김일성종합대학학보
정보과학

주체109(2020)년
제66권 제4호

련관규칙확신도에 의한 련관단어들의 추출과 질문지향단일문서요약에로의 응용

리수정, 정만흥

실마리어에 의한 본문검색에서는 일반적으로 질문단어에 의한 1차검색을 진행하고 검색결과로 얻어지는 문장모임에서 질문단어와 련관되는 단어를 추출하여 질문확장을 진행하는 방법으로 검색의 성능을 높이고있다.

선행연구[1, 2]에서는 단어들사이의 호상정보량과 근접성정도에 따라 질문단어에 대한 련관단어모임을 얻는 방법을 제기하였다. 특히 1차련관단어 및 2차련관단어를 정의하고 그에 토대하여 서술형질문응답을 실현하는 문제를 고찰하였다.

또한 선행연구[3]에서는 질문단어와 기타 단어들의 동시출현빈도수, 단어의 토대득점값 그리고 단어들사이의 최소의존거리에 의해 질문단어와의 련관도득점을 구하고 재귀적으로 전파시키는 방법으로 1차 및 2차련관단어들의 득점값을 계산하는 방법을 고찰하였다.

론문에서는 2－빈발항목모임과 3－빈발부분항목모임우에서 단어들사이에 존재하는 련관규칙에 따라 질문단어에 대한 1차 또는 2차련관단어모임을 추출하는 한가지 방법을 제안하였다.

1. 련관단어추출방법

단어들사이의 련관규칙에 기초하여 질문단어의 련관단어모임을 얻기 위해 표 1과 같은 문장업무표를 작성한다.

표 1. 문장업무표

업무수(문장수)	항목(단어)모임
문장 1	$w_5,\ w_1$
문장 2	$w_3,\ w_2,\ w_6,\ w_1$
문장 3	$w_5,\ w_2,\ w_6,\ w_4$
문장 4	$w_3,\ w_5,\ w_2,\ w_6$
문장 5	$w_3,\ w_5,\ w_2,\ w_4$

표 1은 5개의 문장에 대한 문장업무표로서 이 문장모임에 들어있는 단어전체가 $w_1, w_2, \cdots, w_6$이라는것을 의미한다. 문장목록업무표에서 보는것처럼 단어 w_2, w_3 그리고 w_5들은 거의 모든 문장들에서 출현하는 빈발단어 즉 빈발항목이라는것을 알수 있다. 한편 단어 w_2와 w_5는 3개의 문장들에서 동시출현하는것으로 하여 서로 련관이 있을수 있다는것을 알수 있다.

단어들의 모임을 항목모임이라고 부른다. 특히 1개의 단어로 이루어진 단어모임을 1－항목모임, 일반적으로 k개의 단어쌍들로 이루어지는 단어모임을 k－항목모임이라고 부른다. 빈모임도 고려하여 N개의 단어들에 대한 가능한 항목모임의 총수는 2^N이다.

정의 1 항목모임 I가 속하는 업무의 개수를 업무들의 총수로 나눈 값을 항목모임 I의 지지도라고 부르고 $Support(I)$로 표시한다.

$$Support(I)=\frac{NTCI}{NT}$$

여기서 $NTCI$는 항목모임 I를 포함하는 업무의 개수이며 NT는 업무의 총수이다.

항목모임 I의 지지도 $Support(I)$가 턱값(최소지지도: minimum support)보다 크면 항목모임 I를 빈발항목모임이라고 부른다.

단어 w_i와 w_j사이의 련관규칙을 $w_i \rightarrow w_j$로 표시한다.

련관규칙 $w_i \rightarrow w_j$는 w_i가 문장 d에서 출현하면 w_j도 문장 d에서 출현한다는것을 의미한다.

정의 2 단어 w_i와 w_j들로 이루어진 2－항목모임 $\{w_i, w_j\}$의 지지도를 항목모임 $\{w_i\}$의 지지도로 나눈 값을 련관규칙 $w_i \rightarrow w_j$의 확신도라고 부르고 $Confidence(w_i \rightarrow w_j)$로 표기한다.

$$Confidence\ (w_i \rightarrow w_j) = \frac{Support(\{w_i, w_j\})}{Support(\{w_i\})}$$

실례로 표 1에서 련관규칙 $w_2 \rightarrow w_6$의 확신도는 3/4 = 0.75이다.

분명히 련관규칙은 일반적으로 대칭성을 만족시키지 않는다.

$$Confidence\ (w_i \rightarrow w_j) \neq Confidence\ (w_j \rightarrow w_i)$$

류사하게 3－항목모임 $\{w_i, w_j, w_k\}$에 대한 련관규칙들에 대한 확신도들을 정의할수 있다.

$$Confidence\ (w_i, w_j \rightarrow w_k) = \frac{Support(\{w_i, w_j, w_k\})}{Support(\{w_i, w_j\})}$$

$$Confidence\ (w_i \rightarrow w_j, w_k) = \frac{Support(\{w_i, w_j, w_k\})}{Support(\{w_i\})}$$

련관단어추출알고리듬은 다음과 같다.

이 알고리듬은 질문단어(항목) q와 련관이 있는 1차 및 2차련관단어들을 주어진 문장모임에서 추출하는 알고리듬이다.

걸음 1 1차련관단어

① 문서 D에 대하여 그 문서에 들어있는 단어들을 항목으로 하고 개개의 문장을 업무로 하는 문장업무표를 작성한다. 여기서 우리는 단어를 명사로 제한하였다.

② 질문단어 q와 단어 w에 대한 2－항목모임 $\{q, w\}$가 빈발항목모임이면 2－항목모임 $\{q, w\}$에 대한 련관규칙 $q \rightarrow w$의 확신도를 계산한다.

$$Confidence\ (q \rightarrow w) = \frac{Support(\{q, w\})}{Support(\{q\})}$$

③ 주어진 턱값(최소확신도) *Min-Con*에 대하여

$$Confidence\ (q \rightarrow w) \geq Min - Con$$

이면 단어 w를 질문단어 q의 1차련관단어모임에 출력한다.

걸음 2 2차련관단어

① 3－항목모임 $\{q, w_1, w\}$를 입력한다. 여기서 q는 질문단어, w_1은 질문단어의 1차련관단어이다.

② 3－항목모임 $\{q, w_1, w\}$가 빈발항목모임이면 련관규칙 $q, w_1 \rightarrow w$의 확신도를 계산한다.

$$Confidence\ (q, w_1 \rightarrow w) = \frac{Support(\{q, w_1, w\})}{Support(\{q, w_1\})}$$

③ 주어진 턱값(최소확신도) *Min-Con*에 대하여

$$Confidence\ (q, w_1 \rightarrow w) \geq Min - Con$$

이면 단어 w를 질문단어 q의 2차련관단어모임에 출력한다.

2. 실험 및 결과분석

실험에서는 질문지향단일문서요약에서 질문단어를 확장한 경우와 확장하지 않은 경우, 호상정보량에 기초하여 질문단어를 확장한 경우를 비교하였다.

질문지향단일문서요약법으로는 복합순위화법을 리용하였다.

비교방법들을 다음과 같이 약속한다.

① Baseline: 호상정보량을 리용하여 질문단어를 확장한 경우[2]

② Ex-term: 론문의 방법으로 질문단어를 확장한 경우

③ No-term: 질문단어를 확장하지 않은 경우

비교척도로는 최근에 많이 리용되는 *ROUGE-N*척도를 선택하였으며 련관단어모임을 추출하기 위해 설정되는 최소지지도와 최소확신도를 각각 0.5와 0.7로 설정하였다.

문장의 벡토르표현에서 단어출현확률 *tf*를 리용하였으며 질문단어와 1차 및 2차련관단어들에 대하여 각각 2, 1.5, 1.3의 무게를 고려하였다.

표 2에서 Ex-term(1)은 1차련관단어만을 고려하여 확장한 경우이며 Ex-term(2)는 1차 및 2차련관단어를 모두 고려하여 확장한 경우이다.

표 2. 실험비교표

	ROUGE-1	*ROUGE*-2	*ROUGE-W*
No-term	0.304 70	0.047 64	0.080 84
Baseline	0.373 96	0.072 44	0.098 67
Ex-term(1)	0.384 34	0.073 17	0.102 26
Ex-term(2)	0.385 23	0.073 19	0.102 29

표 2에서 보는바와 같이 질문확장법에 의한 단일문서요약은 질문비확장법에 비해 개선되였다는것을 알수 있다. 또한 질문확장법의 경우 단어들사이의 련관규칙에 의한 질문확장법이 호상정보량에 토대한 방법보다 더 좋으며 1차 및 2차련관단어를 모두 고려하면 보다 더 개선된다는것을 알수 있다.

맺 는 말

단어들사이의 련관규칙에 의해 질문단어의 1차 및 2차련관단어들을 추출하여 질문확장을 진행하는 방법에 의해 질문지향단일문서요약을 실현하는 복합순위화방법을 제안하고 선행한 방법에 비해 론문에서 제안한 질문확장법이 질문지향단일문서요약의 성능을 보다 개선하였다는것을 론증하였다.

참 고 문 헌

[1] **김일성**종합대학학보(자연과학), 62, 10, 37, 주체105(2016).

[2] 전금성, 정만흥; 정보기술 2, 36, 주체108(2019).

[3] Hajime Morita et al.; Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 223, 2011.

주체109(2020)년 8월 5일 원고접수

Extracting Association Words with Confidence of Association Rule and Application to Query-focused Single Document Summarization

Ri Su Jong, *Jong Man Hung*

In this paper, we presented a method that extended query word by extracting the first and second associative words of query word with confidence values of association rule between words. Our method improved the performance of query-focused single document summarization applied with query extension based on mutual information content of words.

Keywords: association rule, query extension, document summarization