NMF 와 코사인유사도를 이용한 질의 기반 문서요약*

박 선, 이주홍, 안찬민, 박태수, 송재원, 김덕환 [1]
인하대학교 컴퓨터정보공학과
[1] 인하대학교 전자전기공학부
e-mail : {parksun[O], juhong, deokhwan[1]}@inha.ac.kr, {ahnch1, taesu,sjw}@datamining.inha.ac.kr

Query-Based Text Summarization Using Cosine Similarity and NMF

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

요 약

인터넷의 발달로 인하여 정보의 양은 시간이 지날수록 폭발적으로 증가하고 있다. 이러한 방대한 정보로부터 정보검색시스템은 사용자에게 너무 많은 검색결과를 제시하여 사용자가 원하는 정보를 찾기 위해 너무 많은 시간을 소요하게 하는 정보의 과적재 문제가 있다. 질의 기반의 문서요약은 정보의 사용자가 원하는 정보의 검색시간을 줄임으로써 정보의 과적재 문제를 해결하는 방법으로서 점차 중요성이 증가하고 있다. 본 논문은 비음수 행렬 인수분해 (NMF, Non-negative Matrix Factorization)과 코사인 유사도를 이용하여 질의 기반의 문서를 요약하는 새로운 방법을 제안하였다. 제안된 방법은 질의와 문서 간에 사전학습이 필요 없다. 또한 문서를 그래프로 변형시키는 복잡한 처리 없이 NMF 에 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

문서 요약은 문서의 기본적인 내용을 유지하면서 문서의 양을 줄이는 작업이다. 문서요약의 접근방법에 따라서 문서내의 여러 통계값들을 이용하는 통계적 방법과 언어학적 지식에 기반한 방법으로 나눌 수 있다. 또한, 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. Berger 와 Mittal 은 FAQ(frequently-asked question)를 이용하여 문서를 요약하는 방법을 제시하였다. 이들의 방법은 기존의 비지도 학습이 많은 문서와 질의로 훈련자료(training data)를 구성하는데 비하여 특정 주제의 질문과 답으로 구성된 FAQ 문서를 훈련자료로 사용하여 훈련자료의 양을 줄였다. 이들의 방법은 사전에 미리 FAQ 가 구성되어 있어야 하며, 훈련 자료

* 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT 연구 센터 육성.지원사업의 연구결과로 수행되었음.

에 따라서 문서 요약 결과가 영향을 받는다[1]. Bosma 는 RST(rhetorical structure theory)를 이용하여 단일 문서를 그래프(graph)로 변형시켜 질의에 가장 접합한 답을 찾는다. 이 방법을 다중문서에 적용 할 때는 RST 에 대한 광범위한 변경(extensive modification)이 필요하다[2]. Varadarajan 과 Hristidis 는 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.

질의의 키워드와 일치하는 문서그래프를 이용하여 문서를 요약한다 [11]. Sakurai 와 Utsumi 는 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]. Sassion 은 주제기반의 다중문서요약 방법을 제안하였다.제안방법은 문장을 제거하여 문서를 요약하는 방법으로 사용자가 지정한 압축율까지 후보 문장집합으로부터 문장을 제거하여 문서를 요약한다 [10]

비음수 행렬 인수분해는 Lee 와 Seung 이 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].

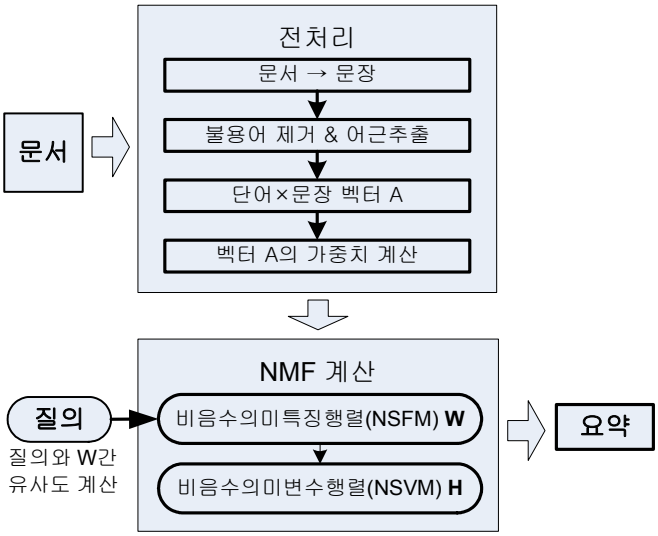
본 논문에서는 NMF 와 코사인 유사도를 이용하여 문서를 요약하는 새로운 질의기반 방법을 제안하였다. 제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 대량의 학습자료 및 사전 학습이 필요 없다. 둘째, 의미특질(semantic feature)들이 문서내의 의미적으로 관련된 단어간의 군집을 이루기 때문에 문서고유의 구조를 쉽게 파악할 수 있고, 이를 이용하여 문서요약의

본 논문의 구성은 다음과 같다. 제 2 장에서는 제안한 문서요약방법을, 제 3 장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제 4 장에서 결론은 맺는다.

2. NMF 에 의한 문서요약

본 장에서는 NMF 를 이용하여 문서로부터 질의기반의 요약을 생성하는 방법을 제안한다. 제안하는 방법은 NMF 에 의하여 계산된 의미 특징 행렬과 질의 사이의 코사인 유사도를 이용하여 문장을 추출한다.

제안방법은 전처리 단계와 문서요약 단계로 구성된다. 다음 (그림 1)은 제안방법에 의한 문서요약 시스템의 개요이다.



(그림 1) NMF 와 코사인유사도를 이용한 질의 기반의 문서요약 시스템 개요

2.1 전처리

전처리 단계는 주어진 문서를 각각의 문장으로 분해 후, 불용어(stopword) 제거, 어근추출(stemming), 가중치 계산으로 이루어진다. 이후 용어-빈도(term-frequency) 벡터를 생성하고 식(1)을 이용하여 가중치를 계산한다[3, 4].

벡터는 $T_i = [\, t_{1i}, t_{2i}, \cdots, t_{ni} \,]^T$ 는 i 번째 문장의 용어의 빈도이다. 여기서 요소 t_{ji} 는 i 번째 절에서 출현한 j 번째 용어의 빈도이다. $\boldsymbol{A}$ 는 m 개의 용어와 n 개의 문장(sentence)으로 이루어진 $m \times n$ 행렬이다. 요소 A_{ji} 는 i 번째 문장에서 j 번째 용어가 출현한 빈도의 가중치이다.

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

여기서 $L(j,i)$는 i 번째 절에서 j 번째 용어를 위한 지역 가중치(local weight)이고, $G(j,i)$는 문서 전체에서 j 번째 용어를 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 i 번째 용어를 포함한 문장의 개수이다.

2.2 NMF 와 코사인유사도에 의한 문서요약

문서요약 단계는 비음수 의미 특징 행렬과 질의 사이의 유사도를 계산하여 유사도가 가장 높은 문장을 추출한다. 주어진 행렬 $\boldsymbol{A}$를 비음수 행렬 인수분해 하여 얻어지는 비음수 의미 특징 행렬(NSFM, non-negative semantic feature matrix) $\boldsymbol{W}$와 비음수 의미 변수 행렬(NSVM, non-negative variable matrix) $\boldsymbol{H}$는 다음 식(4)와 같다.

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

W는 $n \times r$ 행렬이고 **H**는 $r \times m$ 행렬이다. 여기서 r은 일반적으로 n 이나 m 보다 작게 선택하여 행렬 $\boldsymbol{W}$나 행렬 $\boldsymbol{H}$가 행렬 $\boldsymbol{A}$ 보다 작게 한다. NMF 는 $\|A-WH\|^2$ 가 최소화 될 때 까지 식(5), (6)을 이용하며, $\boldsymbol{W}$와 $\boldsymbol{H}$ 행렬 값을 동시에 갱신한다.

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T V)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}} \tag{6}$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(VH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \tag{5}$$

$W_{\cdot l}$ 와 행렬 $\boldsymbol{H}$ 의 요소 h_{lj} 가 선형조합(linear combination)을 이루며 식(6)과 같다.

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

의미 특징 벡터와 질의와의 유사도를 구하는 식은 (8)과 같다. 여기서 w_{ij} 는 j 번째 r 계수 에서의 i 번째 의미 특징인 요소이고 ($w_{ij} \geq 0$), w_{iq} 는 i 번째 의미 특징 요소와 일치하는 q 번째 질의의 용어이다 ($w_{iq} \geq 0$). m은 r열 벡터의 요소들의 수로, 벡터 $\vec{q}$ = (w_{1q}, w_{2q}, …, w_{mq})로 나타낸다[7].

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

비음수 의미 특징 행렬을 이용한 문서요약 방법은 다음과 같다.

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

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

3. 실험 및 평가

본 논문에서 제안한 방법을 실험하기 위하여 '야후 코리아 뉴스'의 100 건의 기사를 실험자료로 사용하였다[12].

<표 1> 야후 코리아 뉴스 자료의 특성표

문서속성	값
문서의 총개수	50
10 문장이상으로 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 50 건의 기사로부터 질의와 관련된 문장을 수동으로 요약 하였다. 다음 표 1 은 실험자료에 대한 특성이다.

본 논문에서는 의미 특징 벡터 $W_{\cdot l}$ 와 질의 벡터 $\vec{q}$ 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.

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

여기서, S_{man}, S_{sum} 는 각각 사람과 제안된 방법에 의해 선택된 문장이다. 다음 그림 1 은 Saggion 의 방법[8]과 제안방법을 비교한 결과이다.

<표 2> 실험 결과

구분	사학법	도청	황우석	축구	저작권
SAggion	0.50	0.50	0.33	0.33	0.67
NMF	0.75	0.75	0.33	0.83	0.67

4. 결론

본 논문은 NMF 와 코사인 유사도를 이용하여 질의 기반의 문서를 요약하는 새로운 방법을 제안하였다.

제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 의미 특질(semantic feature)을 이용하여 문서요약의 질을 높일 수 있다. 둘째, 사전 질의와 문서간의 학습과 학습자료가 필요 없다. 마지막으로 단순한 방법으로 문서를 쉽게 요약할 수 있다.

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)