의사연관피드백과 용어 가중치에 의한 문서요약

김철원* · 박선**

Document Summarization using Pseudo Relevance Feedback and Term Weighting

Chul Won Kim* · Sun Park**

요 약

본 논문은 의사연관피드백과 의미특징기반의 용어 가중치에 의한 문서요약 방법을 제안한다. 제안된 방법은 의사연관피드백을 이용하여 사용자의 간섭을 최소화 시키며, 의미특징으로부터 유도된 용어의 가중치는 문장집합의 내부 특징을 잘 나타내기 때문에 문서요약의 질을 향상할 수 있다. 또한 가중치가 부여된 의미특징과 확장된 질의를 이용하여서 사용자의 요구사항과 제안방법의 요약결과 사이의 의미적 차이를 감소시킨다. 실험결과 제안방법이 용어의 가중치를 부여하지 않은 방법에 비해서 좋은 성능을 보인다.

ABSTRACT

In this paper, we propose a document summarization method using the pseudo relevance feedback and the term weighting based on semantic features. 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. In addition, it uses the semantic feature of term weighting and the expanded query to reduce the semantic gap between the user's requirement and the result of proposed method. The experimental results demonstrate that the proposed method achieves better performance than other methods without term weighting.

키워드

문서요약, 의사연관피드백, 의미특징, 용어 가중치, 비음수행렬분해

Key word

document summarization, pseudo relevance feedback, semantic features, term weighting, non-negative matrix factorization (NMF)

* 종신회원 : 호남대학교 컴퓨터공학과 교수 (cwkim@honam.ac.kr)

** 정회원 : 목포대학교 정보산업연구소 연구전임교수

접수일자 : 2011. 12. 12

심사완료일자 : 2012. 01. 10

I. 서 론

통신기기의 발전은 사용자들로 하여금 인터넷 상의 정보를 더욱 쉽게 접근할 수 있게 하고 있으며, 이로 인하여 인터넷 상의 정보 증가를 더욱 가속화 시키고 있다. 또한 인터넷 사용이 가능한 소형 통신기기들의 보급증가는 대용량의 인터넷 정보를 소형 통신기기의 화면에 효율적으로 표시할 수 있는 정보 요약 기술에 대한 필요성을 증가시키고 있다. 인터넷 정보의 요약 기술은 주로 검색엔진에 의한 검색결과와 요약문인 스니펫(snippet) 생성에 주로 사용되고 있으며, 텍스트 기반의 문서요약이 주로 사용 및 연구되고 있다.

문서요약은 문서의 전반적인 내용을 유지하면서 문서의 양을 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 대상 문서에 따라서 단일문서요약과 다중문서요약으로 구분할 수 있다.

포괄적 문서요약(generic document summarization)은 문서 전체내용을 이해 할 수 있도록 문서의 주요 주제(topics)들을 추출하여 문서를 요약하는 방법이다. 질의 기반 문서 요약(query-based document summarization)은 사용자의 질의에 따라 질의에 관련 있는 주제로 문서를 요약 하는 방법이다[1]. 이외에도 질의 기반의 문서요약의 한 방법으로 인터넷 상의 사용자 로그와 사용자의 흥미와 관련된 특별한 정보로 문서를 요약하는 개인화된 문서요약이 있다[2]. 단일문서요약은 하나의 문서만을 대상으로 문서를 요약하며, 다중문서요약은 여러 개의 문서로부터 문서를 요약한다[1].

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

통계적 접근방법은 문서에 포함된 용어의 출현 빈도에 기반을 둔 방법으로 통계적 접근방법뿐만 아니라 대부분의 접근방법의 기본으로 많이 이용하는 방법이다. 용어의 출현빈도를 이용하기 때문에 문서요약에 쉽게 적용할 수 있으나, 용어와 문서간의 의미적 관계를 고려하지 않는 단점이 있다[1, 2].

그래프기반 접근방법은 문서에 포함된 문장이나 주제를 그래프의 노드 나타내며, 그래프의 에지는 노드

와 노드사이의 가중치를 나타낸다. 이 접근방법의 경우 문서를 그래프로 변환해야 하기 때문에 문서의 용량이 커질 수 록 자원소모가 많아지는 단점을 가진다[1, 3-6].

언어학기반 접근방법은 언어의 언어학적 구조에 기반으로 문서를 요약하는 방법이다. 이 접근방법의 경우 언어학적 분석을 위해서 복잡한 여러 단계를 수행하여야 한다[1, 2].

의미정보기반 접근방법은 문서에 잠재되어 있는 의미특징들을 이용하여 문서를 요약하는 방법이다. 이 접근 방법의 경우 문서 자체에 숨어 있는 구조적 특징을 이용하기 때문에 문서의 자체 구조에 많은 영향을 받는다[1, 8-11].

외부자원기반 접근방법은 위드넷이나 위키피디아와 같은 외부자원을 이용하여 문서를 요약하는 방법이다. 위드넷은 미리 완성된 영어 어휘목록을 가지고 있기 때문에 시간 경과에 따른 정보의 유효성에 대한 제약을 받을 수 있으며, 위키피디아 경우 대량의 자료를 전처리하는데 많은 자원이 소요된다[1, 7].

기타 복합기반 접근방법은 문서요약의 성능향상을 위하여 여러 개의 접근방법을 복합적으로 적용하여 문서를 요약하는 방법이다[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.

본 논문에서 의미특징을 추출하는데 사용하는 비음수 행렬 분해는 Lee와 Seung이 제안한 방법으로 비음수로 구성된 대량의 정보객체를 비음수로 구성된 두개의 객체로 분해하고, 이들의 선형조합으로 원본 객체를 근사적으로 표현할 수 있는 방법이다. 현재 비음수행렬분해는 이미지처리, 음향처리, 문서군집 및 정보요약 분야에서 기본적인 의미특징 추출방법으로 많이 사용하고 있다[12, 13].

연관 피드백은 질의를 관련 문장에는 가깝게 관련이 없는 문장과는 더 멀도록 새로운 질의로 확장하는 방법이다. 새로운 질의로 확장 시 사용자가 개입하여 확장하면 연관 피드백이라 하고, 사용자의 개입 없이 자동으로 질의를 확장하면 의사연관 피드백이라 한다[14-16].

본 논문에서 제안한 방법은 다음과 같은 장점을 갖는다. 첫째, 의사연관 피드백을 이용하여서 사용자가 요구하는 요약정보에 적합한 질의로 확장할 수 있다. 둘째, 확장된 질의와 의미 특징(semantic feature)을 이용하여 용어의 가중치를 계산하기 때문에 질의에 잘 부합되는 문장을 추출함으로써 문서요약의 질을 향상할 수 있다. 마지막으로, 용어 가중치가 반영된 의미특징과 확장된 질의를 이용하여 사용자의 요구사항과 제안방법에 의한 요약결과 간의 의미적 차이를 최소화 시킬 수 있다.

본 논문의 구성은 다음과 같다. 제2장에서는 문서요약에 대한 관련연구와 비음수행렬분해 방법에 대하여 알아보고, 제3장은 본 논문에서 제안하는 문서요약방법을, 제4장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제5장에서는 결론을 맺는다.

II. 관련연구

2.1. 문서요약

본 장에서는 통계학기반의 접근방법 중 연관피드백을 이용한 문서요약 방법과 의미정보기반 접근방법 등의 기타 복합적 접근방법의 문서요약방법의 최근 연구에 대하여 알아본다.

Sanderson은 문서상의 중요한 문장과 사용자가 개입된 질의 확장을 이용하여 문서를 요약 방법을 제안하였다[3]. Tombros와 Sanderson은 문서의 형식에 포함된 정보인 제목, 주제, 용어의 빈도 정보, 질의 등을 점수화하여서 사용자가 보조 정보로 활용할 수 있는 문서 요약 방법을 제안하였다[4]. Varadarajan과 Hristidis는 질의와 가장 관련이 높은 문장과 의미 연관을 이용하여서 문서로부터 추출된 복합 주제를 적용하여서 질의에 특화된 문서 요약 방법을 제안하였다[5]. Diaz와 Gervas는 용어의 위치와 주제 단어를 조합하는 포괄적 문서요약 방법과 사용자의 질의에 가장 적합한 문장을 추출하는 방법을 조합하여서 개인화된 문서요약 방법을 제안하였다[2]. Han 외 저자는 질의 분해와 연관 피드백을 이용한 문서요약 방법을 제안하였다. 이들의 방법은 질의 정보가 부족할 때에 좋은 요약 결과를 보이지 않는다[6].

본 논문의 저자들은 의미특징과 워드넷으로 질의를 확장하여서 문서를 요약하는 방법을 제안하였고[11], 개인화된 문서요약을 위하여 의사연관 피드백과 비음수 행렬분해를 이용한 요약방법을 제안하였으며[8], 비음수 행렬분해와 의사연관 피드백을 이용한 질의 기반의 문서요약 방법을 제안하였다[9].

2.2. 비음수행렬 분해

본 논문에서 행렬 X 의 j 번째 열벡터는 X^{*j} 로, i 번째 행벡터는 X_i^{*} 로, i 번째 행과 j 번째 열의 원소는 X_{ij} 로 표시 한다. 비음수행렬분해(NMF, non-negative matrix factorization)는 비음수로 구성된 원본 행렬을 두 개의 비음수 행렬로 분해하는 방법이다. Lee는 비음수 행렬분해를 위해 거리기반의 목적함수와 다양성기반 목적함수를 사용한 비음수 행렬분해 알고리즘을 제안하였다[12, 13]. 본 논문은 용어행렬 벡터로부터 문서를 요약하기 때문에 거리기반 목적함수를 이용한 비음수행렬분해 알고리즘을 이용한다.

본 논문에서 사용되는 비음수행렬분해 알고리즘은 식(1)의 거리 기반 목표함수 J 가 0에 가깝게 수렴 할 때까지 식(2)를 이용하여 행렬 W 와 H 의 값을 동시에 갱신하여서 원본 행렬 A 를 두개의 W 와 H 행렬로 분해한다. 여기서 Lee는 W 행렬을 의미특징 행렬로 H 행렬을 의미변수 행렬로 이름을 각각 정의하였다[12, 13].

$$J = \|A - WH\|^2 \quad (1)$$

식(1)의 목적은 비음수 행렬 A 에 대하여서 비음수 행렬 W 와 비음수 행렬 H 의 곱의 차가 최소가 될 때까지 식(2)를 반복적으로 계산하여 원본행렬을 분해하는 것이다. 여기서, A 는 m 개의 용어와 n 개의 문장으로 이루어진 $m \times n$ 행렬이고, W 은 $m \times r$ 행렬이며, H 는 $r \times n$ 행렬이다. r 은 의미특징의 개수로 일반적으로 행의 수보다 작게 설정한다.

$$H_{ij} \leftarrow H_{ij} \frac{(W^T A)_{ij}}{(W^T W H)_{ij}}, W_{ij} \leftarrow W_{ij} \frac{(A H^T)_{ji}}{(W H H^T)_{ji}} \quad (2)$$

두개로 분해된 의미특징 행렬 W 와 의미변수 행렬 H 의 선형조합으로 원본행렬 A 의 근사 값을 갖는 행렬로

구성할 수 있다. 이는 행렬 A 의 j 번째 열벡터 A^*j 는 행렬 W 의 1번째 열벡터 W^*1 과 행렬 H 의 요소 Hkj 가 선형조합을 이루며 식(3)과 같다. 즉, 1번째 의미 특징 벡터 W^*1 은 A^*j 의 문장 벡터 내에서의 가중치가 의미변수 Hlj 이다 [12, 13].

$$A^*j = \sum_{l=1}^r H_{lj} W^*l \quad (3)$$

의미특징 행렬 W 는 원본행렬의 내부 구조의 특징 (inherent feature)을 잘 표현할 수 있으며, 의미 변수 행렬 H 는 의미특징의 가중치를 나타낸다. 이러한 특성은 문장집합을 비음수행렬분해하면 W 행렬에 의해 문장집합에 포함되어 있는 주제(topics)들을 쉽게 추출할 수 있으며, H 행렬에 의해 문장집합에 포함된 주제의 중요도를 알 수 있다.

III. 제안방법

본 논문에서 제안한 방법은 전처리, 질의확장, 용어 가중치 계산, 문서요약 단계로 구성된다. 첫 단계는 전처리 단계로 문서를 문장으로 분해한 후 용어를 추출하여 용어문장 행렬을 만든다. 두 번째 단계는 질의 확장 단계로 의사연관 피드백을 이용하여 사용자의 초기질의의를 확장한다. 세 번째 용어 가중치 계산 단계는 확장된 질의와 비음수행렬분해된 의미특징을 이용하여 용어의 가중치를 계산한다. 마지막 단계는 문서요약 단계로 확장된 질의와 용어 가중치가 부여된 의미특징을 이용하여 문서를 요약한다.

3.1. 전처리

본 논문에서 사용하는 전처리 단계는 영문 문서를 위주로 설명한다. 한글 문서는 형태소 분석도구[17]를 이용하여 한글 용어를 추출하여 용어문장 행렬을 구성하면 된다.

전처리 단계는 문장 분해, 불용어(stop-word) 제거, 어근(stemming)을 추출, 용어문장 행렬생성 단계로 구성된다[14-16].

첫 번째 문장 분해 단계에서는 문서를 각각의 문장으

로 분해한다. 문장을 일정 크기를 기준으로 추출하는 방법은 문장이 중첩되거나 도중에 분할되어 정확한 문장의 의미를 전달 할 수 없는 경우가 발생할 수 있다. 이 때문에 본 논문에서는 문서요약에서 일반적으로 많이 사용하는 문장의 마침표를 기준으로 문장을 추출하는 방법을 이용 한다

두 번째인 불용어 제거 단계에서는 영어의 관사나 대명사와 같이 문장에서 불필요한 용어를 제거한다. 본 논문에서는 Rijsbergen의 불용어 목록[16]을 이용하여서 목록에서 정의하고 있는 용어들을 제거한다.

세 번째인 어근추출 단계에서는 Porter의 어근추출 알고리즘[16]을 이용하여서 영어의 파생어들을 가장 중심이 되는 용어인 어근으로 변환한다.

마지막 단계인 용어문장 행렬생성 단계에서는 용어와 문장사이의 빈도 행렬을 구성한다. 생성된 용어문장 행렬 D 에서, j 번째 문장 벡터 D^*j 는 용어문장빈도 벡터 $D^*j = [d1j, d2j, \dots, dnj]^T$ 로 표현되고, 벡터 D^*j 요소인 dij 는 j 번째 문장에서 i 번째 용어를 나타낸다.

3.2. 질의 확장

질의 확장 단계는 의사연관피드백을 이용하여 사용자의 초기질의의를 확장한다. 의사연관 피드백은 질의와의 유사도가 상위 k 개인 문장을 이용하여서 질의를 확장하는 방법이다. 질의 확장을 위해서 사용되는 문장의 개수가 너무 많으면 요약 결과 사용자의 원하는 주제에 대해 너무 포괄적인 모호한 의미를 나타낸다. 또한, 문장의 개수가 너무 적으면 요약의 결과가 사용자가 요구하는 주제에 너무 협소한 요약 결과를 나타내게 된다.

이 때문에 본 논문에서는 실험결과 가장 좋은 결과는 나타내는 상위 5개의 문장을 이용하여서 초기질의의를 확장한다. 의사연관 피드백은 연과 피드백과는 다르게 비연관 문서를 판단 할 수 없기 때문에 식(4)와 같은 양의 연과 피드백을 사용한다[14, 16].

$$\overrightarrow{q^{new}} = \overrightarrow{q} + \sum_{\forall D_j \in D_+} D_j \quad (4)$$

여기서, $\overrightarrow{q^{new}}$ 는 의사연관 피드백을 이용하여 새롭게 확장된 질의이고, $\overrightarrow{q}$ 는 사용자의 초기 질의이다. D_+ 는

연관된 문장을 포함한 연관 문장의 집합이다. 연관 문장은 초기 질의와 유사도가 가장 높은 상위 5개의 문장들로 식(5)의 코사인 유사도를 이용하여 계산한다.

다음은 유사도를 계산하는 코사인 유사도이다[13, 14, 15].

$$\text{sim}(D_{*j}, \vec{q}) = \frac{\sum_{i=1}^n d_{ij} \times q_i}{\sqrt{\sum_{i=1}^n d_{ij}^2} \times \sqrt{\sum_{i=1}^n q_i^2}} \quad (5)$$

여기서 D_{*j} 는 j 번째 문자의 벡터를 나타내고, $\vec{q}$ 는 사용자의 초기질의를 나타내며, n 은 용어의 수를 나타낸다.

3.3. 용어 가중치 계산

본 논문에서 사용되는 용어 가중치 계산방법은 이전에 본 저자들이 제안한 방법[18]을 문서요약에 맞도록 수정하여 사용한다. 용어 가중치 계산 단계에서는 다음 식(7)와 같이 가중치 행렬을 계산하고, 용어문장 행렬 D 에 식(6)과 같이 식(7)의 가중치 행렬을 대입하여 용어에 대한 가중치를 계산한다.

$$\tilde{D} = GD \quad (6)$$

여기서 G 는 용어에 대한 가중치 값을 가지는 대각행렬로 식(7)과 같으며, D 는 용어문장 행렬이고, $\tilde{D}$ 는 용어에 대한 가중치 값이 계산된 용어문장 행렬이다.

$$G = \begin{bmatrix} g_1 & 0 & \cdots & 0 \\ 0 & g_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g_n \end{bmatrix} \quad (7)$$

여기서 가중치 행렬 G 는 용어에 대한 가중치를 원소 값으로 갖는 대각행렬로, a 번째 용어와 일치하는 D_{*a} 의 용어가 존재하는 경우 원소 g_a 는 원소 값을 가지며, 그렇지 않으면 1의 값을 가진다. a 번째 용어에 대한 가중치 g_a 는 다음 식(8)과 같이 계산된다.

$$g_a = g_a^{olg} + \Delta g_a \quad (8)$$

여기서 g_a^{olg} 는 확장된 질의에 포함된 용어가 a 번째 용어와 일치하면 1을 가지며 일치하지 않으면 0을 가진다. Δg_a 는 용어 a 번째 행에 대한 전체 원소의 평균가중치의 변화량으로 식(9)에 의해서 추정된다.

$$\Delta g_a = E(\Delta g_a^i) = \frac{1}{n} \sum_{i=1}^n \frac{1}{D_{ai}} \sum_{k \in I_i} \Delta H_{ki} W_{ak} \quad (9)$$

여기서 n 은 전체문장의 개수이고, I_i 는 의미변수 행렬 H 에서 i 번째 문서의 의미변수벡터 H^*i 에서 $\Delta H_{ki} \neq 0$ 이 아닌 k 들의 집합이고, ΔH_{ki} 는 식(10)에 의해서 유도된다.

$$\Delta H_{ij} = \frac{H_{ji}}{\sum_{k=1}^r H_{ji}} \quad (10)$$

여기서 ΔH_{ji} 는 정규화된 H_{ji} 이다.

3.4. 문서요약

의미특징을 이용한 질의 기반의 문장 요약방법은 이전 저자들이 제안한 방법[10, 11]에 용어가중치를 적용하여 수정한 방법을 이용한다. 이전 장에서 식(6)에 의해 용어 가중치가 계산된 용어문장 행렬 $\tilde{D}$ 를 비음수행렬 분해해서 두개의 의미특징 행렬 W 와 H 를 계산한다. 식(5)을 이용하여 비음수 의미특징 열벡터 W_{*j} 와 의사연관 피드백에 의해서 확장된 질의 간의 유사도를 계산고, 유사도가 가장 높은 의미특징 열벡터 W_{*T} 를 선택한다. 그런 다음에 선택된 의미 특징 열벡터 W_{*T} 와 대응되는 의미 변수 행벡터 H_{T*} 를 선택한다. 마지막으로 선택된 의미 특징 열벡터에서 가장 큰 요소 값과 대응되는 문장을 요약문장으로 추출한다.

IV. 실험 및 평가

본 논문에서 사용되는 실험 자료는 야후코리아 뉴스로부터 20건의 질의 각각에 대하여 순위별로 검색된 50건의 기사를 선택하여 총 1000건의 문서를 준비하였다. 제안 방법을 비교하기 위하여 세 명의 평가자가 각

각의 질의에 관련된 각기 50건의 문서로부터 10개의 문장으로 구성된 20건의 요약문을 수동으로 요약하였다. 수동으로 요약된 요약문과 20건의 질의에 연관된 50건의 문서로부터 문서요약 방법으로 10개의 문장으로 요약된 20건의 요약결과를 비교하여 성능을 평가하였다.

성능평가 척도는 정보검색에서 주로 사용되는 정확률(P; precision), 재현율(R; recall), F-measure(F)를 이용하였으며 평가척도는 다음 식(11)과 같다[14-16].

$$(R) = \frac{|S_{man} \cap S_{sum}|}{|D|}, (P) = \frac{|S_{man} \cap S_{sum}|}{|S_{sum}|}, (F) = \frac{2RP}{R+P} \quad (11)$$

여기서 D는 질의에 연관된 문장집합으로 |D|은 이 집합의 문장 수를 나타내고, S_{man}는 사람이 요약한 요약문이며 S_{sum}은 요약방법에 의하여 요약된 결과이다. |S_{man} ∩ S_{sum}|는 S_{man}와 S_{sum}의 교집합의 문장 수이다.

본 논문에서 제안방법(WPNMF)과 TF, QS, NMF, PNMF 등 다섯 가지 방법을 c언어로 구현하여 평가척도를 이용하여서 비교 평가하였다. TF는 TFISF(term frequency inversed sentence frequency)에 기반을 두어 질의와 문장집합간의 유사도를 이용하여서 문서를 요약하는 방법이다[1, 14, 16]. QS는 Han이 제안한 방법으로 의사연관피드백에 질의 분해를 적용하여서 문서를 요약하는 방법이다[6]. NMF와 PNMF는 저자들이 이전에 제안한 방법으로 NMF는 질의와 비음수행렬분해된 문장집합의 의미특징을 이용하여 문서를 요약하는 방법이며[10], PNMF는 의사연관피드백에 의한 확장질의와 의미특징을 이용하여 문서를 요약하는 방법이다[8]. 다음 표1은 5가지 요약 방법에 대한 평가 결과이다.

표 1. 평가 결과
Table. 1 Result of evaluation

구분	TF	QS	NMF	PNMF	WPNMF
average R	0.29	0.312	0.329	0.342	0.357
average P	0.31	0.351	0.362	0.381	0.395
average F	0.30	0.33	0.34	0.36	0.38

표1에서 보는 것과 같이 요약문에 대한 평가척도 결과 제안 방법인 WPNMF의 평균 재현율이 TF와 비교해서는 6.7%가, QS와 비해서는 4.5%가, NMF와 비교해서는 2.5%가, PNMF와 비교해서 1.5%가 더 높다. 평균 정확률은 TF와 비교해서는 8.5%가, QS와 비해서는 4.4%가, NMF와 비교해서는 3.3%가, PNMF와 비교해서 1.4%가 더 높다. 평균 F-measure는 TF와 비교해서는 7.5%가, QS와 비해서는 4.5%가, NMF와 비교해서는 3.0%가, PNMF와 비교해서 1.5%가 더 높다.

성능 평가 결과 제안방법인 WPNMF가 가장 좋은 결과를 나타내며, 다음으로 PNMF, NMF, QS, TF 순으로 평가 되었다. 이는 단순히 질의와 문장 간의 유사도에 의한 문서를 요약하는 TF 방법보다는 질의를 분해하여서 확장한 QS 방법이 더 좋은 성능을 나타내는 것을 알 수 있다. 또한 QS 방법의 단순한 질의 분해에 의한 질의 확장보다는 질의와 문장집합 내부의 특성을 이용하는 NMF 방법이 더 좋은 성능을 보이는 것을 알 수 있으며, 여기에 질의를 확장하여 문장집합의 내부 특성을 나타내는 의미특징을 이용한 PNMF방법이 NMF 방법보다는 더 좋은 성능을 나타내는 것을 알 수 있다. 특히 제안방법인 WPNMF 방법이 가장 좋은 성능을 나타내는 것은 확장된 질의와 용어의 가중치에 기반을 둔 의미특징을 이용함으로써 사용자의 의도를 가장 잘 반영하여 문서요약의 질을 높이는 것을 판단 된다.

V. 결 론

본 논문은 용어 가중치기반의 의미특징과 의사연관 피드백의 확장된 질의를 이용하여 의미 있는 문장을 추출하여서 문서를 요약하는 질의 기반 문서요약 방법을 제안하였다. 제안 방법은 비음수행렬분해로부터 유도된 용어의 가중치를 이용하여 일반 의미특징을 이용한 문서요약 방법에 비하여서 문서의 중요 주제를 요약 결과에 더 잘 반영하였다. 또한 의사연관 피드백에 의한 확장된 질의와 용어 가중치기반의 의미특징을 이용하여서 사용자가 원하는 요약문과 문서요약 방법에 의한 요약결과 간의 의미적 차이를 감소함으로써 요약문의 질을 높였다.

실험결과 용어의 가중치를 반영하지 않은 방법에 비하여 제안된 방법이 더 좋은 평가 결과를 보이고 있다. 그러나 본 논문의 제안방법은 한글문서를 대상으로 실험을 하였기에 영어나 다른 나라의 언어에 대하여서는 적용이 미흡한 편이다. 앞으로 현재 방법을 다양한 언어를 지원할 수 있도록 확장할 예정이다.

참고문헌

- [1] I. Mani, M. T. Maybury, "Advances 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 제7호, 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] D. D. Lee, H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, pp.788-791, 1999.
- [13] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization," In Advances in Neural Information Processing Systems, vol. 13, pp.556-562, 2001.
- [14] B. Y. Ricardo, R. N. Berthier, "Modern Information Retrieval," ACM Press, 1999.
- [15] S. Chakrabarti, "mining the web: Discovering Knowledge from Hypertext Data," Morgan Kaufmann Publishers, 2003.
- [16] W. B. Frakes, B. Y. Ricardo, "Information Retrieval : Data Structure & Algorithms", Prentice-Hall, 1992.
- [17] 한경한, 남경완, "한국어 정보 처리 입문 : 컴퓨터가 우리말을 이해하려면", 커뮤니케이션북스, 2007.
- [18] 이주홍, 박선, "NMF 기반의 용어 가중치 재산정을 이용한 문서군집", 한국컴퓨터정보학회논문지, 제13권 제54호, pp.11-18, 2008.

저자소개

김철원(Chul-won Kim)

1997년 광운대학교컴퓨터공학과
(공학박사)
1998년 ~ 현재 호남대학교
컴퓨터공학과 교수

※관심분야: XML 응용, 멀티미디어 정보검색,
멀티미디어 정보처리 및 응용

박 선(Sun Park)

1996년 전주대학교 전자계산학과
(이학사)

2001년 한남대학교 정보통신학과
(공학석사)

2007년 인하대학교 컴퓨터정보공학과 (공학박사)

2008년 ~ 2009년 8월 호남대학교 컴퓨터공학과
전임강사

2009년 9월 ~ 2010년 12월 전북대학교 BK21- 전북전자
정보 고급인력양성사업단 박사후 과정

2010년 12월 ~ 현재 목포대학교 정보산업연구소
연구교수

※ 관심분야: 정보검색, 데이터마이닝, 인공지능
IT정보융합, 데이터베이스, 정보보안