT융합, Web Design

신 용 태

1985년 한양대학교 산업공학과(학사)
1990년 Univ. of Iowa Computer Science(석사)
1994년 Univ. of Iowa Computer Science(박사)

1995년~현재 숭실대학교 컴퓨터학부 교수
2014년~현재 숭실대학교 소프트웨어특성화대학원 원장
2015년 개방형컴퓨터통신연구회 회장
2014년 한국인터넷윤리학회 회장
관심분야 : IoT, 정보보호, 방송 컨텐츠 보안, 차세대 인터넷 기술, IT융합