비음수 행렬 인수분해를 이용한 질의 기반의 문서 요약*

박선° 이주홍 안찬민 박태수 김덕환[1]
인하대학교 컴퓨터정보공학과, [1]인하대학교 전자전기공학부
{parksun°, juhong, deokhwan[1]}@inha.ac.kr, {ahnch1, taesu}@datamining.inha.ac.kr

Query-Based Summarization using Non-negative Matrix Factorization

Sun Park°, Ju-Hong Lee, Chan-Min Ahn, Tae-Su Park, Deok-Hwan Kim[1]
Dept. of Computer Science and Information Engineering, Inha University
[1]School of Electronic and Electrical Engineering, Inha University

요 약

기존 질의기반의 문서요약은 질의와 문서간의 사전 학습으로 요약의 질을 높이거나, 문서의 고유 구조(inherent structure)를 반영하여 요약의 정확도를 높이기 위하여 문서를 그래프로 변환한다. 본 논문은 비음수 행렬 인수분해 (NMF, Non-negative Matrix Factorization)를 이용하여 질의 기반의 문서를 요약하는 새로운 방법을 제안하였다. 제안된 방법은 질의와 문서간에 사전학습이 필요 없다. 또한 문서를 그래프로 변형시키는 복잡한 처리 없이 NMF에 의해 얻어진 의미 특징(semantic feature)과 의미 변수(semantic variable)로 문서의 고유 구조를 반영하여 요약의 정확도를 높일 수 있다. 마지막으로 단순한 방법으로 문장을 쉽게 요약할 수 있다.

1. 서 론

인터넷 상의 가용 정보량이 증가하면서 특정 사용자의 관심사항에 중점을 두는 문서요약의 필요성이 점점 증가하고 있다. 문서 요약은 문서의 기본적인 내용을 유지하면서 문서의 양을 줄이는 작업이다. 문서의 요약은 제시되는 방법에 따라서 문서 내용전체를 요약하는 포괄적 요약(Generic Summary)과 사용자의 질의에 따라 질의와 관련 있는 내용만을 포함하는 질의 기반 요약(Query-based Summary)으로 나눌 수 있다[7].

질의 기반의 문서요약에 대한 최근 연구는 다음과 같다. Berger와 Mittal은 FAQ(frequently-asked question)를 이용하여 문서를 요약하는 방법을 제시하였다. 이들의 방법은 특정 주제의 질문과 답으로 구성된 FAQ문서를 훈련자료로 사용하여 훈련자료의 양을 줄였다. 이들의 방법은 사전에 미리 FAQ가 구성되어 있어야 하며, 훈련 자료에 따라서 문서 요약 결과가 영향을 받는다[1]. Bosma는 RST(rhetorical structure theory)를 이용하여 단일 문서를 그래프(graph)로 변형시켜 질의에 가장 접합한 답을 찾는다. 이 방법을 다중문서에 적용 할 때는 RST에 대한 광범위한 변경(extensive modification)이 필요하다[2]. Varadarajan과 Hristidis는 구조기반의 질의기반(query specific) 문서요약 방법을 제안하였다. 이들의 방법은 문서를 상호 연결된 문장의 집합으로 보고, 문서 그래프(document graph)를 만든다. 문서 그래프는 각각의 문장으로 노드(node)가 구성되며, 에지(edge)는 문장의 의미적 관계나 인접한 문장에 따라 가중치가 부여된다.

*본 연구는 대학 IT연구 센터 육성 지원사업의 연구결과로 수행되었음

질의의 키워드와 일치하는 문서그래프를 이용하여 문서를 요약한다 [11]. Sakurai와 Utsumi는 정보검색을 위한 질의 기반의 다중 문서요약 방법을 제안하였다. 이들이 제안한 방법은 먼저 질의와 가장 관련이 있는 문서로부터 문서요약의 핵심부분을 생성하고, 나머지 문서들로부터 요약을 보충할 부분을 생성하여 문서를 요약하였다. 이들의 방법은 긴 문서를 요약 할 때 효과적이나 요약 문장이 짧을 때는 좋은 성능을 보장하지 못한다[9]. Sassion은 주제기반의 다중문서요약 방법을 제안하였다.제안방법은 문장을 제거하여 문서를 요약하는 방법으로 사용자가 지정한 압축율까지 후보문장집합으로부터 문장을 제거하여 문서를 요약한다[10]

· 비음수 행렬 인수분해는 Lee와 Seung이 제안한 방법으로 인간이 객체를 인식할 때 객체의 부분정보의 조합으로 인식하는 것에 착안하여, 객체정보를 부분정보인 의미 특징(semantic feature)과 의미 변수(semantic variable)로 나누어 표현한다. 이러한 부분정보의 조합으로 전체 객체를 표현하는 방법으로, 대량의 정보를 효율적으로 표현 할 수 있는 방법이다[5,6].

본 논문에서는 의미 특징 행렬과 의미 변수 행렬을 이용하여 문서를 요약하는 새로운 질의기반 방법을 제안하였다. 제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 대량의 학습자료 및 사전 학습이 필요 없다. 둘째, 의미 특질(semantic feature)들이 문서내의 의미적으로 관련된 단어간의 군집을 이루기 때문에 문서고유의 구조를 쉽게 파악할 수 있고, 이를 이용하여 문서요약의

질을 높일 수 있다. 마지막으로 단일문서뿐만 아니라 다중문서에서도 간단한 방법으로 쉽게 문장을 요약 할 수 있다.

본 논문의 구성은 다음과 같다. 제2장에서는 제안한 문서요약방법을, 제3장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제4장에서 결론은 맺는다.

2. NMF에 의한 문서요약

본 장에서는 NMF를 이용하여 문서로부터 질의기반의 요약을 생성하는 방법을 제안한다. 그리고, 의미 특징 행렬과 질의 사이의 유사도를 이용하여 문장을 추출한다. 제안된 방법은 질의와 관련된 한 개 또는 여러 개의 문서에 대하여 전처리를 하고 문서요약을 찾는다. 전처리 단계는 주어진 문서를 각각의 문장으로 분해 후, 불용어(stopword) 제거, 어근추출(stemming), 가중치 계산으로 이루어진다. 이후 term-frequency 벡터를 생성하고 식(1)을 이용하여 가중치를 계산한다[3, 4].

벡터는 $T_i = [\ t_{1i},\ t_{2i},\ \cdots,\ t_{ni}\ ]^T$는 i번째 문장의 term-frequency 이다. 여기서 요소 t_{ji}는 i번째 절에서 j번째 term의 빈도이다. A는 m개의 term과 n개의 sentence로 이루어진 $m \times n$ 행렬이다. 요소 A_{ji}는 i번째 문장에서 j번째 term이 출현한 빈도의 가중치이다.

$$A_{ji} = L(j, i) \cdot G(j, i) \tag{1}$$

여기서 $L(j,i)$는 i번째 절에서 j번째 term을 위한 지역 가중치(local weight)이고, $G(j,i)$는 문서 전체에서 j번째 term을 위한 전역 가중치(global weight)로 다음 식(2), (3)과 같이 정의된다.

$$L(j,i) = t_{ji} \tag{2}$$

$$G(i) = \log(N/n(i)) \tag{3}$$

여기서, N은 문서에서 문장의 총 개수이다. $n(i)$는 i번째 term을 포함한 문장의 개수이다.

문서요약 단계는 비음수 의미 특징 행렬과 질의 사이의 유사도를 계산하여 유사도가 가장 높은 문장을 추출한다. 주어진 행렬 A를 비음수 행렬 인수분해 하여 얻어지는 비음수 의미 특징 행렬(NSFM, non-negative semantic feature matrix) W와 비음수 의미 변수 행렬(NSVM, non-negative variable matrix) H는 다음 식(4)와 같다.

$$A \approx WH \tag{4}$$

W는 $n \times r$ 행렬이고 H는 $r \times m$ 행렬이다. 여기서 r은 일반적으로 n이나 m보다 작게 선택하여 행렬 W나 행렬 H가 행렬 A보다 작게 한다. NMF는 $\|A-WH\|^2$가 최소화 될 때 까지 식(5)을 이용하며, W와 H 행렬 값을 동시에 갱신한다.

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T WH)_{a\mu}}, \quad W_{ia} \leftarrow W_{ia} \frac{(VH^T)_{ia}}{(WHH^T)_{ia}} \tag{5}$$

$W_{.l}$와 행렬 H의 요소 h_{lj}가 선형조합(linear combination)을 이루며 식(6)과 같다.

$$A_{.j} = \sum_{l=1}^{r} h_{lj} W_{.l} \tag{6}$$

$$sim(W_{.l}, \vec{q}) = \frac{W_{.l} \cdot \vec{q}}{|W_{.l}| \times |\vec{q}|} \tag{7}$$

비음수 의미 특징 행렬을 이용한 문서요약 방법은 다음과 같다.

1. 문서 D를 개개의 문장(sentences)으로 분해한다. 요약할 문장의 개수를 k라 한다.
2. 각각의 문장에 대한 불용어 제거 및 어근추출 한다.
3. 식(1)를 이용하여 term-frequency 벡터의 가중치를 계산하여 term-sentence 행렬 A를 구성한다.
4. 행렬 A에 식(5)를 적용하여 식(4)과 같은 비음수 행렬 W, H로 인수분해 한다.
5. 식(7)을 이용하여 행렬 W의 열 벡터들과 질의 간 유사도를 계산하여 가장 유사도가 높은 p번째 열 벡터 $W_{.p}$를 찾는다.
6. 행렬 H에서 p번째 행에 포함된 행 벡터 $H_{p.}$에서 가장 큰 요소 값을 가진 q열과 같은 열에 있는 행렬 A의 문장 벡터 $A_{.q}$에 대응되는 문장을 선택한다.
7. 만약 미리 정의된 k의 수 만큼 문장이 선택되면 알고리즘을 종료하고, 그렇지 않으면 5단계로 가서 다음으로 가장 큰 열 벡터 $W_{.p}$를 찾는다.

위의 4번째 단계에서 질의와 유사도가 가장 높은 열 벡터 $W_{.p}$는 질의와 연관이 있는 가장 중요한 의미 특징이다.

3. 실험 및 평가

본 논문에서 제안한 방법을 실험하기 위하여 ' 야후 코리아 뉴스' 의 기사를 실험자료로 사용하였다[12]. 10개의 질의에 대하여 각각 10건의 기사를 ' 야후 코리아 뉴스' 에서 검색하였다.

[표1] 야후 코리아 뉴스 자료의 특성표

문서속성	값
문서의 총개수	50
10문장이상으로 구성된 문서의 개수	42
평균 문장 개수	16
최소 문장 개수	2
최대 문장 개수	29

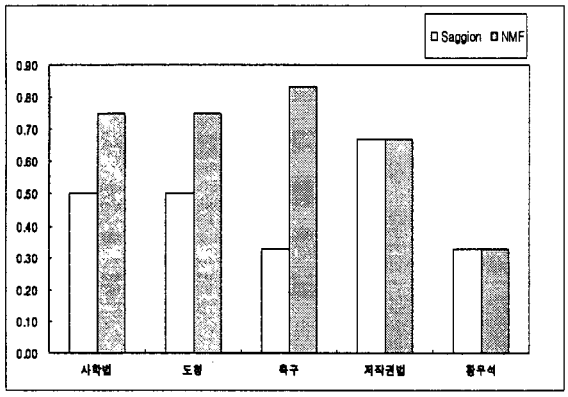
성능 평가는 질의 기반의 문서요약에서 주로 사용되는 정확률(Precision)을 이용하였다[3,8]. 정확률을 계산하기 위하여 50건의 기사로부터 질의와 관련된 문장을 수동으로 요약 하였다. 다음 표1은 실험자료에 대한 특성이다.

본 논문에서는 의미 특징 벡터 $W_{\cdot l}$와 질의 벡터 $\tilde{q}$와의 유사도는 두 벡터의 상관도로 구할 수 있으며, 이 상관도는 식(7)과 같이 두 벡터간 사이각의 코사인 값으로 정량화 할 수 있다[8].

평가척도는 다음 식 (8)이다.

$$P = \frac{|S_{man} \cap S_{sum}|}{|S_{sum}|} \quad (8)$$

여기서, S_{man}, S_{sum}는 각각 사람과 제안된 방법에 의해 선택된 문장이다. 다음 그림1은 Saggion의 방법[8]과 제안방법을 비교한 결과이다.



[그림 1] 실험 결과

4. 결론

본 논문은 NMF를 이용하여 질의 기반의 문서를 요약하는 새로운 방법을 제안하였다. 제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 사전 질의와 문서간의 학습과 학습자료가 필요 없다. 둘째, 의미 특질(semantic feature)을 이용하여 문서요약의 질을 높일 수 있다. 마지막으로 단일문서뿐만 아니라 다중문서에서도 쉽게 문장을 추출할 수 있다.

앞으로 제안 방법의 성능 향상을 위하여 다양한 종류의 가중치 및 전처리 방안에 대한 연구를 진행 시켜야 하며, 문서요약의 성능 향상을 위하여 비음수 행렬 W, H간의 관계에 대한 깊이 있는 연구가 진행 되어야 할 것이다.

5. References

1. Berger, A., Mittal, V. O.: Query-Relevant Summarization using FAQs. In Proceeding of the 38th Annual Meeting on Association for Computational Linguistics ACL' 00 (2000)
2. Bosma, W.: Query-based Summarization using Rhetorical Structure Theory. The Proceeding of CLIN (2004)
3. Frankes, W. B., Ricardo, B. Y.: Information Retrieval : Data Structure & Algorithms, Prentice-Hall (1992)
4. Kang, S. S.: Information Retrieval and Morpheme Analysis. HongReung Science Publishing Co. (2002)
5. Lee, D. D. and Seung, H. S.: Learning the parts of objects by non-negative matrix factorization, Nature, 401:788-791 (1999)
6. Lee, D. D. and Seung, H. S.: Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems, volume 13, pages 556-562 (2001)
7. Mani, I.: Automatic Summarization. John Benjamins Publishing Company (2001)
8. Ricardo, B. Y., Berthier, R. N.: Moden Information Retrieval, ACM Press (1999)
9. Sakurai, T., Utsumi, A.: Query-based Multidocument Summarization for Information Retrieval. The Proceeding of NTCIR-4 (2004)
10. Sassion. H.: Topic-based Summarization at DUC 2005. In Proceedings of the Document Understanding Conference 2005 (DUC' 05), (2005)
11. Varadarajan, R., Hristidis, V.: Structure-Based Query-Specific Document Summarization. (2005)
12. Http://kr.news.yahoo.com (2005)