클라우드 기반의 용어가중치 재산정을 이용한 문서요약

박선 · 원정호 · 바트 · 양진호 · 최상길 · 추종윤 · 최호수 · 이성로
국립목포대학교
sunpark@mokpo.ac.kr,

Document Summarization using Term Reweighting based on Cloud

Park Sun · Won Jong Ho · Battsetsrg Ganbaatar · Yang Jin Ho · Choi Sang Gil · Choi Ho Su · Lee Sung Ro
Mokpo National Univ.

요 약

본 논문은 클라우드 기반의 연관피드백과 비음수행렬분해의 의미특징에 의한 용어 가중치 재 산정에 의한 문서요약 방법을 제안한다. 제안된 방법은 연관피드백을 이용하여 사용자의 의도를 문서요약 결과에 반연하며, 클라우드 기반의 비음수행렬분해의 의미특징으로 용어의 가중치를 재 산정함으로서 문장집합의 내부 특징을 잘 나타나기 때문에 문서요약의 질을 향상할 수 있다. 또한 클라우드 기반으로 대량의 빅데이터로부터 효율적으로 문서를 요약할 수 있다.

Ⅰ. 서 론

정보통신과 인터넷의 발전으로 정보의 량은 계속해서 폭발적으로 증가하다. 그러나 대량의 빅데이터 정보로부터 수집되는 정보들은 사용자가 원하는 정보를 쉽게 접근하지 못하도록 한다. 이 때문에 사용자들이 원하는 정보를 쉽게 사용자에게 보여줄 수 있는 요약에 대한 필요성을 증가시키고 있으며, 문서요약에 대한 많은 연구가 진행되고 있다.

문서요약은 요약의 목적에 따라서 다음과 같이 구분할 수 있다. 첫째, 일반요약으로 또는 포괄적 문서요약으로 문서의 전체 내용을 이해할 수 있도록 요약한다. 둘째, 질의기반의 문서요약으로 사용자의 질의에 관한 내용으로 문서를 요약한다. 셋째, 다중 문서요약으로 여래개의 문서로 부터 동일 주제에 관련된 내용을 요약하며, 단일문서요약으로 단일문서로 부터 관련 주제를 요약한다. 마지막으로 개인문서요약으로 인터넷 상의 사용자 로그와 사용자의 요구에 관련된 특별한 정보로 개인의 특성에 맞도록 요약한다[1]. 또한 문서요약에 적용되는 기술에 따라서 요약의 방법을 다음과 같이 분류할 수 있다. 통계적 방법, 그래프기반 방법, 언어학기반 방법, 의미정보기반 방법, 외부자원기반 방법, 기타 복합기반 방법을 이용한 문서요약이 있다[1-4].

본 논문은 클라우드 기반의 비음수 행렬분해로부터 추출된 의미특징과 의사연관피드백을 이용하여 문장을 추출하여서 문서를 요약하는 질의기반 문서요약 방법을 제안한다. 제안된 방법은 문서의 내부 구조를 나타내는 의미특징을 이용하여 용어의 가중치를 재 산정함으로써 요약의 질을 향상시킨다. 또한 의사연관피드백을 이용하여 사용자의 의도를 요약의 결과에 잘 반영 할 수 있다.

Ⅱ. 본론

본 논문에서 제안한 방법은 전처리, 질의확장, 클라우드 기반의 용어 가중치 재 산정, 문장추출에 의한 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. 세 번째 용어 가중치 계산 단계는 클라우드 기반의 비음수행렬분해된 의미특징을 이용하여 용어의 가중치를 재 산정한다. 마지막 단계는 문서요약 단계로 확장된 질의와 재 산정된 용어 가중치를 이용하여 문장을 추출하여서 문서를 요약한다.

2.1 전처리

전처리 단계는 문장 분해, 불용어(stop-word) 제거, 어근(stemming)을 추출, 용어문장 행렬생성 단계로 구성된다. 불용어는 Rijsbergen의 불용어

목록과 어휘 분석에 의한 불용어 제거방법을 이용한다. 또한, 어근 추출은 Porter 스테밍 알고리즘을 이용하여 어근을 추출한다[5].

2.2 질의확장

질의 확장 단계는 연관피드백 중에 의사연관 피드백을 이용하여 사용자의 질의를 확장한다. 의사연관 피드백은 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. 의사연관 피드백은 연과 피드백과는 다르게 비연관 문서를 판단 할 수 없기 때문에 식(1)과 같은 양의 연과 피드백을 사용한다[5]. 여기서 qnew는 확장된 질의, q는 질의 D는 문서의 집합이다.

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

2.3 클라우드 기반의 용어 가중치 계산

클라우드 기반의 용어 가중치 재 산정 단계는 다음 식(2)와 같이 가중치 행렬을 계산한다. 용어문장 행렬 D에 식(2)와 같이 식(3)의 가중치 행렬을 대입하여 용어에 대한 가중치를 계산한다.

$$\tilde{D} = \overline{W}D \quad (2)$$

여기서 $\overline{W}$는 용어에 대응하는 비음수행렬분해된 의미특징 값의 합을 가지는 대각행렬로 식(3)과 같으며, $\overline{w}$는 식 (4)에 의해서 계산된다. D는 용어문장 행렬이고, $\tilde{D}$는 용어에 대한 가중치 값이 계산된 용어문장 행렬이다.

$$\overline{W} = \begin{bmatrix} \overline{w_1} & 0 & \cdots & 0 \\ 0 & \overline{w_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \overline{w_n} \end{bmatrix} \quad (3)$$

용어 가중치 재 산정을 위한 의미특징의 합은 다음과 같으며, 여기서 w는 다음 비음수행렬분해 알고리즘에 의해 계산된다.

$$w_i = \sum_{k=1}^{l} w_{ik} \quad (4)$$

비음수 행렬 분해 알고리즘[6]은, 목표함수에 의한 유클리디안 거리가 0에 가깝게 수렴 할 때까지 의미 특징 행렬 W와 H의 값을 동시에 갱신한다[6]. 본 논문에서는 클라우드 기반으로 비음수 행렬 분해를 수행한다. 이를 위해서 Liu외 저자들이 제안한 맵리두스 기반의 GNMF 방법을 적용한다[7].

2.4 문장 추출에 의한 문서요약 단계

문장 추출에 의한 문서요약 단계는, 용어 가중치 재 산정 단계에서 용어의 가중치가 재 산정된 용어문장 행렬에 의사연관피드백에 의해서 확장된 질의를 이용한다. 이들 간의 코사인 유사도를 이용하여 유사도가 가장 높은 상위 문장들을 추출하여서 문서를 요약한다.

III. 결론

본 논문은 클라우드 기반의 용어 가중치의 재 산정과 의사연관 피드백의 확장된 질의를 이용하여 의미 있는 문장을 추출하여서 문서를 요약하는 질의 기반 문서요약 방법을 제안하였다. 제안 방법은 클라우드 기반의 비음수행렬분해로부터 유도된 용어의 가중치를 이용하여 의미 있는 주제를 요약에 더 잘 반영하였다. 또한 의사연관 피드백에 의한 확장된 질의와 재 산정된 용어 가중치를 이용하여 사용자가 사용자의 의도를 요약결과에 반영함으로써 요약문의 질을 높였다. 또한 클라우드를 이용하여 대량의 빅데이터로부터 사용자가 원하는 요약을 효율적으로 처리할 수 있다.

ACKNOWLEDGMENT

"본 연구는 미래창조과학부 및 정보통신산업진흥원의 IT융합 고급인력과정 지원사업의 연구결과로 수행되었음"(NIPA-2013-H0401-13-2006), 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2009-0093828)

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] 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] B. Y. Ricardo, R. N. Berthier, "Moden Information Retrieval," ACM Press, 1999.
[6] 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.
[7] C. Liu, H. C. Yang, J. Fan, L. W. He, Y. M. Wang, "Distributed Nonnegative Matrix Factorization for Web-Scale Dyadic Data Analysis on MapReduce", In proceeding of the IW3C2, pp.23-32, 2010.