클러스터 기반 키워드 랭킹 기법

유한묵*, 김한준*
*서울시립대학교 전자전기컴퓨터공학과
e-mail:khj@uos.ac.kr

Cluster-based keyword Ranking Technique

Han-mook Yoo*, Han-joon Kim*
*School of Electrical and Computer Engineering, University of Seoul

요 약

본 논문은 기존의 TextRank 알고리즘에 상호정보량 척도를 결합하여 군집 기반에서 키워드 추출하는 ClusterTextRank 기법을 제안한다. 제안 기법은 k-means 군집화 알고리즘을 이용하여 문서들을 여러 군집으로 나누고, 각 군집에 포함된 단어들을 최소신장트리 그래프로 표현한 후 이에 근거한 군집 정보량을 고려하여 키워드를 추출한다. 제안 기법의 성능을 평가하기 위해 여행 관련 블로그 데이터를 이용하였으며, 제안 기법이 기존 TextRank 알고리즘보다 키워드 추출의 정확도가 약 13% 가량 개선됨을 보인다.

1. 서론

최근 인터넷 기술의 발달에 따라 웹에 등록되는 문서의 양이 급격하게 증가하고 있다. 이러한 문서들로부터 핵심 정보를 추려내는 기술의 필요성이 증대되고 있으며, 키워드 추출(keyword extraction)이 그 중 하나이다. 키워드는 문서의 내용을 포괄할 수 있는 대표 단어를 의미하며, 이는 문서 요약 또는 단어망 구성에 활용될 수 있다. 이러한 키워드를 추출하는 방법에는 크게 단어 통계량 방식과 단어 그래프 방식이 있다. 첫째, 단어 통계량 방식은 단어의 중요도(importance)를 문서로부터 통계적 정보를 활용하여 평가하는 방식이다. 둘째, 단어 그래프 방식은 단어의 중요도를 그래프상에서의 주변 단어들과의 관계를 고려하여 평가하는 방식이다. 최근에 이 방식에 속하는 기법으로서 잘 알려진 것이 TextRank 알고리즘이다[1]. TextRank 알고리즘은 그래프 기반 키워드 랭킹 알고리즘으로서, 특정 단어의 중요도를 이에 이웃한 단어들의 중요도를 고려하여 평가하는 알고리즘이다. 여기서 그래프는 문서의 각 단어가 노드(node)로 표현되고 단어 간의 관계가 간선(edge)으로 표현된다. 이 키워드 추출 방법은 시간의 흐름에 따라 트렌드가 변화하고 문서 내 고유명사를 포함하고 있는 문서 데이터(예: 관광 관련 소셜 데이터)에 매우 효과적이다. 그러나 이 기법은 현재 대중적이지 않거나 성장 가능성이 있지만 아직 대중화 되지 않은 주제를 담고 있는 문서들의 경우에는 키워드 추출의 정확도가 높지 못하다. 이는 전체 문서 집합에서 상대적으로 작은 비율을 가지는 주제의 문서에 포함된 단어들 간의 관계가 다른 주제의 문서들에 비해 상대적으로 약하기 때문이다. 이러한 TextRank 알고리즘의 문제점을 극복하기 위해서 본 논문은 유사한 문서들이 동일한 집단에 속하도록 군집화(clustering)하여 각 군집별로 키워드 추출을 수행하는 기법을 제안한다. 기존 TextRank 알고리즘은 단어 간의 관계를 주어진 문서 집합 전체에서 정의하기 때문에, 정확한 키워드 추출을 위해 군집화된 문서집합에서 각 군집 정보량을 산출하는 것이 중요하다.

제안 기법은 문서들을 k-means 군집화 알고리즘을 활용하여 여러 군집으로 나누고, 각 군집에 포함된 단어들을 최소신장트리(Minimal Spanning Tree) 그래프로 표현한 후, 이에 근거한 군집 정보량을 고려하여 단어의 중요도를 평가한다. 여기서 단어의 중요도는 상호정보량(Mutual Information)척도를 이용하여 군집 정보량을 고려하여 계산한다. 따라서 본 제안 기법에 의해 단어의 중요도는 유사한 문서들이 속해있는 군집 내에서 평가될 수 있으며, 이에 따라 전체 문서 집합에서 상대적으로 작은 비율을 가지는 문서에 대해 키워드 추출 정확도를 높일 수 있다.

2. 배경지식

본 절은 그래프에 기반을 둔 키워드 추출 방법인 TextRank 알고리즘과 클래스 정보를 고려한 키워드 추출 기법에 사용되는 상호정보량 척도에 대해 기술한다.

2.1. TextRank

TextRank 알고리즘은 Google에서 제안한 PageRank 알고리즘[2]을 자연어 처리에 응용한 그래프 기반 키워드 랭킹 알고리즘으로서, 특정 키워드 V_i에 대한 중요도가 식 (1)과 같이 표현된다.

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

여기서 V_i는 i번째 노드를 의미하며, $TR(V_i)$는 i번째 노드의 중요도인 TextRank score를 뜻한다. $In(V_i)$는 V_i를 가리키는 이웃한 노드들의 집합을 의미하며, $Out(V_j)$는 V_j가 가리키는 이웃한 노드들의 집합을 의미한다. w_{ji}는 V_j와 V_i를 연결하는 간선의 가중치(weight)이며, d는 사용자가 특정 페이지를 클릭할 확률(damping factor)를 나타낸다. 여기서 d 값의 범위는 0과 1사이 이며, 일반적으로 0.85로 둔다[3].

2.2. 상호정보량(Mutual Information)

상호정보량은 확률론에 기반을 두어 항목 상호간에 연관되는 정보량을 계산하는 방법으로서 두 단어 x,y간의 상호 정보량은 아래와 같이 표현한다.

$$I(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

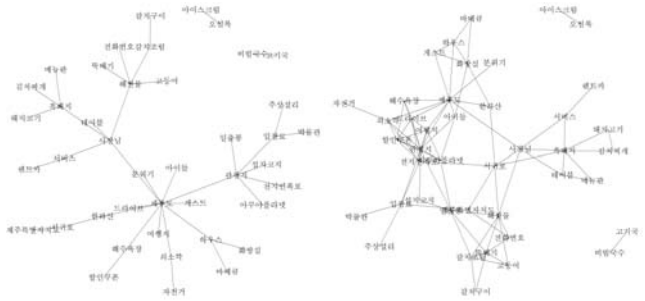
$$MI(x,y) = \sum_{x \in X,\ y \in Y} p(x,y) I(x,y) \quad (2)$$

이 상호정보량은 단어 x, y가 동일한 문서 내에서 출현할 확률에 비해 각 단어가 문서에서 독립적으로 출현할 확률이 얼마나 큰지를 표현한 것이다.

3. ClusterTextRank

본 절에서는 군집 기반에서 TextRank 알고리즘을 변형한 키워드 추출 기법인 ClusterTextRank를 제안한다. 2장에서 설명한 식 (1)에 따르면, TextRank에서 노드 V_i 의 중요도를 계산하기 위해 먼저 가중치 w를 정의해야한다. 기본적으로 TextRank 알고리즘에서 가중치 w는 두 단어의 co-occurrence를 이용하여 산출된다[2]. 그 다음 각 노드의 중요도를 임의의 초기값으로 설정한 후 0에 수렴하거나 임계값에 도달할 때까지 식 (1)을 반복적으로 수행한다. 그러나 이와 같이 군집 기반에서 TextRank 알고리즘을 수행할 때 가중치 w를 단어 간의 co-occurrence로 정의하면, 특정 군집에서 중요하다고 평가된 단어가 다른 군집에서 많이 출현한 보편적인 단어일 가능성이 높을 문제가 있다. 따라서 군집 기반에서 가중치 w를 정의할 때, 특정 군집에 편향된 co-occurrence로 정의한다면 더 효율적으로 군집 기반에서 키워드 추출을 수행할 수 있다. 또한, 단어 간의 co-occurrence를 이용하여 그래프를 생성하면 단어 쌍의 수만큼 간선을 형성하게 되므로 매우 복잡한 형태의 그래프를 형성하게 된다. 이러한 특성으로 노이즈와 계산량이 증가하는데 이를 간략화할 필요가 있다.

본 논문은 그래프를 간략하게 하기 위해 최소신장트리(Minimal Spanning Tree)를 이용한다. 일반적으로 그래프의 크기를 축소하는 방법은 두 단어를 연결하는 선의 가중치 w에 임계값을 적용하여 임계값 미만의 선을 제거한다[7]. 그러나 이와 같은 방법은 그래프에서 연결되지 않은 고립된 노드가 생길 수 있다. 따라서 본 논문은 최소의 비용으로 모든 노드를 연결하는 최소신장트리를 이용하여 상대적으로 다른 노드와의 co-occurrence가 낮은 노드도 그래프에 포함될 수 있도록 한다. (그림 1)은 최소신장트리를 적용한 방법과 적용하지 않은 방법에서 그래프의 차이를 보여준다. 최소신장트리를 적용한 그래프에 비해 적용하지 않은 방법은 노드들 간의 연결이 복잡하게 형성되어 있는 것을 볼 수 있다. 본 논문은 최소신장트리를 이용한 방법을 제안하여 키워드 추출의 정확도를 높인다.



(그림 1) $co-occurrence_{MST}$(좌)와 $co-occurrence$(우) 샘플

다음으로 단어 그래프에서 간선의 가중치 w를 정의할 때 상호 정보량을 이용하여 군집 정보를 고려하는 방법에 대해서 설명한다. 2장에서 설명한 식 (2)는 두 변수 x와 y 간의 상호정보량을 고려하고 있으므로, 두 단어와 군집 정보량을 함께 고려해야하는 본 논문에서는 적합하지 않다. 따라서 두 단어의 관계와 군집 간의 상호정보량을 계산하기 위해 두 변수 쌍 (x_1,x_2)와 변수 y와의 정보량을 계산하는 다변량 상호 정보량(Multivariate Mutual Information)을 사용한다[8]. 이는 구체적으로 식 (3)과 같이 정의한다.

$$I(S, y) = \log \frac{p(x_1,x_2,y)}{p(x_1,x_2)p(y)} \ (S = \{x_1,x_2\})$$

$$MI(S, y) = \sum_{y \in Y,\ x_1 \in X_1,\ x_2 \in X_2} p(x_1,x_2,y) I(S,y) \quad (3)$$

식 (3)에서 $I(S, y)$는 두 변수 쌍 (x_1,x_2)가 변수 y에 종속적이면 값이 증가하고 독립이면 0에 가까워진다. 본 논문은 그래프 상에 두 단어와 군집 간에 상호 정보량을 계산하므로 두 변수 쌍 (x_1, x_2)를 두 단어 쌍 (t_1,t_2)로 두고, 변수 y를 군집 c_p로 둔다. 그래프 상에 두 단어 (t_1,t_2)와 군집 c_p와의 상호 정보량을 계산하기 위해 식 (3)을 아래의 식과 같이 정의한다.

$$I(S,c_p) = \log \frac{p(t_1,t_2,c_p)}{p(t_1,t_2)p(c_p)} \ (S = \{t_1,t_2\})$$

$$MI(S,c_p) = \sum_{p=1}^{L} p(c_p) I(S,c_p) \quad (4)$$

여기서 c_p는 p번째 군집을 의미하며, 두 단어 t_1과 t_2가 동일한 군집 내에서 출현할 확률과 특정 군집 내에서 출현할 확률로 표현된다. 따라서 식 (4)가 양수 값을 가지면 두 단어가 특정 군집에 종속성이 높다는 것을 의미하고 0에 가까우면 군집에 독립적이라는 것을 의미한다. 이를 기존 TextRank 기법의 아이디어에 적용하면 아래와 같이 특정

노드의 중요도 값을 평가할 수 있다.

$$TR(V_i) = (1-d) + d \sum_{Vj \in \in (V_i)} \frac{MI(S_{ji}, c_p)}{\sum_{V_k \in Out(V_j)} MI(S_{jk}, c_p)} TR(V_j) \quad (S_{ij} = \{V_i, V_j\}) \quad (5)$$

식 (5)에 따르면, i번째 노드 V_i의 중요도는 해당 노드를 가리키는 j번째 노드의 중요도와 상호정보량 그리고 j번째 노드가 가리키는 k번째 노드들과의 상호정보량으로 계산되어진다. 따라서 기존 TextRank 알고리즘에서 두 단어를 연결하는 선의 가중치를 계산할 때 일반적으로 사용하는 co-occurrence 대신에 상호정보량 척도를 사용함으로써, 군집 정보량을 고려하여 가중치를 계산할 수 있다.

위 ClusterTextRank를 활용한 키워드 추출 절차는 아래와 같다.

1. k-means 알고리즘을 이용하여 전체 문서집합을 m 개의 군집으로 군집화
2. 각 군집별로 단어 간의 co-occurrence를 이용하여 최소신장트리 그래프 생성
3. m번째 군집에 속하는 모든 노드의 초기 TextRank score 값을 1로 초기화
   $TR(V_{j,i}) = 1 \ (i = 1,2, \ldots, n)(j = 1,2, \ldots, m)$
4. 식 (5)를 이용하여 TextRank값을 계산
5. 단계 (4)를 TextRank score 값이 수렴하거나 임계값(0.0001)에 도달할 때까지 반복
6. 군집별로 전체 노드에 대한 TextRank score 값이 계산되면, 각 군집에 속한 노드들과 TextRank score 값들을 하나의 테이블로 통합
7. TextRank score 값 기준으로 내림차순 정렬하여 본 실험에서 지정한 개수만큼 키워드 선택

4. 실험 및 결과

4.1. 실험 방법 및 환경

실험에 사용된 문서집합은 2015년 1월부터 2015년 12월까지 네이버(http://www.naver.com)에 등록된 여행 관련 블로그 267,730건으로 선정하였다. 여기서 문서집합에는 국내 여행관련 블로그뿐만 아니라 해외, 여행사, 여행상품에 관한 글을 포함하고 있으므로 ('여행','맛집','펜션','한국지명')을 포함한 문서들만 필터링하여 총 37,985건의 문서들을 실험 대상으로 하였다. 다음으로 k-means 알고리즘을 이용하여 전체 문서집합을 군집화하기 위해 엘보우 방법(Elbow method)을 적용하여 군집의 수 k 값을 결정하였다. 엘보우 방법은 임의로 설정한 k 값 범위 내에서 순차적으로 군집을 생성하고 군집 비용을 계산하여 군집의 수 k 값을 결정하는 방법이다. 이를 통해 k 값을 15로 설정하여 k-means 알고리즘을 통해 군집화를 수행하고, 각 군집별로 co-occurrence를 이용하여 그래프를 생성하였다. 여기서 그래프를 생성할 때, 단어 간의 co-occurrence가 0.10 미만인 단어들은 모두 제외하였다.

다음으로 각 군집에서 생성한 그래프를 이용하여 군집별로 TextRank 알고리즘을 수행한다. ClusterTextRank는 본 논문에서 제안한 최소신장트리와 상호정보량 척도를 이용한 방법이고, co-occurrence$_{MI}$는 co-occurrence를 이용하여 생성한 그래프와 상호정보량 척도만 이용한 방법을 의미한다. baseline은 기존 TextRank에서 일반적으로 사용되는 co-occurrence를 이용한 방법을 의미한다. 이를 표로 정리하면 <표 1>과 같다.

<표 1> 각 방법별 간선 가중치

method	w
ClusterTextRank	Mutual Information
co-occurrence$_{MI}$	Mutual Information
baseline	co-occurrence

<표 2>는 각 방법별로 단어들을 TextRank score 값 기준으로 순위를 매긴 결과에서 상위 15개의 단어들을 보여준다. <표 2>에서 보는바와 같이 최소신장트리를 이용한 ClusterTextRank방법과 최소신장트리를 이용하지 않은 co-occurrence$_{MI}$방법에서 선택된 후보 키워드들은 서로 다르다. 이는 TextRank 알고리즘을 이용하여 단어의 중요도를 계산할 때 동일한 단어라도 이웃한 단어들의 집합이 상이하여 다르게 측정되어지기 때문이다. 또한 '화이트', '횡단보도', '체크인'과 같은 불용어 수준에 가까운 단어들이 상위에 랭크되어 있음을 확인할 수 있다.

<표 2> 각 방법별 Top 15 후보키워드

rank	Cluster TextRank	co-occurrence$_{MI}$	baseline
1	호스텔	휴게실	화장실
2	해운대	화장실	화장품
3	화장실	호스텔	제주도
4	휴게실	휴게소	흑돼지
5	홈페이지	휴지통	하이킹
6	칼국수	해운대	해운대
7	토요일	홍합탕	일주문
8	화이트	횡단보도	체크인
9	휴양지	홈페이지	한라산
10	순천만	휴가철	홈페이지
11	홍합탕	해변가	케이블카
12	횡단보도	후라이팬	파충류
13	하와이	휴양지	호스텔
14	이벤트	황급연휴	임진왜란
15	사이트	해수욕장	징검다리

4.2. 성능평가

본 논문에서 제안한 ClusterTextRank 키워드 추출 성능을 평가하기 위해 DCG(Discounted Cumulative Gain) 척도를 이용하였다[4]. DCG는 웹 검색엔진 알고리즘의 효율성을 측정하는 도구로 많이 쓰이는 방법 중 하나이며, 실제 검색된 문서들과 검색어와 관련된 문서들 간의 관련도(relevance)를 계산하는 방식이다. 이러한 방식에는 먼저 p개의 검색된 문서들의 관련도를 순위정보를 고려하지 않

고 모두 더하여 계산하는 CG(Cumulative Gain) 방식이 있으며 식은 아래와 같다.

$$CG_p = \sum_{i=1}^{p} rel_i \quad (6)$$

여기서 rel_i은 실제 검색된 문서들 중에서 i번째 문서의 관련도를 나타낸다. 다른 방식으로는 실제 검색된 문서들의 순위정보(rank)를 이용하여 -상위에 랭크된 문서들은 하위에 랭크된 문서보다 중요하다.- 페널티(penalty)를 부여한 후 정규화(normalized)한 nDCG(Normalized Discounted Cumulative Gain) 방식이 있으며 식은 아래와 같다.

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i} \quad (7)$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (8)$$

여기서 식 (7)은 실제 검색된 문서들의 관련도를 정량화할 때 단순히 더하는 것이 아니라 순위정보에 따라 페널티를 부여하여 계산한다. IDCG는 관련도를 내림차순으로 정렬하여 DCG를 계산한 값이다. 따라서, 식 (8)은 DCG를 0과 1사이로 정규화한 방식이며 1에 가까울수록 좋은 검색 결과임을 의미한다.

본 논문에서는 nDCG를 계산하기 위해 관련도는 2진 관련도(binary relevance) $rel_i = \{0,1\}$를 이용하였다. 2진 관련도는 추출한 키워드 - 본 논문은 문서들의 검색 성능을 평가하는 것이 아니라 단어들의 키워드 추출 성능을 평가하는 것이 주목적이므로 문서 대신 단어를 사용한다. - 가 비교 대상에 존재하면 1을 부여하고 존재하지 않으면 0을 부여하는 방법이다[6]. 비교 대상이 되는 단어들은 연구원들이 평가한 키워드들을 활용하였으며, 이 중 상위 10개의 단어들은 다음과 같다.

호스텔	센트럴파크	순천만	제주도	조계종
주차장	천왕문	첨성대	태종대	특산물

각 방법별 키워드 추출 성능 평가를 위해 상위 T개의 결과에 대해서 nDCG 값을 측정한다. [3]에서 T 값의 범위를 5에서 20 사이로 권장하지만, 본 논문은 T = (10, 20, 30) 범위 안에서 수행한다.

<표 4> 각 방법별 성능평가 결과

$nDCG_p$	Cluster TextRank	co-occurrence$_{MI}$	baseline
$nDCG_{10}$	**0.65051**	0.63093	0.63093
$nDCG_{20}$	**0.77845**	0.53380	0.54001
$nDCG_{30}$	**0.81605**	0.71647	0.67287

<표 4>는 각 방법별 Top T 개의 nDCG 결과 값들을 보여준다. 표에서 보는 바와 같이 $nDCG_{10}$ 에서는 본 논문에서 제안한 ClusterTextRank 방법이 0.65051로 다른 방법보다 근소하게 향상된 약 2% 높은 결과를 보여준다. 이에 반해 $nDCG_{20}$은 0.77845로 약 24% 향상된 결과를 보여주며, $nDCG_{30}$은 0.81605로 약 12% 향상된 결과를 보여주고 있다. 결과적으로 본 논문에서 제안한 방법이 키워드 수에 따라 약 2%~24% 정도 높은 우수한 성능을 보여주고 있음을 확인할 수 있다. 이는 전체 문서집합에서 유사한 문서들이 동일한 집단에 속하도록 군집화하고 군집 정보를 고려하여 단어 간의 가중치를 계산함으로써, 기존 방법보다 더욱 세밀 하게 단어의 중요도를 평가할 수 있기 때문이다.

5. 결론

본 논문은 기존의 TextRank 알고리즘에 상호정보량 척도를 결합하여 군집 기반에서 키워드 추출하는 방법을 제안하였다. 본 논문은 각 군집에 포함된 단어들을 최소신장트리 그래프로 표현하여 그래프를 간략화 했으며, 이에 근거한 군집 정보량을 상호정보량 척도를 이용하여 산출하고 이를 통해 단어의 중요도를 평가하였다. 실험 결과에 따르면, 본 논문에서 제안한 방법이 기존 방법보다 평균적으로 약 13.4% 향상된 결과를 보여준다. 따라서 본 논문에서 제안한 방법이 기존 방법에 비해 더 우수한 성능을 보여주고 있음을 알 수 있다.

6. 감사의 글

이 논문은 2015년도 한국연구재단의 개인기초연구사업 (NRF-2015R1D1A1A09061299) 지원으로 이루어졌음.

참고문헌

[1] Hasan, Kazi Saidul, and Vincent Ng. "Automatic Keyphrase Extraction: A Survey of the State of the Art." *ACL (1)*. 2014.

[2] Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the web." (1999).

[3] Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into texts." Association for Computational Linguistics, 2004.

[4] Discounted cumulative gain, https://en.wikipedia.org/wiki/Discounted_cumulative_gain

[5] Minimum spanning tree, https://en.wikipedia.org/wiki/Minimum_spanning_tree

[6] Wang, Yining, et al. "A theoretical analysis of NDCG ranking measures." Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013). 2013.

[7] Y. J. Lee, S. D. Kim, S. H. Kang and H. G. Cho . "Characteristics Analysis on Keyword Network Obtained from Twitter." 한국정보과학회 학술발표 논문집, (2015. 06)

[8] Timme, Nicholas, et al. "Multivariate information measures: an experimentalist's perspective." arXiv preprint arXiv:1111.6857 (2011).