논문 2012-49SP-2-7

비음수행렬분해와 위키피디아를 이용한 사용자기반의 문서요약

(User-based Document Summarization using Non-negative Matrix Factorization and Wikipedia)

박 선*, 정 민 아**, 이 성 로***

(Sun Park, Min-A Jeong, and Seong Ro Lee)

요 약

본 논문은 위키피디아의 외부지식을 이용하여 사용자의 질의를 확장하고, 확장된 질의와 문서집합의 내부구조를 표현하는 의미특징을 이용하여 문서를 요약하는 새로운 방법을 제안한다. 제안된 방법은 사용자의 초기 질의에 위키피디아 기반의 연관 피드백을 적용하여 사용자가 요구하는 요약문장을 추출할 수 있도록 질의를 확장하며, 비음수 분해된 문서의 의미특징을 이용함으로써 문서의 내부 구조를 잘 표현 할 수 있다. 확장된 질의와 의미특징을 이용하여 의미 있는 문장을 추출함으로써 사용자의 요구사항과 제안방법의 요약결과 사이의 의미적 차이를 감소시킨다. 실험결과 제안방법이 기존방법에 비해서 문서요약에 대해 더 좋은 성능을 보인다.

Abstract

In this paper, we proposes a new document summarization method using the expanded query by wikipedia and the semantic feature representing inherent structure of document set. The proposed method can expand the query from user's initial query using the relevance feedback based on wikipedia in order to reflect the user require. It can well represent the inherent structure of documents using the semantic feature by the non-negative matrix factorization (NMF). In addition, it can reduce the semantic gap between the user require and the result of document summarization to extract the meaningful sentences using the expanded query and semantic features. The experimental results demonstrate that the proposed method achieves better performance than the other methods to summary document.

Keywords : 사용자 기반 문서요약(user-based document summarization), 연관 피드백(relevance feedback), 위키피디아(wikipedai), 의미특징(semantic features), 비음수행렬분해(non-negative matrix factorization,).

I. 서 론

인터넷상의 기하급수적인 정보 증가와 통신기술의 발전은 사용자들이 접근할 수 있는 정보의 량을 점차 증가 시키고 있다. 이러한 정보의 증가는 인터넷 사용자가 원하는 정보를 효율적으로 검색할 수 있는 기술을 요구하고 있다. 특히, 검색기반 기술 중의 하나인 문서요약기술은 사용자가 원하는 정보에서 중요한 요점만을 요약함으로써 검색결과를 효율적으로 활용할 수 있도록

* 정회원-교신저자, 목포대학교 정보산업연구소
(Institute Research of Information Science and Engineering, Mokpo National University)
** 정회원, 목포대학교 컴퓨터공학과
(Department of Computer Engineering, Mokpo National University)
*** 정회원, 목포대학교 정보전자공학과
(Department of Information Electronic Engineering, Mokpo National University)
※ 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 대학중점연구소 지원사업으로 수행된 연구임(2010-0028295), 본 논문은 한국통신학회 지원에 의하여 연구되었음.
접수일자: 2011년12월6일, 수정완료일: 2011년12월27일

해준다. 또한 인터넷 휴대용 단말기의 사용 증가는 소형의 화면상에 대량을 정보를 효율적으로 표시할 수 있도록 하는 방법의 하나로 요약기술을 요구하고 있다.

문서요약은 문서에 포함된 핵심적인 주제를 나타내도록 문서의 량을 자동으로 줄이는 작업으로 현재까지 지속적으로 연구가 이루어지고 있다. 문서의 요약은 요약의 목적에 따라서 일반 문서요약(generic document summarization)과 사용자기반 문서요약(user-based document summarization)으로 구분할 수 있으며, 요약 대상에 따라서 단일 문서요약(single document summarization)과 다중 문서요약(multi-document summarization)으로 나눌 수 있다[1].

일반 문서요약은 문서의 구조를 유지하면서 문서전체를 파악할 수 있도록 문서에 포함된 중요한 주제들로 문서를 요약하는 방법이다. 사용자기반 문서요약은 사용자가 제시하는 질의에 부합되는 핵심적인 내용만으로 문서를 요약하는 방법이다. 또한 하나의 문서만을 대상으로 요약하면 단일문서 요약이라 하며, 하나의 주제가 시간의 경과에 따라서 여러 개의 문서를 구성하는 신문기사와 같은 다중 문서로부터 문서를 요약하면 다중 문서요약이라 한다[1]. 이외에도 요즘은 문서 자체의 정보를 요약하는 것 보다는 인터넷의 사용자 로그기록과 사용자 흥미에 관련된 특별한 정보를 유지하면서 문서를 요약하는 개인화 문서요약이 많이 연구되고 있다[2].

문서요약에 대한 접근방법은 크게 다섯 가지로 통계적 모델, 그래프기반 모델, 언어학 및 의미정보기반 방법, 외부자원기반 방법, 기타 복합 모델 방법이 있다.

문서요약에서 통계적 모델 접근방법은 문서에 포함된 용어의 출현 빈도를 이용한 방법으로 모든 접근방법의 가장 기본이 되는 방법이다. 문서요약에 쉽게 적용할 수 있으나 용어의 출현빈도만을 이용하기 때문에 용어와 문서간의 의미적 관계를 전혀 고려하지 않는 단점을 가지고 있다[1~2].

그래프기반 모델은 문서를 그래프로 변환하여 문서를 요약하는 방법으로, 그래프의 노드는 문서에 포함된 문장이나 주제를 나타내며, 그래프의 에지는 노드와 노드사이의 가중치를 나타낸다. 그래프기반 모델의 경우 문서를 그래프로 변환해야 하기 때문에 문서의 용량이 커질 수 록 계산이 복잡해진다[1, 3].

언어학 및 의미정보기반 방법은 언어의 언어학적 구조나 문서에 잠재되어 있는 의미특징들을 이용하여 문서를 요약하는 방법이다. 언어학적 방법의 경우 언어학적 분석을 위해서 복잡한 여러 단계를 수행하여야 하며, 의미정보 방법의 경우 문서 자체에 숨어 있는 구조적 특징을 이용하기 때문에 문서의 자체 구조에 많은 영향을 받는다[4~7].

외부자원기반 방법은 워드넷이나 위키피디아와 같은 외부자원을 이용하여 문서를 요약하는 방법이다. 온라인으로 전세계사용자에 의해서 실시간으로 갱신되는 위키피디아는 미리 완성된 영어 어휘목록의 워드넷과는 다르게 시간경과에 따른 정보의 유효성에 대한 제약을 덜 받으나, 대량의 위키피디아 자료를 전처리하는데 많은 자원이 소요된다[3, 8~9].

기타 복합 모델 방법은 문서요약의 성능향상을 위하여 여러 개의 모델을 복합적으로 적용하여 문서를 요약하는 방법이다[10~13].

본 논문에서 접근하는 문서요약 방법은 통계적 모델, 의미정보기반 방법, 외부자원기반 방법을 복합적으로 사용하는 복합모델 방법이다. 본 논문은 비음수 행렬분해로부터 추출된 의미특징과 위키피디아 기반의 연관피드백을 이용하여 문장을 추출하여서 문서를 요약하는 사용자 기반의 문서요약 방법을 제안한다.

본 논문에서 사용하는 비음수행렬분해는 Lee와 Seung이 제안한 방법으로 비음수로 된 원본 행렬을 두 개의 비음수 행렬로 분해하는 방법이다. 분해된 두개의 행렬은 원본 행렬의 내부 구조적 특성을 쉽게 파악할 수 있다. Lee는 분해된 첫번째 행렬을 비음수 의미 특징(NSF, non-negative semantic features)로 두 번째 행렬을 비음수 의미 변수(NSV, non-negative semantic variable)로 정의하였다[14~15].

위키피디아는 웹을 기반으로 한 다국 언어의 온라인 백과사전으로 전세계 온라인 사용자들이 광범위한 지식정보를 직접 생산 및 갱신하고 있으며, 위키피디아의 내용은 지금 이 시간에도 지속적으로 새로운 내용이 추가되고 있다. 또한 전문적인 사전이나 서적, 기사, 연구문헌 등을 참고하여 위키피디아 문서가 작성도기 때문에 문서의 개념과 내용에 대한 신뢰성을 가지고 있다[16~17].

연관피드백은 가장 널리 사용되는 질의 재작성 방법으로, 사용자에게 검색된 문서목록을 보여주고, 사용자가 그 문서들을 검사하여 어떤 문서가 연관된 문서인가를 판단하도록 하면서 질의를 확장하는 방법이다. 연관피드백은 사용자가 직접 개입하여 질의를 확장하는 연관피드백과 사용자의 개입 없이 자동으로 질의를 확장

하는 의사연관 피드백으로 구분된다[18~19].

다음은 본 논문에서 제안한 문서요약방법에 대한 설명이다. 사용자의 초기 질의와 위키피디아를 이용하여 질의를 확장하고, 확장된 질의와 비음수행렬분해에 의해 추출된 의미특징을 이용하여 문서를 요약한다. 제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 위키피디아를 초기 질의와 관련해서 질의 확장 시에만 사용하기 때문에 유사방법[3]에 비하여 자원을 효율적으로 사용하며, 사용자가 요구하는 요약정보에 적합한 질의로 확장할 수 있다. 둘째, 의미 특징(semantic feature)에 의해서 문서의 내부 특징(inherent feature)[14]을 요약 문서에 잘 반영할 수 있다. 마지막으로, 의미특징과 확장된 질의의 조합으로 문서요약의 질을 높일 수 있다.

본 논문의 구성은 다음과 같다. 제Ⅱ장에서는 문서요약에 대한 관련연구를, 제Ⅲ장은 제안된 요약방법에 대하여 설명하며, 제Ⅳ장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제Ⅴ장에서 결론은 맺는다.

Ⅱ. 관련 연구

다음은 본 논문의 제안방법과 관련된 최근의 문서요약 연구들이다.

Nastase는 위키피디아와 워드넷의 외부지식과 확장 활동에 의한 주제기반 다중문서요약 방법을 제안하였다. 이들의 방법은 외부지식을 이용하여 사용자의 질의를 확장하였으며, 확장된 질의와 관련된 문서를 문법적으로 연결된 용어의 그래프로 표현하여 문서를 요약하였다[3]. Ramanathan외 저자들은 언어에 독립적인 단일 문서요약 방법을 제안하였다. 이들의 방법은 문서의 문장과 위키피디아의 의미적 개념과 일치시켜 문서를 요약한다[8]. Ye와 저자들은 위키피디아 기반으로 새로운 문서 개념격자를 이용한 문서요약 방법을 제안하였다. 이들이 제안한 방법은 문장 간의 연관관계를 얻기 위하여 위키피디아의 내부 연결에 의한 위키피디아 개념을 소개하였으며, 위키 개념과 비 텍스트 특징을 이용하여 확장된 문서 개념 격자 모델(extended document concept lattice model)을 제안하였다[9]. Gong외 저자들은 위키피디아를 이용한 요약 시스템을 제안하였다. 이들이 제안한 방법은 위키피디아의 개념을 인식한 후에, 개념에 가중치를 주고 특징을 개산하여 요약문을 생성한다[20]. 위키피디아를 이용한 문서요약 방법은 문서요약 하기 전에 대량의 위피키디아 자료를 전처리하거나 학습함으로써 많은 자원을 소모해야 하는 문제를 가지고 있다.

Sanderson은 문서상의 중요한 문장과 사용자가 확장한 질의를 이용하여 문서를 요약 방법을 제안하였다[10]. Tombros와 Sanderson은 문서의 형식에 포함된 정보인 제목, 주제, 용어의 빈도 정보, 질의 등을 점수화 하여서 사용자가 보조 정보로 활용할 수 있는 문서 요약 방법을 제안하였다[11]. Varadarajan과 Hristidis는 질의와 가장 관련이 높은 문장과 의미 연관을 이용하여서 문서로부터 추출된 복합 주제를 적용하여서 질의에 특화된 문서 요약 방법을 제안하였다[12]. Han 외 저자는 질의 분해와 연관 피드백을 이용한 문서요약 방법을 제안하였다[13]. 연관피드백을 이용한 문서요약 방법들은 질의 정보가 부족하거나 사용자의 의도가 정확하기 반영되지 않을 때에 좋은 요약 결과를 보이지 않는 문제점을 가지고 있다.

본 논문의 저자들은 의미특징과 워드넷으로 질의를 확장하여서 문서를 요약하는 방법을 제안하였고[5], 개인화된 문서요약을 위하여서 의사연관 피드백과 비음수행렬분해를 이용한 요약방법을 제안하였으며[6], 비음수행렬분해와 의사연관 피드백을 이용한 질의 기반의 문서요약 방법을 제안하였다[7]. 의미특징을 이용한 문서요약 방법의 경우 원본문서의 특성에 요약 결과가 많은 영향을 받는다.

Ⅲ. 제안방법

본 논문에서 제안한 방법은 다음과 같이 세 단계로 구성된다. 첫 단계는 전처리 단계로 문서를 문장으로 분해해서 용어문장 빈도 행렬을 만든다. 두 번째 단계는 질의 확장 단계로 위키피디아와 연관 피드백을 이용한다. 마지막 단계는 문서요약 단계로 연관 피드백에 의해 확장된 질의와 비음수행렬분해된 문장집합의 의미특징을 이용하여서 문서를 요약한다.

3.1 전처리

문서요약에서 현재 많이 사용하고 있는 평가 자료는 대부분 영문 문서로 이루어져 있다. 이 때문에 본 논문에서는 주로 영문문서에 대한 문서요약 방법을 설명한다. 만약 한글문서에 본 논문에서 제안방법을 적용하려면 전처리 단계에서 한글 형태소 분석 도구[21]로 한글 용어를 추출하여 용어문장 빈도 행렬을 생성하여 적용

하면 된다.

전처리 단계는 문장추출, 불용어(stop-word) 제거, 어근(stemming)을 추출, 용어문장 빈도 행렬 생성 단계로 구성된다[19, 22]. 첫 번째인 문장추출 단계는 요약대상인 문서를 각각의 문장으로 분해하여 문장집합을 추출한다. 일반적으로 많이 사용하는 문장추출 방법은 정해진 크기로 문장을 추출하는 방법과 문장의 마침표를 기준으로 문장을 추출하는 방법이 있다. 일정 크기를 기준으로 문장을 추출 분해하는 방법은 문장이 중첩되거나 도중에 문장이 분할되어 정확한 문장의 의미를 전달할 수 없는 경우가 발생할 수 있다. 이 때문에 본 논문에서는 마침표를 기준으로 문장을 추출한다.

두 번째인 불용어 제거 단계에서는 영어의 관사나 대명사 같이 불필요한 용어들을 제거한다. 본 논문에서는 Rijsbergen의 불용어 목록[22]을 이용하여서 이 목록에서 정의하고 있는 용어만 제거한다. 세 번째인 어근추출 단계에서는 Porter의 어근추출 알고리즘[22]을 이용하여서 영어의 파생어들을 가장 중심이 되는 용어인 어근으로 변환한다.

마지막 단계인 용어문장 빈도 행렬생성 단계에서는 한 문장에 포함된 용어가 그 문장에서 출현한 발생 빈도를 이용하여 행렬을 구성한다. 본 논문에서 이용하는 용어문장 빈도 행렬은 다음과 같다. 생성된 용어문장 빈도 행렬 D는, j번째 문장 벡터 D_{*j}는 용어문장빈도 벡터 $D_{*j} = [\ t_{j1}, t_{j2}, \cdots, t_{jn}\ ]^T$로 표현되고, 여기서 T는 전치행렬을 나타내며, 문장 벡터 D_{*j}의 요소인 t_{ij}는 j번째 문장에서 i번째 용어를 나타낸다. 본 논문에서 행렬 X의 j번째 열벡터는 X_{*j}로, i번째 행벡터는 X_{i*}로, i번째 행과 j번째 열의 원소는 X_{ij}로 표시 한다.

3.2 질의 확장

본 논문의 질의 확장 단계는 위키피디아와 연관피드백을 이용하여 질의를 확장한다. 질의확장 단계는 질의확장을 위해서 위키피디아로부터 적합한 용어를 추출과 연관피드백에 의한 질의 확장으로 구성된다.

현재 영문 위키피디아의 경우 각각의 주제에 대한 사전 문서의 개수는 3,811,688개(2011년 12월 1일 시점)로 구성되어 있으며 문서의 개수를 지속적으로 증가하고 있다[16]. 이 때문에 위키피디아로 부터 질의 확장에 사용할 수 있는 용어는 너무 많아져서 포괄적인 의미를 포함하는 질의로 확장될 수 있다. 이 문제를 해결하기 위해서 본 논문에서는 질의 확장에 사용될 후보 용어로 부터 적합성을 검사하여 확장에 사용될 용어를 선택한다.

3.2.1 적합용어 추출

확장에 사용되는 후보용어 집합은 초기 질의에 서 단축이름으로 된 용어를 제외한 모든 포함 용어와 일치하는 위키피디아의 항목 제목, 동음이의어 목록의 용어와 설명을 이용하여서 추출한다.

여기서 단축이름 용어를 제외하는 이유는 위키피디아에는 단축이름에 대한 전체이름의 용어들이 너무 많이 존재하기 때문이다. 예를 들어 위키피디아에는 PCA(principal component analysis)의 용어에 대한 전체이름의 수는 사회, 문화, 과학 등의 전반적인 분야에 대하여 55개의 항목이 존재한다.

추출된 후보용어집합과 문장집합 간에 식(1)의 코사인 유사도를 계산하여 질의 확장에 적합한 용어들을 추출한다. 추출되는 용어들은 유사도가 높은 순으로 결정하여 다음 장에서 연관피드백을 이용하여 질의를 확장한다. 다음 표 1은 sentence를 초기 질의로 가지는 샘플 문서집합에 위키피디아를 이용한 후보용어집합과 유사도를 나타낸다.

다음 식(1)은 본 논문에서 유사도 계산에 사용하는 코사인 유사도이다[18~19, 22].

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

여기서 D_{*j}는 j번째 용어문장 빈도 벡터를 나타내고,

표 1. 후보용어집합과 유사도
Table 1. Candidate term set and similarity.

초기 질의	위키피디아의 유의어	유사도
sentence	linguistics, a grammatical unit of language	17.4%
	mathematical logic, a formula with no free variables	1.7%
	music, a particular type of musical phrase	1.2%
	law, a penalty applied to a person or entity found gilty of a criminal act	0.3%
	the Wire, the thirteenth episode of the Wire	0%
	a 12-century book of theology by Peter Lombard	0%
	The Life of MF Grimm, an autobiographical graphic novel by the MF Grimm, published by Verigo in 2007	0%

q는 초기질의를 나타내며, n은 용어의 수를 나타낸다.

3.2.2 연관 피드백

연관 피드백의 기본이 되는 방법은 Rocchio의 방법으로, 원래의 질의 벡터 $\vec{q}$에 연관된 문서의 문장집합 $D+$에 대응하는 벡터의 가중치 합을 단순히 더하고, 비연관 문서의 문장집합 $D-$의 가중치 합을 빼는 방법으로 식(2)와 같다[18~19].

$$\overrightarrow{q^{new}} = \alpha\vec{q} + \beta \sum_{\forall \vec{d_{*j}} \in D_+} \overrightarrow{t_{*j}} - \gamma \sum_{\forall \vec{d_{*j}} \in D_-} \overrightarrow{d_{*j}} \quad (2)$$

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

질의 확장은 이전 장에서 위키피디아와 유사도를 이용하여 추출한 적합 용어를 이용하여 사용자의 초기 질의를 확장한다. 본 논문에서는 의사연관 피드백을 변경하여 사용한다. 의사연관 피드백은 질의와 문장사이의 유사도를 계산하여서 질의와의 유사도가 상위 k개인 문장을 이용하여서 질의를 확장하는 방법이다. 그러나 위키피디아의 후보용어 집합에는 질의확장에 필요한 충분한 용어를 포함하고 있기 때문에 본 논문에서 각 항목 각각에 대하여 질의를 확장한다. 의사연관 피드백은 연과 피드백과 달리 비연관 문서를 판단 할 수 없기 때문에 식(3)과 같은 양의 연과 피드백을 사용한다.

$$\overrightarrow{q^{new}} = \vec{q} + \sum_{\forall t_{*j} \in D_+} t_{*j} \quad (3)$$

여기서, $\overrightarrow{q^{new}}$는 의사연관 피드백을 이용하여 새롭게 확장된 질의이고, $\vec{q}$는 초기 질의이다. t는 연관된 문서에 포함된 연관된 문장이다.

그림 1은 의사연관 피드백을 이용하여서 질의를 확장하는 예를 나타낸 것이다.

$$\underbrace{\begin{bmatrix}2\\0\\1\\1\\1\\0\\1\end{bmatrix}}_{\overrightarrow{q^{new}}} = \underbrace{\begin{bmatrix}1\\0\\0\\0\\0\\0\\0\end{bmatrix}}_{\vec{q}} + \underbrace{\begin{bmatrix}1\\0\\1\\1\\1\\0\\1\end{bmatrix}}_{t_{*1}}$$

그림 1. 의사연관 피드백을 이용한 질의 확장
Fig. 1. Query expansion by pseudo relevance feedback

3.3 문서요약

본 논문의 문서요약 단계에서는 확장된 질의와 문장집합의 의미특징을 이용하여 문서를 요약한다. 본 논문에서는 이전 저자들이 제안한 질의 기반의 문서요약 방법[23, 24]을 수정하여 사용한다. 문서요약 방법은 다음과 같다.

전처리된 용어문장 빈도 행렬 D를 식(4)와 식(5)를 이용하여 비음수 행렬 분해하여 두개의 의미특징 행렬을 계산한다. 식(1)을 이용하여 비음수 의미특징 열벡터와 확장된 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. 본 논문에서 요약문의 크기는 확장된 질의의 수로 조절할 수 있다. 즉, 유사도가 높은 확장 질의에 의해 추출된 문장을 요약문의 앞에 순차적으로 배치해 감으로써 미리 설정한 크기의 요약문을 생성할 수 있다.

본 논문에서 사용되는 비음수행렬분해 알고리즘은 LEE가 제안한 방법[14, 15]으로, 식(4)의 목표함수 J가 0에 가깝게 수렴 할 때까지 식(5)를 이용하여 행렬 W와 H의 값을 동시에 갱신한다.

$$J = \|A - WH\|^2 \quad (4)$$

식(5)의 목적은 행렬 A를 비음수 $m\times r$ 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.

$$H_{ij} \leftarrow H_{ij} \frac{(W^T A)_{ij}}{(W^T WH)_{ij}}, W_{ij} \leftarrow W_{ji} \frac{(AH^T)_{ji}}{(WHH^T)_{ji}} \quad (5)$$

Ⅳ. 실험 및 평가

본 논문에서 사용되는 실험 자료는 야후코리아 뉴스로 부터 20건의 질의에 대여서 각각의 질의 순위별로

20건의 기사 선택하였다. 제안 방법을 비교하기 위하여 세 명의 평가자가 문서를 수동으로 요약하여 요약방법으로 부터 요약된 요약결과와 비교하였다. 즉, 수동으로 요약한 요약문과 요약방법의 요약문에 대해서 성능을 비교 평가 하였다. 성능 평가 방법으로는 정보검색에서 주로 사용되는 정확률(Precision, P), 재현율(Recall, R), F-measure(F)등의 평가 척도를 이용하였다[18~19, 22]. 이들에 대한 평가 척도는 다음 식(6)과 같다.

$$(R)=\frac{|S_{man}\cap S_{sum}|}{|D|}\ ,\ (P)=\frac{|S_{man}\cap S_{sum}|}{|S_{sum}|}\ ,\ (F)=\frac{2RP}{R+P} \quad (6)$$

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. $|S_{man}\cap S_{sum}|$ 는 S_{man}와 S_{sum}의 교집합의 문장 수이다.

문서요약 결과를 비교평가하기 위하여 제안방법(WKEQ, Wikipedia Expanded Query)에 TFISF(Term Frequency Inverse Sentence Frequency), NMF(Non-negative Matrix Factorization), QS(Query Split), WDEQ(Wordnet Expanded Query)의 네 가지 방법을 비교하였다. 여기서 TFISF는 문서요약의 통계적 모델 방법으로 초기질의와 문장집합 간의 코사인 유사도를 이용하여 문서를 요약한다[1, 18, 22]. NMF는 초기질의와 비음수행렬분해를 이용한 방법으로 이전 저자들이 제안한 방법이다[23, 24]. QS는 Han이 제한한 방법으로 초기 질의와 변형된 의사연관 피드백을 이용하여 질의를 분할하여 확장하는 방법으로 문서를 요약하는 방법이다[13]. WDEQ는 역시 이전 저자들이 제안한 방법으로 워드넷과 의미특징을 이용하여 문서를 요약하는 방법이다[25]. WKEQ는 본 논문에서 제안한 방법이다.

본 논문에서는 질의 확장이 요약 결과에 얼마나 영향을 미치는 가를 다음 표2와 같이 비교평가 하였다. 표 2에서 보는 것과 같이 평가 결과에서는 제안 방법인 WKEQ의 평균 재현율이 TFISF와 비교해서는 17.7%가, QS와 비해서는 12.6%가, NMF와 비교해서는 6.1%가, WDEQ와 비교해서 1.1%가 더 높다. 평균 정확률은 WKEQ가 TFISF와 비교해서는 12.6%가, QS와 비해서는 9.6%가, NMF와 비교해서는 6.3%가, WDEQ와 비교해서 4.3%가 더 높다.

표 2. 평가 결과
Table 2. Result of evaluation.

구분	TFISF	QS	NMF	WDEQ	WKEQ
average P	0.321	0.372	0.437	0.487	0.498
average R	0.291	0.321	0.354	0.374	0.417
average F	0.305	0.345	0.3911	0.423	0.454

표 2에서 성능 평가 결과 제안방법인 WKEQ가 가장 좋은 결과를 보인다. 다음으로 WDEQ, NMF, QS, TFISF 순으로 평가 되었다. 이는 단순히 질의와 문장간의 유사도에 의한 문서를 요약하는 TFISF보다는 질의를 확장하여 문서를 요약하는 QS방법이 더 좋은 성능을 나타냄을 알 수 있으며, QS보다 문서 내부의 고유의미 특징을 이용한 NMF방법이 더 좋은 성능을 나타내는 것을 알 수 있다. 또한 단 순한 문서의 내부 특징을 이용하는 것보다 외부 정보인 WordNet과 의미특징을 이용한 방법이 더좋을 것을 알 수 있다. 특히 제안방법인 WKEQ는 위키피디아를 이용하여 질의를 확장한 후에 문서의 내부 특징을 나타내는 의미특징을 이용함으로써 시간 경과에 따른 정보의 변화를 반영함으로써 가장 좋은 성능을 보이는 것으로 판단된다.

V. 결 론

본 논문에서는 의미특징과 위키피디아 기반의 의사연관 피드백을 이용한 질의 기반 문서요약 방법을 제안하였다. 제안 방법은 위키피디아를 이용하여 질의를 확장함으로써 사용자 의도 변화를 문서요약 결과에 반영할 수 있으며, 문장집합으로 부터 비음수행렬 분해된 의미특징을 사용하여서 문서가 포함하고 있는 내부 특징으로부터 중요한 주제를 요약에 잘 반영할 수 있다. 또한 확장된 질의와 의미특징을 사용하여 문서를 요약함으로써 사용자의 요구사항을 잘 반영한 요약문을 생성 할 수 있다. 실험결과 이전에 제안된 질의기반의 연관피드백을 사용한 문서요약 방법에 비하여 더 좋은 평가 결과를 보였다.

참 고 문 헌

[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] V. Nastase, “Topic-Driven Multi-Document

Summarization with Encyclopedic Knowledge and Spreading Activation", In proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp.763-772, 2008.
[4] 박선, 김경준, 이진석, 이성로, "군집 주제의 유의어와 유사도를 이용한 문서군집 향상 방법", 한국전자공학회 논문지, 제48권 제5호, 2011.
[5] 박선, 김철원, 임향석, "의미특징과 워드넷을 이용한 문서요약", 2010 한국통신학회춘계학술대회, 2010.
[6] S. Park, D. U. An, "Automatic Query-based Personalized Summarization that uses Pseudo Relevance Feedback with NMF", In proceeding of ACM ICUIMC2010, 2010.
[7] S. Park, "User-focused Automatic Document Summarization using Non-negative Matrix Factorization and Pseudo Relevance Feedback", In proceeding of ICCEA2009, 2009.
[8] K. Ramanathan, Y. Sankarasubramaniam, N. Mathur, A. Gupta, "Document Summarization using Wikipedia", In proceedings of the First International Conference on HCI, 2009.
[9] S. Ye, T. S. Chua, J. Lu, "Summarization Definition from Wikipedia", In proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pp. 199-207, 2009.
[10] M., Sanderson, "Accurate user directed summarization from existing tools", In proceeding of the international conference on information and knowledge management, pp.45-51, 1998.
[11] A., Tombros, M., Sanderson, "Advantages of Query Biased summaries in Information Retrieval", In proceeding of ACM SIGIR, pp.2-10, 1998.
[12] R., Varadarajan, V., Hristidis, "A System for Query Specific Document Summarization", In proceeding of the CIKM, pp.622-631, 2006.
[13] 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.
[14] D. D. Lee, H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, pp.788-791, 1999.
[15] 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.
[16] 위키피디아, "www.wikipedia.com", 2011.
[17] Miller G. "WordNet: A lexical databased for english", CACM, 38(11), pp.39-41, 1995.
[18] B. Y. Ricardo, R. N. Berthier, "Moden Information Retrieval," ACM Press, 1999.
[19] S. Chakrabarti, "mining the web: Discovering Knowledge from Hypertext Data," Morgan Kaufmann Publishers, 2003.
[20] S. Gong, Y. Qu, S. Tian, "Summarization using Wikipedia", In proceedings of Text Analysis Conference 2010, 2010.
[21] 한경한, 남경완, "한국어 정보 처리 입문 : 컴퓨터가 우리말을 이해하려면", 커뮤니케이션북스, 2007.
[22] W. B. Frankes, B. Y. Ricardo, "Information Retrieval : Data Structure & Algorithms", Prentice-Hall, 1992.
[23] 박선, "의미 특징 행렬과 의미 가변행렬을 이용한 질의 기반의 문서 요약", 한국항행학회 논문지, 제12권, 제4호, 2008.
[24] 박선, 이주홍, "비음수 행렬 분해와 K-means를 이용한 주제기반의 다중문서요약", 한국정보과학회 논문지, 제35권, 제4호, 2008.
[25] 김철원, 박선, "의미특징과 워드넷 기반의 의사 연관 피드백을 사용한 질의 기반의 문서요약", 한국해양정보통신학회 논문지, 제15권 제7호, 2011.

저 자 소 개

박 선(정회원)-교신저자

1996년 전주대학교 전자계산학과 학사 졸업.

2001년 한남대학교 정보산업 대학원 정보통신학과 석사 졸업.

2007년 인하대학교 컴퓨터정보 공학과 박사 졸업.

2008년～2009년 호남대학교 컴퓨터공학과 전임강사.

2010년 전북대학교 전기전자정보인력양성사업단 박사후과정.

2011년～현재 목포대학교 정보산업연구소 연구교수.

<주관심분야 : 정보검색, 데이터마이닝, 데이터베이스, 해양생물 IT정보융합>

이 성 로(정회원)

1987년 고려대학교 전자공학과 졸업

1990년 한국과학기술원 전기및 전자공학과 석사

1996년 한국과학기술원 전기및 전자공학과 박사

1997년 9월～현재 목포대학교 공과대학 정보전자공학과 교수

<주관심분야 : 디지털통신시스템, 이동 및 위성통신시스템, USN/텔레미틱스응용분야, 임베디드시스템>

정 민 아(정회원)

1994년 2월 전남대학교 전산통계학과 석사

2002년 2월 전남대학교 전산통계학과 박사

2005년 3월～현재 목포대학교 컴퓨터공학과 부교수

<주관심분야 : 데이터베이스/데이터마이닝, 생체인식시스템, 무선통신응용분야(RFID, USN, 텔레메틱스), 임베디드시스템>