의미특징과 워드넷 기반의 의사 연관 피드백을 사용한 질의기반 문서요약

김철원* · 박선**

Query-based Document Summarization using Pseudo Relevance Feedback based on Semantic Features and WordNet

Chul-won Kim* · Sun Park**

요 약

본 논문은 의미특징과 워드넷 기반의 의사연관피드백을 이용하여 사용자의 질의에 관련 있는 의미 있는 문장을 추출하여 문서요약을 하는 새로운 방법을 제안한다. 제안된 방법은 비음수 행렬 분해로 부터 유도된 의미특징이 문서의 잠재의미를 잘 나타나기 때문에 문서요약의 질을 향상할 수 있다. 또한 의미특징과 워드넷 기반의 의사연관피드백을 이용하여서 사용자의 요구사항과 제안방법의 요약결과 사이의 의미적 차이를 감소시킨다. 실험결과 제안방법이 유사도, 비음수행렬분해를 이용한 방법들에 비하여 좋은 성능을 보인다.

ABSTRACT

In this paper, a new document summarization method, which uses the semantic features and the pseudo relevance feedback (PRF) by using WordNet, is introduced to extract meaningful sentences relevant to a user query. The proposed method can improve the quality of document summaries because the inherent semantic of the documents are well reflected by the semantic features from NMF. In addition, it uses the PRF by the semantic features and WordNet to reduce the semantic gap between the high level user's requirement and the low level vector representation. The experimental results demonstrate that the proposed method achieves better performance than the other methods.

키워드

질의 기반 문서요약, 의사 연관 피드백, 워드넷, 의미특징, 비음수 행렬 분해

Key word

Query-based document summarization, pseudo relevance feedback, WordNet, semantic features, non-negative matrix factorization

* 종신회원 : 호남대학교 컴퓨터공학과 교수

** 정회원 : 목포대학교 정보산업연구소 연구교수

접수일자 : 2011. 03. 08

심사완료일자 : 2011. 03. 16

Ⅰ. 서 론

인터넷 상의 지속적인 정보의 증가는 사용자들에게 필요한 정보만을 검색할 수 있는 방법을 요구하고 있다. 특히, 인터넷 사용 시 고정된 일반 단말기로부터 이동용 소형화된 인터넷 사용 단말기로 이용이 증가 되고 있다. 이러한 소형 인터넷 단말기의 폭발적인 사용은 대량의 정보로부터 사용자가 필요로 하는 정보를 소형의 화면 상에 표시할 수 있도록 효율적으로 정보를 요약할 수 있는 방법을 더욱 필요로 하고 있다. 또한 사용자가 필요로 하는 사용자의 질의(query)에 대한 검색 정보의 맞춤형 검색을 지원하도록 하는 문서 요약의 필요성을 점차 증가 시키고 있다.

문서 요약은 문서의 기본적인 내용을 유지하면서 문서의 량을 자동으로 줄이는 작업으로 인터넷의 폭발적인 문서 및 문자 정보의 증가로 많은 연구가 지속적으로 이루어지고 있다[1]. 이러한 자동 문서 요약은 포괄적 문서요약 방법과 질의 기반의 문서요약 방법으로 구분할 수 있다.

포괄적 문서 요약(generic document summarization)방법은 사용자의 개입 없이 문서 내용전체를 필요한 정보로 자동 요약하는 방법이다. 질의 기반 문서 요약(query-based document summarization) 방법은 사용자가 요구하는 질의에 따라 질의에 관련 있는 내용만으로 문서를 자동으로 요약 하는 방법이다[1]. 요약을 할 대상 문서의 종류에 따라서 하나의 문서만을 기반으로 할 경우는 단일문서요약으로, 신문의 기사와 같이 하나의 주제에 시간의 경과에 따라서 여러 개의 문서들로부터 요약하는 경우는 다중문서요약 이라고 한다. 또한 질의 기반의 문서요약 방법의 한 종류로 문서 자체의 정보를 요약하는 것 보다는 사용자의 흥미와 관련된 특별한 정보를 유지하면서 문서를 요약하는 개인화된 문서요약 방법이 있다[2].

다음은 질의 기반의 문서요약에 대한 최근의 연구들이다. 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 접합한 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[7], 개인화된 문서요약을 위하여서 의사연관 피드백과 비음수 행렬분해를 이용한 요약방법을 제안하였으며[8], 비음수 행렬분해와 의사연관 피드백을 이용한 질의 기반의 문서요약 방법을 제안하였다[9].

본 논문은 비음수 행렬분해로부터 추출된 의미특징과 워드넷 기반의 의사연관 피드백을 이용하여 문서를 요약하는 질의 기반의 새로운 문서요약 방법을 제안한다.

본 논문에서 의미특징을 추출하는데 사용하는 비음수 행렬 분해는 Lee와 Seung이 제안한 방법으로 비음수의 자료행렬을 비음수 의미 특징(NSF, non-negative semantic features)과 비음수 의미 변수(NSV, non-negative semantic variable)의 두개의 의미특징 행렬로 인수 분해하는 방법이다[10, 11].

워드넷은 영어의 어휘목록으로 영어 단어를 유의어 집단으로 분류하여서 정의제공하고, 어휘목록 사이의 다양한 의미 관계를 나타낼 수 있도록 한 워드넷은 어휘 데이터베이스로 온라인 서비스나 프로그램 안에서 사용될 수 있도록 설계되었다. 워드넷은 어휘 개념으로 영어의 명사, 동사, 접속사, 부사 등의 이음동의어 집합을 지원한다[12].

연관 피드백은 질의를 관련 문장과 가깝게 관련이 없는 문장과는 더 멀도록 새로운 질의로 확장하는 방법이다. 새로운 질의로 확장 시 사용자가 직접 개입하여 확장하면 연관 피드백이라 하고, 사용자의 개입 없이 자동으로 질의를 확장하면 의사연관 피드백이라 한다[13, 14].

제안된 문서요약 방법은 다음과 같다. 사용자의 질의를 워드넷을 이용하여서 1차 확장하고, 요약할 문서를 문장으로 분해하고, 의사연관 피드백을 이용하여

서 문장과 1차 확장 질의로 부터 2차 질의로 확장한다. 최종 확장된 질의와 의미특징을 이용하여 문서를 요약한다.

제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 워드넷과 의사연관 피드백을 이용하여서 사용자가 요구하는 요약정보에 적합한 질의로 확장할 수 있다. 둘째, 의미 특징(semantic feature)에 의해서 문서의 고유 의미 구조(inherent semantic structure)[10]를 잘 반영하기 때문에 문서로 부터 사용자가 요구하는 정보를 잘 반영할 수 있다. 마지막으로, 의미특징과 확장된 질의의 조합으로 문서요약의 질을 높일 수 있다.

본 논문의 구성은 다음과 같다. 제2장에서는 의사연관 피드백과 비음수 행렬 분해 방법을, 제3장은 제안한 요약방법을, 제4장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제5장에서 결론은 맺는다.

Ⅱ. 관련연구

2.1. 연관 피드백

연관 피드백의 기본이 되는 방법은 Rocchio의 방법으로, 원래의 질의 벡터 $\vec{q}$에 연관된 문서 D+에 대응하는 벡터의 가중치 합을 단순히 더하고, 비연관 문서 D-의 가중치 합을 빼는 방법으로 식(1)과 같다[13, 14].

본 논문에서 행렬 X의 j번째 열벡터는 X*j로, i번째 행벡터는 Xi*로, i번째 행과 j번째 열의 원소는 Xij로 표시 한다.

$$\overrightarrow{q^{new}} = \alpha\vec{q} + \beta \sum_{\forall d_{*j} \in D_+} \overrightarrow{d_{*j}} - \gamma \sum_{\forall d_{*j} \in D_-} \overrightarrow{d_{*j}} \tag{1}$$

여기서, $\overrightarrow{q^{new}}$는 새롭게 확장된 질의이고, α, β, γ는 조정이 가능한 매개변수들로 일반적으로 $\alpha=\beta=\gamma=1$로 고정하여 사용하며, $\overrightarrow{d_{*j}}$는 j번째 문서의 벡터이다. D_+와 D_-는 각각 연관 문서 및 비연관 문서 집합으로서, 사용자에 의서 수동으로 선택되면 연관 피드백이라 하고, 자동으로 선택되면 의사연관 피드백이라 한다.

2.2. 비음수행렬 분해

이번 장에서는 비음수행렬 분해(NMF, non-negative matrix factorization)의 기본적인 개념과 알고리즘에 대하여서 알아본다. 비음수행렬분해는 비음수로 구성된 대량의 객체정보로부터 두 개의 행렬로 구성된 비음수로 된 부분정보를 추출하고, 이들의 선형 조합으로 객체를 표현할 수 있도록 하는 방법이다. 추출된 첫 번째 행렬은 객체의 부분정보를 두 번째 행렬은 부분정보에 대한 가중치로 나타낼 수 있다[10, 11].

비음수 행렬 분해 알고리즘은 주어진 비음수 행렬로부터 두 개의 비음수의 인수를 찾는 행렬 로 분해한다[10, 11]. 비음수 행렬 분해 알고리즘은, 식(2)의 목표함수 J가 0에 가깝게 수렴 할 때까지 식(3)과 식(4)을 이용하여 행렬 W와 H의 값을 동시에 갱신한다.

$$J = \|A - WH\|^2 \tag{2}$$

식(3)과 식(4)의 목적은 행렬 A를 비음수 m×r 행렬 W와 비음수 $r \times n$ 행렬 H로 분해하는 것이다. 여기서, A는 m개의 용어와 n개의 문장으로 이루어진 $m \times n$ 행렬이고, r은 의미특징의 개수로 일반적으로 행의 수보다 작게 설정한다.

$$H_{ij} \leftarrow H_{ij} \frac{(W^T A)_{ij}}{(W^T WH)_{ij}} \tag{3}$$

$$W_{ij} \leftarrow W_{ji} \frac{(AH^T)_{ji}}{(WHH^T)_{ji}} \tag{4}$$

예1) 다음 그림1은 4×5 행렬의 Matlab 7.8의 *nnmf()*함수를 이용하여서 비음수행렬 분해한 결과이다.

$$\underset{A}{\begin{bmatrix} 4 & 2 & 0 & 1 & 0 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 2 & 5 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}} = \underset{W}{\begin{bmatrix} 1.018 & 4.456 & 0 \\ 2.883 & 0 & 2.442 \\ 4.997 & 0 & 1.499 \\ 0 & 1.307 & 0.785 \end{bmatrix}} \times \underset{H}{\begin{bmatrix} 0.025 & 0 & 0.145 & 0.989 & 0.035 \\ 0.879 & 0.477 & 0 & 0 & 0 \\ 0 & 0.130 & 0.756 & 0.054 & 0.662 \end{bmatrix}}$$

그림 1. 비음수 행렬분해에 의한 행렬 분해 결과
Fig. 1. Result of factorization by NMF

행렬 A의 j번째 열벡터 A_{*j}는 행렬 W의 1번째 열벡터 W_{*l}과 행렬 H의 요소 H_{kj}가 선형조합을 이루며 식(5)과 같다. 즉, 1번째 의미 특징 벡터 W_{*l}은 A_{*j}의 문장 벡터 내에서의 가중치가 의미변수 H_{lj}이다.

$$A_{*j} = \sum_{l=1}^{r} H_{lj} W_{*l} \tag{5}$$

예2) 그림1의 4×5 행렬을 4개의 용어와 5개의 문장으로 구성된 용어문장 빈도 행렬로 가정할 때, 그림2에서와 같이 하나의 문장을 분해된 두 개의 행렬의 선형조합으로 나타낼 수 있다.

$$\underset{A_{*1}}{\begin{bmatrix}4\\0\\0\\1\end{bmatrix}} \approx \underset{H_{11}}{0.026} \times \underset{W_{*1}}{\begin{bmatrix}1.018\\2.883\\4.997\\0\end{bmatrix}} + \underset{H_{21}}{0.879} \times \underset{W_{*2}}{\begin{bmatrix}4.456\\0\\0\\1.307\end{bmatrix}} + \underset{H_{34}}{0} \times \underset{W_{*3}}{\begin{bmatrix}0\\2.442\\1.499\\0.785\end{bmatrix}}$$

그림 2. H와 W행렬 원소의 선형 조합에 의한 문장의 표현
Fig. 2. Representing of sentence by linear combination of elements of matrix H and W.

Ⅲ. 제안방법

본 논문에서 제안한 방법은 다음과 같다. 첫 단계는 전처리 단계로 문서를 문장으로 분해해서 용어문장 행렬 벡터를 만든다. 두 번째 단계는 질의 확장 단계로 워드넷과 의사연관 피드백을 이용한다. 마지막 단계는 문서요약 단계로 확장된 질의와 의미특징을 이용하여서 문서를 요약한다.

3.1. 전처리

전처리 단계는 문장 분해, 불용어(stop-word) 제거, 어근(stemming)을 추출, 용어빈도 벡터 생성 단계로 구성된다[14, 15].

첫 번째인 문장 분해 단계에서는 주어진 문서를 각각의 문장으로 분해하는 단계이다. 일반적으로 문서요약 방법에서 사용되는 문장을 분해하는 방법은 일정한 문장의 크기로 지정하여서 모든 문장을 같은 크기로 추출 분해하는 방법과, 문서상에서 문장의 마침표를 기준으로 문장이 끝나는 부분까지 추출 분해하는 방법이 있다.

일정 크기를 기준으로 문장을 추출 분해하는 방법은 문장이 중첩되거나 도중에 분할되어 정확한 문장의 의미를 전달 할 수 없는 경우가 많이 발생한다. 그렇기 때문에 본 논문에서는 마침표를 기준으로 완전한 한 문장씩 추출 분해하는 방법을 이용한다.

두 번째인 불용어 제거 단계에서는 Rijsbergen의 불용어 목록[15]을 이용하여서 목록에서 정의하고 있는 용어들을 제거한다.

세 번째인 어근추출 단계에서는 Porter의 어근추출 알고리즘[15]을 이용하여서 영어의 파생어들을 가장 중심이 되는 용어인 어근으로 변환한다.

마지막 단계인 용어빈도 벡터생성 단계에서는 용어와 문장사이의 빈도 행렬을 다음과 같이 구성한다. 생성된 용어빈도 행렬 D는, j번째 문장 벡터 D_{*j}는 용어문장빈도 벡터 $D_{*j} = [\, d_{1j}, d_{2j}, \cdots, d_{nj} \,]^T$로 표현되고, 벡터 D_{*j} 요인 d_{ij}는 j번째 문장에서 i번째 용어를 나타낸다.

3.2. 질의 확장

질의 확장 단계도 워드넷을 이용한 1차 질의확장과 의사연관 피드백을 이용한 2차 2질의확장 단계로 나눈다.

3.2.1. 1차질의 확장

1차질의 확장은 워드넷을 이용하여 사용자의 기본 질의를 유의어로 확장한다. 1차질의 확장은 워드넷의 명사 관련 유의어 많을 이용하여서 질의를 확장한다. 워드넷의 동사 및 다른 품사의 유의어를 사용하여서 질의를 확장하면 너무 많은 의미를 포함하기 때문에 오히려 요약의 질이 떨어 질 수 있다. 다음 표1은 워드넷 2.1을 이용하여서 'document'의 명사 유의어로 확장한 것이다.

표 1. 워드넷을 이용한 1차질의 확장
Table 1. First query expansion by WordNet

기본 질의	1차 확장 질의
document	writing representation communication computer file

3.2.1. 2차질의 확장

워드넷 만을 이용하여서 1차 질의를 확장할 경우 사용자의 질의가 문서에 충분히 반영될 수 없을 수 있다. 이 때문에 본 논문에서는 사용자의 질의가 요약 문서에 충분히 반영할 수 있도록 의사 연과 피드백을 이용하여서 질의를 2차로 확장한다.

의사연관 피드백은 질의와 문장사이의 유사도를 계산하여서 질의와의 유사도가 상위 k개인 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 상위 5개의 문장을 이용하여서 2차 질의를 확장한다.

다음 식(6)은 질의와 문장 간의 유사도를 계산하는 코사인 유사도이다[13, 14, 15].

$$sim(d_{*j},q)=\frac{\sum_{i=1}^{n} d_{ij}\times q_i}{\sqrt{\sum_{i=1}^{n} d_{ij}}\times\sqrt{\sum_{i=1}^{n} q_i}} \tag{6}$$

여기서 d_{*j}는 j번째 문자의 벡터를 나타내고, q는 질의를 나타내며, n은 용어의 수를 나타낸다.

의사연관 피드백은 연과 피드백과 같이 비연관 문서를 판단 할 수 없기 때문에 식(7)과 같은 양의 연과 피드백을 사용한다.

$$\overrightarrow{q^{new2}}=\overrightarrow{q^{new1}}+\sum_{\forall s_{*j}\in D_{+}}\overrightarrow{s_{*j}} \tag{7}$$

여기서, $\overrightarrow{q^{new2}}$는 의사연관 피드백을 이용하여 새롭게 확장된 2차 질의이고, $\overrightarrow{q^{new1}}$는 워드넷에 의해 확장된 1차 질의이다. s는 연관된 문서에 포함된 연관된 문장이다. 연관 문장은 1차 질의와 유사도가 가장 높은 상위 5개의 문장들이다.

다음 그림3은 의사연관 피드백을 이용하여서 1차 질의를 2차 질의로 확장하는 예를 나타낸 것이다.

$$\underset{q^{new2}}{\begin{bmatrix}3\\1\\2\\1\\3\\1\\1\end{bmatrix}}=\underset{q^{new1}}{\begin{bmatrix}1\\1\\1\\0\\0\\0\\0\end{bmatrix}}+\underset{s_{*1}}{\begin{bmatrix}1\\0\\0\\0\\1\\0\\0\end{bmatrix}}+\underset{s_{*2}}{\begin{bmatrix}0\\0\\1\\0\\0\\0\\0\end{bmatrix}}+\underset{s_{*3}}{\begin{bmatrix}0\\0\\0\\1\\1\\0\\0\end{bmatrix}}+\underset{s_{*4}}{\begin{bmatrix}0\\0\\0\\0\\1\\1\\0\end{bmatrix}}+\underset{s_{*5}}{\begin{bmatrix}1\\0\\0\\0\\0\\0\\1\end{bmatrix}}$$

그림 3. 의사연관 피드백을 이용한 2차질의 확장
Fig. 3. Second query expansion by pseudo relevance feedback

3.3. 문서요약

의미특징을 이용한 질의 기반의 문장 요약방법[16, 17]은 다음과 같다. 전처리된 행렬 A를 비음수 행렬 분해하여 두개의 의미특징 행렬을 계산한다. 이 결과 계산된 비음수 의미 특징 행렬 W와 비음수 의미 변수 행렬 H는 식(3)과 식(4)과 같다[10, 11]. 식(6)을 이용하여 비음수 의미특징 열벡터와 질의 간의 유사도를 계산고, 유사도가 가장 높은 의미특징 열벡터를 선택한다. 그런 다음에 선택된 의미 특징 열벡터와 대응 되는 의미 변수 행벡터를 선택한다. 마지막으로 선택된 의미 특징 열벡터에서 가장 큰 요소 값과 대응되는 문장을 요약문장으로 추출한다.

Ⅳ. 실험 및 평가

본 논문에서 사용되는 실험 자료는 야후코리아 뉴스로 부터 20건의 질의에 대여서 각각의 질의 순위별로 20건의 기사 선택하였다. 다음 표2는 평가 자료에 대한 특

성을 나타낸다. 제안 방법을 비교하기 위하여 세 명의 평가자가 문서를 수동으로 요약하여 요약방법으로 부터 요약된 요약결과와 비교하였다. 즉, 수동으로 요약한 요약문과 요약방법의 요약문에 대해서 성능을 비교 평가하였다.

성능 평가 방법으로는 정보검색에서 주로 사용되는 정확률(Precision), 재현율(Recall), F-measure등의 평가 척도를 이용하였다[13, 14, 15]. 이들에 대한 평가 척도는 다음 식(8), 식(9) 식(10)과 같다.

표 2.평가 자료의 특성
Table 2. Property of the test data set

문서의 속성	야후 코리아
문서의 수	400
30문장 이상인 문서의 수	74
문서당 평균 문장의 수	23
최소 문장의 수	3
최대 문장의 수	116

$$\text{Recall } (R) = \frac{|S_{man} \cap S_{sum}|}{|S_{man}|} \quad (8)$$

$$\text{Precision } (P) = \frac{|S_{man} \cap S_{sum}|}{|S_{sum}|} \quad (9)$$

$$\text{F-measure}(F) = \frac{2RP}{R+P} \quad (10)$$

여기서 S_{man}, S_{sum}은 각각 사람과 요약방법에 의하여 요약된 문장이다.

본 논문에서는 질의 확장이 요약 결과에 얼마나 영향을 미치는 가를 실험하였다. 비교는 비음수행렬 분해만을 이용한 질의 기반의 문서요약 방법을 기준으로 기본 질의(BQ), 워드넷에 의한 확장 질의(WQ), 워드넷 및 의사연관 피드백에 의한 확장질의를 사용한 방법(WPQ), 질의 분할과 연관 피드백을 이용한 방법(QS)에 대해여서 비교평가 하였다. 여기서 BQ는 저자가 이전에 제안한 방법[16]이고, WQ와 WPQ는 본 논문에서 제안한 방법이다. QS는 Han이 제한한 방법[6]이다.

다음 표3은 4가지 요약 방법에 대한 평가 결과 이다. 그림4는 표3을 막대그래프로 도식화한 것이다.

표 3. 평가 결과
Table 3. Result of evaluation

구분	QS	BQ	WQ	WPQ
average R	0.372	0.328	0.399	0.487
average P	0.321	0.29	0.341	0.374
average F	0.345	0.308	0.368	0.423

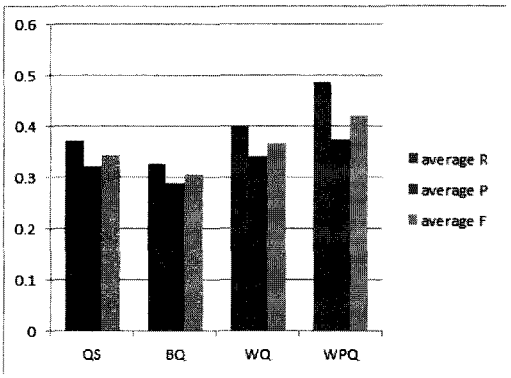

그림 4. 평가 결과의 도식
Fig. 4. Plot of evaluation result

표3에서 보는 것과 같이 야후 코리아 뉴스를 이용한 평가 결과에서는 제안 방법인 WPQ의 평균 재현율이 QS와 비교해서는 11.5%가, BQ와 비해서는 15.9%가, WQ와 비교해서는 8.8%가 더 높다. 평균 정확률은 WPQ가 QS와 비교해서는 5.3%가, BQ와 비해서는 8.4%가, WQ와 비교해서는 3.3%가 더 높다. 평균 F-measure는 WPQ가 QS와 비교해서는 7.8%가, BQ와 비해서는 11.5%가, WQ와 비교해서는 5.5%가 더 높다.

성능 평가 결과 제안방법인 WPQ가 가장 좋은 결과를 나타내며, 다음으로 WQ, QS, BQ 순으로 평가 되었다. 이는 단순히 문서 내부의 고유 의미 특징을 이용한 BQ 방법 보다는 질의를 분해하여서 확장한 QS방법이 더 좋은 성능을 나타내는 것을 알 수 있다. 또한 단 순한 질의 분해에 의한 질의 확장보다는 외부 지식을 이용한 질의 확장이나 WQ방법이 더 좋은 성능을 보이는 것을 알 수 있으며, 외부 지식과 의사연관 피드백을 사용한 제안 방법이 가장 좋은 성능을 보이는 것을 알 수 있다.

V. 결 론

본 논문은 의미특징과 워드넷 기반의 의사연관 피드백을 이용한 질의 기반 문서요약 방법을 제안하였다. 제안 방법은 비음수행렬 분해로부터 계산된 의미특징을 사용하여서 문서가 포함하고 있는 고유의 구조로부터 중요한 주제 및 세부 주제를 요약문에 잘 반영할 수 있다. 또한 외부 지식인 워드넷과 의사연관 피드백을 이용하여서 질의를 확장함으로써 사용자의 요구사항을 잘 반영한 요약문을 생성 할 수 있다. 실험결과 이전에 제안된 개인화된 문서 요약 방법에 비하여 더 좋은 평가 결과를 보였다.

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] 박선, 김철원, 임향석, "의미특징과 워드넷을 이용한 문서요약", 2010 한국통신학회춘계학술대회, 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] D. D. Lee, H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, pp.788-791, 1999.

[11] 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.

[12] Miller G. "WordNet: A lexical databased for english", CACM, 38(11), pp.39-41, 1995.

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

[14] S. Chakrabarti, "mining the web: Discovering Knowledge from Hypertext Data," Morgan Kaufmann Publishers, 2003.

[15] W. B. Frankes, B. Y. Ricardo, "Information Retrieval : Data Structure & Algorithms", Prentice-Hall, 1992.

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

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

저자소개

김철원(Chul-won Kim)

1997년 광운대학교 컴퓨터공학과 (공학박사)

1998년~현재 호남대학교 컴퓨터공학과 교수

※관심분야 : XML 응용, 멀티미디어 정보검색, 멀티미디어 컨텐츠 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- 전북전자정보 고급인력양성사업단 박사후과정
2010년 12월 ~ 현재 목포대학교 정보산업연구소 연구교수
※관심분야 : 정보검색. 데이터마이닝, 인공지능데이터베이스, 정보보안