김일성종합대학학보
정보과학

주체108(2019)년
제65권 제4호

실마리어추출을 위한 개선된 TextRank알고리듬

김예화, 정만흥

인터네트의 급속한 발전으로 망상의 자료들이 폭발적으로 늘어나고있으며 이러한 자료들을 어떻게 자동적으로 처리하여 리용하겠는가 하는 문제는 매우 중요하게 제기되고 있다.

실마리어는 분문분류, 본문무리짓기를 비롯한 본문처리, 질문응답체계와 다중문서요약을 비롯한 정보검색체계들에서 유용하게 리용되며 실마리어를 추출하기 위한 많은 방법들이 제기되였다.

실마리어추출알고리듬[1]은 크게 감독학습 혹은 무감독학습방법으로 나누어볼수 있다.

감독학습방법에 기초한 실마리어추출알고리듬에서는 단어출현빈도와 TF-IDF, *n*-그람, 단어길이, 위치정보, 공기빈도수와 같은 특징들을 효과적으로 리용하였다.

무감독학습방법에 기초한 실마리어추출알고리듬에서는 문맥을 고려하고 실마리어들의 순위를 매기는데 point wise KL분리를 리용하였다.

선행연구[1]에서는 범위, 위치정보와 단어출현빈도의 영향을 TextRank모형에 도입하였다. 전통적인 TextRank모형에서는 단어그라프를 리용하는데 단어들사이의 의미를 고려하지 않고 평균적으로 처리하였다.

우리는 다중문서요약이나 정보검색에서 중요한 실마리어를 추출하기 위하여 단어들의 영향득점을 고려하여 TextRank알고리듬을 개선하였으며 그 성능을 평가하였다.

1. 실마리어추출

전통적인 TextRank알고리듬은 그라프에 토대한 알고리듬으로서 여기서 단어들은 그라프에서의 마디점들과 같고 단어들사이의 관계는 그라프의 릉과 같으며 단어들은 고정된 크기의 창문에서 단어들의 출현번호로 나타난다.

PageRank알고리듬[1]은 웨브페지들사이에 이어지는 웨브페지의 가능성이 우연적이고 동등하다는 원리에 기초하여 페지의 득점을 얻는다. 그러나 대다수 본문망들에서 단어들사이에는 관계가 있으며 이로부터 단어들사이의 관계의 세기를 고찰할 필요가 있게 된다.

전통적인 TextRank모형[1]은 무방향그라프로 표현한다.

$$G = (V, E)$$

여기서 V는 점모임, E는 릉모임($V \times V$ 의 부분모임)이다.

w_{ij} : 임의의 두점 v_i, v_j의 릉의 무게

$In(v_i)$: 점 v_i를 가리키는 점들의 모임

$Out(v_i)$: 점 v_i가 가리키는 점들의 모임

$TR(v_i)$: TextRank모형에 의해 얻어진 마디 v_i의 득점

$$TR(v_i) = (1-d) + d \sum_{v_i \in In(v_j)} \frac{w_{ji}}{\sum_{v_k \in Out(v_j)} w_{jk}} TR(v_j) \quad (1)$$

d는 0~1의 값으로서 PageRank에서의 d와 같다.

TextRank알고리듬에서 그라프의 매 마디점에 대한 득점계산은 그라프의 매 점이 수렴할 때까지 그라프상에서 반복된다.

실마리어추출에서 단위는 단어들의 모임이다. 매 단위에 대하여 최종득점값이 얻어지면 실마리어로서 가장 높은 득점값을 가지는 웃준위 m개 단위가 선택된다.

그러나 TextRank는 영향－평형－전달단어그라프만을 작성한다.

문서의 단어그라프를 그림에 보여주었다. TextRank에서 3개의 마디가 단어 b와 련결되여있기때문에 단어 b가 단어 a에 기여하는 영향은 b의 1/3이다. 그러나 련결되여있는 단어들사이의 의미적관계는 모두 같지 않을수 있다.

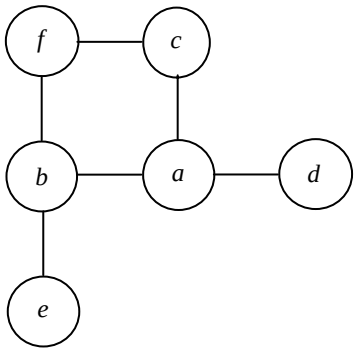

그림. 문서의 단어그라프

이로부터 우리는 단어들사이의 의미적관계를 고려하여 단어그라프에서 단어들의 득점을 평가하기 위한 방법을 제안하였다.

선행연구[2]에서는 word2vec의 단어들사이의 의미적관계를 평가하기 위한 방법으로서 대단히 큰 자료모임에 대한 련속벡토르표현을 제기하였다. 여기서 벡토르들은 의미류사성을 계산하는데 리용한다.

$W = \{w_0, w_1, \cdots, w_{N-1}\}$을 문서 d의 단어모임이라고 하자.

이때 W'는 단어모임 W에서 명사, 동사들로만 된 단어모임, p는 W'의 단어개수라고 하자.

W의 매 단어를 word2vec를 가지고 코퍼스에서 훈련한 후 K차원벡토르로 표현한다. 즉 단어 a의 벡토르표현은

$$vec_a = [v_0, v_1, \cdots, v_{k-1}],\ 0 \le k < N$$

이다.

단어의 벡토르표현으로부터 단어 a와 b의 의미류사도를 코시누스류사도로 계산한다.

$$sem_{ab} = \frac{\sum_{k=0}^{k-1} vec_a[k] \times vec_b[k]}{\sqrt{\sum_{k=0}^{k-1} vec_a[k]^2} \sqrt{\sum_{k=0}^{k-1} vec_b[k]^2}} \quad (2)$$

단어 b가 단어 a에 주는 영향전파식은 의미류사도에 의해 다음과 같이 정의한다.

$$S(b,\ a) = \frac{sem_{ab}}{\sum_{c \in Out(b)} sem_{cb}} S(b) \quad (3)$$

여기서 $Out(b)$는 단어그라프에서 단어 b가 지적하는 단어들의 색인모임으로서 TextRank와 같다. 그리고 $S(b)$는 단어 b의 영향득점이고 이외에 단어는 린접한 높은 출현빈도를 가지

는 단어들에 더 많은 영향을 전파한다고 가정한다.

이로부터 영향전파식 (2)에 tf-idf를 도입한다.

최종적인 영향전파식은 다음과 같다.

$$S(b,\ a)=\frac{sem_{ab}}{\sum_{c\in Out(b)} sem_{cb}\times tfidf}S(b) \tag{4}$$

단어 a의 영향득점은 다음과 같이 계산한다.

$$Score(a)=(1-d)+d\times\sum_{b\in In(a)} S(b,\ a) \tag{5}$$

여기서 $In(a)$는 단어 a를 지적하는 단어들의 색인모임이다.

실마리어를 추출하는 과정은 다음과 같다.

먼저 주어진 문서로부터 단어분리를 진행한다.

분리된 단어들에 품사를 부여한다. 이때의 단어모임을 W로 한다.

품사부여된 단어모임으로부터 명사, 동사의 품사를 가진 단어이외의 기타 단어들을 W에서 제거하여 W'를 생성한다.

W'에 대한 단어그라프 G를 작성한다.

높은 출현빈도를 가지는 린접한 단어들에 더 큰 영향을 주도록 식 (5)를 반복실행한다.

그라프에 들어있는 매 단어에 대하여 최종득점값이 계산되면 단어들을 득점값에 따라 거꾸로 정렬한다.

순위화목록에서 웃준위의 m개 단어들을 실마리어로 등록한다. m은 추출하려는 실마리어의 개수이다.

2. 실마리어추출에 대한 실험결과

명사, 동사, 사용자사전단어들을 제외하고 다른 단어들을 려과한다.

TextRank의 d값은 0.85로 설정하고 창문크기는 10으로 설정하였다. TR에서는 영향전파식으로 식 (1)을 리용하며 TR_score에서는 식 (5)를 리용한다.

우리가 사용한 평가척도로는 적중률, 완전률, F-척도이다.

전통적인 TextRank와 제안한 방법을 비교하였다.

실험은 3개의 실마리어가 추출된 경우와 비교하였다.

적중률, 완전률, F-척도는 다음과 같이 계산된다.

적중률: $P=\dfrac{|S_s\cap S_h|}{|S_s|}$

완전률: $R=\dfrac{|S_s\cap S_h|}{|S_h|}$

F-척도: $F=\dfrac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}$

여기서 S_h는 정답문장모임, S_s는 체계가 출력한 대답문장모임이며 β는 적중률과 완전

률의 중요도를 조절하는 상수이다.($\beta = 2$)

$m = 3$(실마리어개수가 3)일 때 실마리어추출알고리듬의 성능을 비교하였다.(표)

표. 실마리어추출알고리듬의 성능비교

실험방법	*P*	*R*	*F*
TR	0.407	0.290	0.307 69
TR_score	0.510	0.349	0.372 52

표에서 보여주는바와 같이 제안한 방법이 전통적인 TextRank알고리듬에 비하여 우월하다는것을 알수 있다.

맺 는 말

실마리어추출을 위하여 단어그라프에서 단어들사이의 의미관계를 고려하여 단어의 득점계산식을 제안하고 그에 기초하여 TextRank알고리듬을 개선하였으며 제안된 알고리듬의 효과성을 검증하였다.

참 고 문 헌

[1] X. Tian; New Technology of Library and Information Service, 9, 30, 2013.

[2] T. Mikolov et al.; Advances in Neural Information Processing Systems, 26, 3111, 2013.

주체108(2019)년 8월 5일 원고접수

Advanced TextRank Algorithm for Keywords Extraction

Kim Ye Hwa, Jong Man Hung

In this paper, we improved the word score calculation in TextRank algorithm.

Considering the relation between the words in the document, we proposed the semantic similarity and the effect propagation formula between the words and so using it, we improved the score calculation of each word.

Key words: TextRank algorithm, word graph