김일성종합대학학보 주체108(2019)년
정보과학 제65권 제2호

다중문서요약을 위한 문장순위화의 한가지 방법

정만흥, 김예화

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《프로그람을 개발하는데서 기본은 우리 식의 프로그람을 개발하는것입니다. 우리는 우리 식의 프로그람을 개발하는 방향으로 나가야 합니다.》(《**김정일**선집》 증보판 제21권 42페지)

다중문서요약[2]은 두가지 과제수행을 통해 실현된다. 그중 하나는 다중문서모임에서 요약문장들을 추출하는 과제이며 다른 하나는 추출된 요약문장들을 순위화하는 과제이다.

선행연구[1, 3]에서는 문장순위화를 위한 bottom-up방법을 제기하였다.

론문에서는 문장순위화를 진행하기 위해 문장류사도를 무리짓기의 기준으로 정하고 요약된 문장들에 대한 무리짓기를 한 다음 요약문장의 순위화를 진행하는 한가지 방법을 제기하였다.

1. 무리짓기에 의한 문장순위화방법

무리짓기를 실현하기 위해 문장들을 단어벡토르로 표현한다.

문장의 단어벡토르표현을 실현하는데서 부족점은 문장내용에 기여하는 매개 내용단어들을 동등하게 고찰하는데 있다.

그러나 실제적으로 내용단어들의 기여정도는 서로 다르다. 실례로 실체내용단어들을 들수 있다. 특히 새 소식을 알리는 기사본문들에서 실체내용단어는 다른 개념단어들보다 중요하다.

실체내용단어벡토르로 문장을 표현하기 위해 두 종류의 실체내용단어(일반실체와 고유실체)를 생각한다. 여기서 일반실체는 실체를 표현하는 내용단어이며 고유실체는 이름을 가진 단어들이다. 사람, 기관 및 기업소, 장소 및 위치, 수량 등을 나타내는 단어들은 고유실체이다.

일반적으로 일반실체와 고유실체는 동등하다. 그러나 문장들의 린접관계를 나타낼 때 고유실체는 일반실체에 비해 보다 중요한 역할을 하게 된다.

이와 같은 해석에 기초하여 문장 S_i를 다음과 같은 벡토르로 표현한다.

$$S_i = (w_1 \times f(e_{i1}),\ w_2 \times f(e_{i2}), \ldots, w_m \times f(e_{im}))$$

여기서 $f(e_{ik})$는 실체내용단어 e_{ik}의 출현여부(1 혹은 0), w_k는 단어 e_{ik}에 부여되는 무게, m은 실체단어의 개수이다.

무게 w_k를 다음과 같이 계산한다. e_{ik}가 일반실체이면 $w_k = 1$, 고유실체이면 $w_k = 2$이다.

새 기사를 비롯하여 많은 본문들에서 고유실체의 의미가 중요하기때문에 그것의 무게를 일반실체무게의 2배로 한다.

요약된 문장들의 무리짓기에 기초한 무리들의 순위화와 무리들에서의 개별적문장들

의 순위화단계를 거쳐 문장순위화를 진행한다.

1) 요약된 문장들에 대한 무리짓기

문장무리짓기를 위해 개선된 *k*-평균무리짓기알고리듬을 리용하였다.

*i*번째 클라스를 C_i라고 할 때 그것의 류사도를 $Sim(C_i)$로 표시하고 다음과 같이 정의한다.

$$Sim(C_i) = \min_{s_m,\ s_n \in C_i} \{Sim(S_m,\ S_n)\}$$

여기서 $Sim(S_m,\ S_n)$은 문장 S_m과 S_n사이의 코시누스류사도이다.

이때 개선된 *k*-평균무리짓기알고리듬은 다음과 같다.

모든 요약문장을 포함하는 클라스 C_1을 생각한다.

클라스 C_i에 대하여 $\min(Sim(C_i)) < T_0$(턱값)이면 다음과 같이 반복한다.

① $Sim(S_m, S_n) = \min(Sim(C_i))$인 S_m과 S_n을 찾고 이것을 새로운 2개 무리의 중심으로 한다.

② 새로 창조된 중심을 가지고 일반*k*-평균무리짓기를 수행한다.

우의 알고리듬은 매개 클라스에 속하는 문장쌍들의 류사성이 턱값 T_0보다 크거나 1개의 문장만이 속하는 경우에 중지된다.

2) 순위화알고리듬

요약문장들에 대한 무리짓기를 수행한 후 무리짓기에 기초하여 다음과 같이 문장들의 순위화를 진행한다.

(1) 무리수준의 순위화

무리수준의 순위화는 무리짓기의 결과로 얻어진 무리들사이의 순위화로서 대역적특징을 가진다. 무리수준의 순위화를 대역적순위화라고 부른다.

알고리듬은 다음과 같다.

① 첫번째 순위무리 G_1을 선택한다.

동일한 문서에 속하는 요약문장들에 문서자체에서의 문장순위에 따라 순위번호를 결정한다.

매개 무리에서 동일한 순위번호를 가진 문장들의 개수를 계수한다.

순위번호가 1인 문장개수가 제일 큰 클라스를 첫번째 순위무리로 결정한다. 개수가 같은 경우 순위번호 2를 론의한다.

② *i*번째 순위의 무리 G_i를 선택한다.

$i-1$개의 무리들이 $G_1, G_2, \cdots, G_{i-1}$과 같이 순위화되였다고 할 때 이미 순위화된 무리들과의 류사성이 최대로 되는 무리를 *i*번째 무리로 결정한다.

$$G_i = \arg\max_{G} \sum_{j=1}^{i-1} Sim(G_j,\ G)(i > 1)$$

여기서 G는 순위화되지 않은 무리이다.

(2) 문장수준의 순위화

문장수준의 순위화 역시 무리수준의 순위화와 같은 원리에 따라 진행한다. 문장수준의 순서는 국부적특징을 반영하며 결과 문장수준의 순위화를 국부적순위화라고 부른다.

알고리듬은 다음과 같다.

i=1, 2, …, k에 대하여 다음의 단계를 수행한다.(여기서 k는 요약문장들을 무리짓기하였을 때의 무리의 개수.)

① i번째 무리 G_i에 속하는 문장들이 모두 동일한 문서내의 문장들이라면 해당 문서에서의 본문문장순위에 따라 무리 G_i안의 문장들을 순위화한다.

② i번째 무리 G_i에 속하는 문장들이 여러 문서들에 분산되여있으면 순위화를 다음과 같이 진행한다.

ㄱ) i=1번째 무리에서의 첫번째 순위문장 S_{11}을 선택한다.

기타 다른 모든 문장들과의 류사성이 최대로 되는 문장을 S_{11}로 결정한다.

$$S_{11} = \arg\max_{S \in G_1} \sum_{S' \in G_1, S' \neq S} Sin(S,\ S')$$

ㄴ) $i \neq 1$번째 무리에서의 첫번째 순위문장 S_{i1}을 선택한다.

첫번째 무리 G_1에 속하는 모든 문장들과의 류사성이 최대로 되는 문장을 S_{i1}로 결정한다.

$$S_{i1} = \arg\max_{S \in G_i} \sum_{S' \in G_1} Sin(S,\ S')$$

여기서 $Sim(S,\ S')$는 문장 S와 S'사이의 코시누스류사도이다.

ㄷ) p번째 문장 S_{ip}를 선택한다.

앞서 $p-1$개의 문장들이 $S_1, S_2, \cdots, S_{p-1}$과 같이 순위화되였다고 할 때 이미 순위화된 문장들과의 류사성이 최대로 되는 문장을 찾고 그 문장을 순위화의 p번째 문장으로 결정한다.

$$S_{ip} = \arg\max_{S} \sum_{j=1}^{p-1} Sim(S_{ij},\ S)(p > 1)$$

여기서 S는 i번째 무리 G_i에 속하는 문장으로서 아직 순위화되지 않은 문장이다.

결국 요약문장들의 순위화과정을 종합하면 다음과 같다.

첫째로, 개선된 k-평균법에 의해 얻어진 요약문장들을 무리짓기하여 k개의 요약문장들의 무리를 얻는다.

둘째로, 개선된 k-평균법에 의해 얻어진 문장들의 무리를 무리순위화알고리듬(대역적순위화알고리듬)을 리용하여 k개의 무리들로 순위화한다. 즉 $G_1, G_2, \cdots, G_k$이다.

셋째로, 문장순위화알고리듬(국부적순위화알고리듬)을 리용하여 무리내에서의 문장들을 순위화한다.

이때 순위화된 문장렬은 다음과 같다.

$$S_{11}, S_{12}, \cdots, S_{1p1}, S_{21}, S_{22}, \cdots, S_{2p2,} \cdots, S_{k1}, S_{k2}, \cdots, S_{kpk}$$

여기서 p_i는 i번째 무리에 속하는 문장의 개수이다.

2. 실험결과 및 분석

순위화의 효과성평가에서 일반적으로 쓰이고있는 두가지 거리척도에 준하여 선행방법[2]과 비교분석하였다. 거리척도에는 τ 거리척도와 AC 거리척도가 있다.

1) τ 거리척도

τ 거리척도를 Kendall의 척도라고도 부르며 다음과 같이 계산한다.

$$\tau = 1 - \frac{4m}{N(N-1)}$$

여기서 N 은 요약문장의 개수, m 은 순위화된 요약문장렬을 참고순위의 문장렬로 변환하기 위해 린접한 문장들끼리 순서를 바꾸는 총회수이다.(참고순위란 정답순위를 의미한다.)

τ 의 값은 -1부터 1까지 변한다. 여기서 1은 $m=0$ 인 경우로서 요약문장들의 순서와 참고문장들의 순서가 일치하는 경우이다. -1은 요약문장들의 순서와 참고문장들의 순서가 완전히 거꿀순서인 경우로서 최대로 나쁜 경우이다. 그러므로 우연적인 순서는 보통 $\tau=0$ 인 경우이다.

τ 거리척도의 의미를 명백히 하기 위한 실례를 표 1에 보여주었다.

표 1. τ 거리척도의 의미

요약문장들의 순서렬	참고문장들의 순서렬	N	m	τ 의 값
1 2 4 3	1 2 3 4	4	1	0.67
1 5 2 3 4	1 2 3 4 5	5	3	0.40
2 1 3	1 2 3	3	1	0.33

표 1에서 보는바와 같이 m 이 같다고 하더라도 N 이 클수록 τ 거리척도가 크다는것을 알수 있다.

2) AC 거리척도

AC (Average Continuity)거리척도는 평균련속성거리척도이다. 이 거리척도의 의미는 순위화의 정확도가 정확하게 순서화된 련속적인 문장들의 개수에 의해 평가된다는데 있다.

AC 거리의 계산식은 다음과 같다.

$$\mathrm{AC} = \exp\left(1/(k-1)\sum_{n=2}^{k}\log(P_n+\alpha)\right)$$

여기서 k 는 정확하게 순서화된 련속적인 문장들의 최대개수, α 는 $P_n=1$ 일 때 로그함수가 값을 가지도록 정의되는 작은 상수값이다.(론문에서는 $\alpha=0.01$ 로 하였다.)

P_n 은 련속문장의 길이 n 의 비률로서 다음과 같이 계산된다.

$$P_n = \frac{m}{N-n+1}$$

순위화의 비교실험을 위해 요약문장의 개수 N 의 각이한 값범위(5～12)에 따르는 평균값을 선택하였다.

순위화의 비교실험결과를 표 2에 보여주었다.

표 2에서 보는것처럼 론문에서 제기한 무리짓기에 토대한 순서화알고리듬은 기준선법에 기초한 방법에 비해 τ 거리척도와 AC 거리척도의 의미에서 효과적이라는것을 알수 있다. 특히 론문에서 제기한 방법은 개개의 단일문서요약의 길이가 크게 차이날 때보다 효과적이다.

표 2. 순위화의 비교실험결과

방법	τ 거리척도	AC거리척도
선행방법[2]	0.647 6	0.440 2
론문의 방법	0.758 2	0.572 1

맺 는 말

개선된 k-평균무리알고리듬에 기초하여 요약문장모임에 대한 무리짓기를 진행한 후 국부적 및 대역적의미의 순위화에 따르는 요약문장의 순위화를 실현하는 방법을 제기하고 방법의 효과성을 검증하였다.

참 고 문 헌

[1] D. Bollegala et al.; Information Processing & Management, 46, 89. 2010.

[2] Feng Jin et al.; Coling, 525, 2010.

[3] L.Peifeng et al.; School of Computer Science and Technology, 215006{pfli, gxdeng, qmzhu}@suda.edu.cn., 2014.

주체108(2019)년 2월 5일 원고접수

A Method of the Sentence Ordering for the Multi-Document Summary

Jong Man Hung, Kim Ye Hwa

In this paper we considered the method performing the ordering for the summary sentences by the local and global ordering method based on the clustering concepts.

The method proposed in this paper is more effective when the length of each single document summary has a big difference.

Key words: document summary, clustering, sentence ordering