# 용어 가중치에 의한 문서요약

박선* · 김철원*

*국립목포대학교, **호남대학교

# Document Summarization using Term Weighting

Sun Park* · Chul Won Kim**

*Mokpo National orea Maritime University, **Honam University

E-mail : sunpark@mokpo.ac.kr, cwkim@honam.ac.kr

## 요 약

본 논문은 용어 가중치에 의한 문서요약 방법을 제안한다제안된 방법은 의사연관피드백을 이용하여 사용자의 간섭을 최소화 시키며, 의미특징으로부터 유도된 용어의 가중치는 문장집합의 내부 특징을 잘 나타나기 때문에 문서요약의 질을 향상할 수 있다

## ABSTRACT

In this paper, we proposes a document summarization method using the term weighting. The proposed method can minimize the user intervention to use the pseudo relevance feedback. It also can improve the quality of document summaries because the inherent semantic of the sentence set are well reflected by term weighting derived from semantic feature.

## 키워드

문서 요약(document summarization), 의미특징(sematic features), 용어 가중치(term weighting), 비음수 행렬분해(NMF)

## Ⅰ. 서 론

문서요약은 문서의 전반적인 내용을 유지하면서 문서의 량을 자동으로 줄이는 작업으로 문서 및 문자 정보 증가로 인하여 많은 연구가 이루어지고 있다. 문서요약은 요약 목적에 따라서 포괄적 문서요약과 질의 기반의 문서요약의 구분할 수 있다. 또한 요약 대상문서에 따라서 단일문서 요약과 다중문서요약으로 구분할 수 있다.

문서요약에 대한 접근방법은 통계적 방법, 그래프기반 방법, 언어학기반 방법, 의미정보기반 방법, 외부자원기반 방법, 기타 복합기반 방법이 있다[1-11].

본 논문의 접근방법은 통계적 접근방법을 기반으로 의미정보기반 접근방법을 복합적으로 사용한다. 본 논문은 비음수 행렬분해로부터 추출된 의미특징과 이에 기반을 둔 용어의 가중치의 의미특징과 의사연관피드백을 이용하여 문장을 추출하여서 문서를 요약하는 질의기반 문서요약 방법을 제안한다.

## Ⅱ. 본 론

본 논문에서 제안한 방법은 전처리, 질의확장, 용어 가중치 계산, 문서요약 단계로 구성된다.

첫 단계는 전처리 단계로 문서를 문장으로 분해한 후 용어를 추출하여서 용어문장 행렬을 만든다.

두 번째 단계는 질의 확장 단계로 의사연관 피드백을 이용하여 사용자의 초기질의를 확장한다 (1)과 같은 양의 연과 피드백을 사용한다[12].

$$\overrightarrow{q^{new}} = \vec{q} + \sum_{\forall \vec{D_{*j}} \in D_+} D_{*j} \qquad (1)$$

여기서, $\overrightarrow{q^{new}}$는 의사연관 피드백을 이용하여 새롭게 확장된 질의이고, $\vec{q}$는 사용자의 초기 질의이다. D+는 연관된 문장을 포함한 연관 문장의 집합이다.

세 번째 용어 가중치 계산 단계는 확장된 질의와 비음수행렬분해된 의미특징을 이용하여 용어의 가중치를 계산한다. 마지막 단계는 문서요약 단계로 확장된 질의와 용어 가중치가 부여된 의미특징을 이용하여서 문서를 요약한다.

다음 식(2)와 같이 가중치 행렬을 계산하고, 용어문장 행렬 D에 식(3)와 같이 식(4)의 가중치 행렬을 대입하여 용어에 대한 가중치를 계산한다.

$$\tilde{D} = GD \tag{2}$$

여기서 G는 용어에 대한 가중치 값을 가지는 대각행렬로 식(3)과 같으며, D는 용어문장 행렬이고, $\tilde{D}$는 용어에 대한 가중치 값이 계산된 용어문장 행렬이다.

$$G = \begin{bmatrix} g_1 & 0 & \cdots & 0 \\ 0 & g_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g_n \end{bmatrix} \tag{3}$$

여기서 가중치 행렬 G는 용어에 대한 가중치를 원소 값으로 갖는 대각행렬로, a번째 용어와 일치하는 $D_{a*}$의 용어가 존재하는 경우 원소 ga는 원소 값을 가지며, 그렇지 않으면 1의 값을 가진다. a번째 용어에 대한 가중치 ga는 다음 식(4)과 같이 계산된다.

$$g_a = g_a^{olg} + \Delta g_a \tag{4}$$

여기서 $g_a^{olg}$는 확장된 질의에 포함된 용어가 a번째 용어와 일치하면 1을 가지며 일치하지 않으면 0을 가진다. $\Delta g_a$는 용어 a번째 행에 대한 전체 원소의 평균가중치의 변화량이다.

## III. 실험 및 평가

본 논문에서 제안방법(WPNMF)과 TF, QS, NMF등 세 가지 방법과 평가척도를 비교하였다. TF는 TFISF(term frequency inversed sentence frequency)에 기반을 두어 질의와 문장집합간의 유사도를 이용하여서 문서를 요약하는 방법이다[1]. QS는 Han이 제안한 방법으로 의사연관피드백에 질의 분해를 적용하여서 문서를 요약하는 방법이다[6]. NMF는 저자들이 이전에 제안한 방법으로 NMF는 질의와 비음수행렬분해된 문장집합의 의미특징을 이용하여 문서를 요약하는 방법이다[10].

요약문에 대한 평가척도 결과 제안 방법인 WPNMF의 평균 재현율이 TF와 비교해서는 6.7%가, QS와 비해서는 4.5%가, NMF와 비교해서는 2.5%가 더 높다. 평균 정확률은 TF와 비교해서는 8.5%가, QS와 비해서는 4.4%가, NMF와 비교해서는 3.3%가 더 높다. 평균 F-measure는 TF와 비교해서는 7.5%가, QS와 비해서는 4.5%가, NMF와 비교해서는 3.0%가 더 높다.

## V. 결 론

본 논문은 용어 가중치기반의 의미특징과 의사연관 피드백의 확장된 질의를 이용하여 의미 있는 문장을 추출하여서 문서를 요약하는 질의 기반 문서요약 방법을 제안하였다.

## 참고문헌


[1] I. Mani, M. T. Maybury, "dvances in Automatic Text," The MIT Press, 1999.

[2] A., Diaz, P., Gservas, "User-model based personalized summarization", Information Processing and Management, 43, pp.1715-1734, 2007.

[3] M., Sanderson, "Accurate user directed summarization from existing tools", In proceeding of the international conference on information and knowledge management, pp.45-51, 1998.

[4] A., Tombros, M., Sanderson, "Advantages of Query Biased summaries in Information Retrieval", In proceeding of ACM SIGIR, pp.2-10, 1998.

[5] R., Varadarajan, V., Hristidis, "A System for Query Specific Document Summarization", In proceeding of the CIKM, pp.622-631, 2006.

[6] Han, K. S., Bea, D. H., Rim, H. C., "Automatic Text Summarization Based on Relevance Feedback with Query Splitting", In proceedings of the 5th International Workshop on Information Retrieval with Asian Language, pp.201-202, 2000.

[7] 김철원, 박선, "의미특징과 워드넷 기반의 의사연관 피드백을 사용한 질의 기반의 문서요약", 한국해양정보통신학회논문지, 제15권 제

7호, 2010.

[8] S. Park, D. U. An, "Automatic Query-based Personalized Summarization that uses Pseudo Relevance Feedback with NMF", In proceeding of ACM ICUIMC2010, 2010.

[9] S. Park, "User-focused Automatic Document Summarization using Non-negative Matrix Factorization and Pseudo Relevance Feedback", In proceeding of ICCEA2009, 2009.

[10] 박선, "의미 특징 행렬과 의미 가변행렬을 이용한 질의 기반의 문서 요약", 한국항행학회 논문지, 제12권, 제4호, 2008.

[11] 박선, 이주홍, "비음수 행렬 분해와 K-means를 이용한 주제기반의 다중문서요약", 한국정보과학회 논문지, 제35권, 제4호, 2008.

[12] B. Y. Ricardo, R. N. Berthier, "Moden Information Retrieval," ACM Press, 1999.