비음수 행렬 인수분해를 이용한 일반적 문서 요약*

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

Generic Text Summarization Using Non-negative Matrix Factorization

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

요 약

본 논문은 비음수 행렬 인수분해(NMF, non-negative matrix factorization)를 이용하여 문장을 추출하여 문서를 요약하는 새로운 방법을 제안하였다. 제안된 방법은 문장추출에 사용되는 의미 특징(semantic feature)이 비 음수 값을 갖기 때문에 잠재의미분석에 비해 문서의 내용을 정확하게 요약한다. 또한, 적은 계산비용을 통하여 쉽게 요약 문장을 추출할 수 있는 장점을 갖는다.

1. 서 론

인터넷의 발달로 인하여 정보의 양은 폭발적으로 증가하고 있다. 이러한 대량의 정보로부터 필요한 정보를 정확하게 찾는 정보검색의 중요성이 점차 증가하고 있다. 그러나 인터넷 상의 방대한 정보들에서 원하는 정보를 찾기 위해서 정보검색시스템이 주로 이용된다. 정보검색시스템이 제시하는 검색결과를 사용자가 자세히 확인하기에는 정보의 양이 너무 많다. 이러한 문제를 해결하기 위하여 일반적으로 정보검색시스템들은 문서의 서론부분만 제시하여 해결하는 경향이 있다. 그러나 이와 같은 단편적인 정보만을 이용하여 사용자가 원하는 정보의 적합성을 판단하기에는 부족하다. 자동 문서요약은 사용자가 원하는 정보의 검색시간을 줄임으로써 정보의 과적재 문제를 해결하는 방법으로 점차 중요성이 증가하고 있다[14].

문서 요약은 문서의 기본적인 내용을 유지하면서 문서의 량을 줄이는 작업이다[13]. 문서의 요약은 제시된 방법에 따라서 문서 내용전체를 요약하는 일반적 요약(Generic Summary)과 사용자의 질의에 따라 질의에 관련 있는 내용만을 포함하는 질의 중심 요약(Query-focused Summary)으로 나눌 수 있다[6, 11].

문장추출을 이용한 문서요약의 최근 연구는 다음과 같다. 잠재의미분석(LSA, latent semantic Analysis)을 이용한 방법으로, Gong 과 Liu 는 문서를 요약하였고[4], Sum 과 Shen 은 웹 페이지를 요약하였다[8]. 잠재의미분석은 문장을 선택하기 위하여 복합(multiple) 고유벡터(singular vector)의 구성요소(component)를 이용한다. 고유벡터에 일치하는 고유 값(singular value)은 양수와 음수의 구성요소를 갖고, 이러한 의미가 작은 고유벡터의 구성요소 값에 의해서 추출 문장의 순위가 구성될 수 있다[11].

문서의 주제(topic)을 이용한 방법으로, Nomoto 와 Matsumoto 는 변형된 k-means 를 이용하여 문서에서

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

다양한 주제를 찾은 후, 각 주제에 일치하는 문장을 선택하여 문서를 요약 하였다[7]. Zha 는 문장과 용어(terms)로부터 특성점수(saliency scores)를 계산하고, 이를 이용하여 문장들을 주제그룹(topical groups)들로 군집하여 문서를 요약하였다[11]. Harabagiu 와 Lacatusu 는 다중문서(multi-document) 요약을 위하여 다섯 개의 주제를 이용한 문서요약방법에 대하여 비교 평가하였다[5]. 이들 방법은 먼저 주제를 추출 한 다음 문서를 요약하기 때문에 계산비용이 많이 드는 단점이 있다.

비음수 행렬 인수분해(NMF, non-negative matrix factorization)는 Lee 와 Seung 이 제안한 방법으로 다변량 자료를 유용하게 분해하는 알고리즘이다[2,3,10].

본 논문은 비음수 행렬 인수분해를 이용하여 문장을 추출하여 문서를 요약하는 새로운 방법을 제안하였다.

제안된 방법은 다음과 같은 장점을 갖는다.

(1) NMF 에 의해 찾아지는 의미 특징(semantic feature)들이 비 음수 값을 갖기 때문에 잠재의미분석에 비해 의미 있는 문서요약 결과를 갖는다.

(2) 적은 계산비용으로 쉽게 문장을 추출할 수 있다.

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. 제안방법

본 장에서는 NMF 를 기반으로 문장을 추출하여 일반적 문서요약을 할 수 있는 방법을 제안한다. 제안방법은 전처리 단계와 문서요약 단계로 이루어진다. 전처리 단계는 주어진 문서를 각각의 문장으로 분해 후, 불용어(stopword) 제거, 어근추출(stemming), 가중치 계산으로 이루어진다.

2.1 전처리

본 논문에서는 Rijsbergen 의 불용어 목록을 이용한 불용어 제거 및 Porter 스테밍 알고리즘을 이용하여 스테밍 하였다[9]. 이후 term-frequency 벡터를 생성하고 식(1)을 이용하여 가중치를 계산하였다[4].

벡터는 $T_i = [\, t_{1i}, t_{2i}, \ldots, t_{ni} \,]^T$ 는 i 번째 문장(sentences)의 term-frequency 이다. 여기서 요소(element) t_{ji} 는 i 번째 문장에서 출현한 j 번째 term 의 빈도이다. i 번째 문장 A_i 는 가중치가 부여된 term-frequency 벡터 $A_i = [\, a_{1i}, a_{2i}, \ldots, a_{ni} \,]^T$ 로 표현되고, 벡터 A_i 요인 a_{ji} 는 식(1)와 같이 정의 된다.

$$a_{ji} = L(t_{ji}) \cdot G(t_{ji}) \tag{1}$$

여기서 $L(t_{ji})$는 i 번째 문장에서 j 번째 term 을 위한 지역 가중치(local weight)이고, $G(t_{ji})$는 문서 전체에서 j 번째 term 을 위한 전역 가중치(global weight)로 다음 식(2), (3)과 같이 정의된다.

$$L(i) = tf(i) \tag{2}$$

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

여기서, $tf(i)$는 문장에서 i 번째 term 이 출현한 빈도, N 은 문서에서 문장의 총 개수이다. $n(i)$는 i 번째 term 을 포함한 문장의 개수이다.

2.2 NMF 를 이용한 문서요약

NMF 를 적용한 문서요약 단계는 다음과 같다. 문서에서 총 m 개의 term 과 n 개의 sentence 로 이루어진 $m \times n$ 행렬 **A** 는 요소 A_{ji} 로 구성되며, 요소 A_{ji} 는 i 번째 문장과 j 번째 term 의 가중치가 부여된 빈도이다. 주어진 행렬 **A** 를 비음수 행렬 인수분해 하면 다음 식(4)과 같다.

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

여기서, **A** 는 $m \times n$ 행렬이고, $m \times r$ 행렬 **W** 와 $r \times n$ 행렬 **H** 는 행렬 **A** 로부터 근사값으로 인수분해 된 행렬이다. r 은 의미특징의 개수로 지정된다. **W**, **H** 를 구하기 위해서 $\|A - WH\|^2$ 가 수렴 할 때까지 식(5), (6)을 이용하며, **W** 와 **H** 행렬 값을 동시에 갱신 한다.

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T V)_{\alpha\mu}}{(W^T W H)_{\alpha\mu}} \tag{5}$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(V H^T)_{i\alpha}}{(W H H^T)_{i\alpha}} \tag{6}$$

행렬 **A** 의 j 번째 열벡터 $A_{\cdot j}$ 는 행렬 **W** 의 l 번째 열벡터 $W_{\cdot l}$ 와 행렬 **H** 의 요소 h_{kj} 가 선형조합(linear combination)을 이루며 식(7)과 같다.

$$A_{\bullet j} = \sum_{l=1}^{r} h_{kj} W_{\bullet l} \tag{7}$$

비음수 행렬 W 와 H 에 대한 의미적 해석은 다음과 같다. 모든 의미 변수는 각 문장을 표현 할 수 있다. 직관적으로 단지 하나의 주제나 모든 주제 보다는 광범위하게 배열된 주제와 연관된 작은 부 집합이 각 문장을 더욱 의미 있게 한다. 각각의 의미 특징은 NMF 에 의해 의미적으로 관련 있는 용어로 군집화된다. 의미적으로 관련된 군집이 의미 특징으로 결합하여, 문맥상에서 동음이의어를 구별하는데 NMF 를 사용한다[8].

비음수 행렬 인수분해를 이용한 문서요약 방법은 다음과 같다.

1. 문서 D 를 개개의 문장(sentences)으로 분해하고, 추출문장의 개수 k 를 지정한다.
2. 각각의 문장에 대한 불용어 제거 및 어근추출 후, 식(1),(2),(3)을 이용하여 term-frequency 벡터의 가중치 계산, term-sentence 행렬 **A** 를 구성한다.
3. 행렬 **A** 에 식(5)과 식(6)를 적용하여 식(4)과 같은 비음수 행렬 **W**, **H** 로 인수분해 한다.
4. 행렬 **H** 에서 p 번째 행에 포함된 행 벡터 H_p. 의 요소의 합 $\sum_{i=1}^{n} H_{pi}$ 을 각각 행 벡터 별로 계산한다. 행 벡터의 요소의 합 값이 큰 순서로 k 개의 행 벡터 H_p. 를 선택한다.
5. 선택된 k 개의 행 벡터 각각에서, 행에서 가장 큰 요소 값을 가진 q 열과 같은 열에 있는 행렬 **A** 의 문장 벡터 $A_{\cdot q}$ 에 대응되는 문장을 선택한다.

3. 실험 및 평가

본 논문에서 제안한 방법을 실험하기 위하여 Reuters-21578 [12] 컬렉션 중 129 건의 기사를 무작위로 선택하여 테스트 자료로 사용하였다. 이들 중 실제 평가에 사용된 문서는 81 건으로 세 명의 평가자에 의해 수동으로 요약 되었다. 평가자는 문서당 평균 3.8 문장을 선택하였으며, 이중 평가자 두 명 이상이 공통으로 선택된 문장을 포함하는 81 건의 문서를 실험에 사용하였다. 성능 평가는 문선요약에서 주로 사용되는 정확률(Precision), 재현율(Recall), F-measure 를 이용하였다[3,9]. 평가척도는 다음 식 (7), (8), (9) 이다.

$$R = \frac{|S_{man} \cap S_{sum}|}{|S_{man}|} \tag{7}$$

$$P = \frac{|S_{man} \cap S_{sum}|}{|S_{sum}|} \tag{8}$$

$$F = \frac{2RP}{R+P} \tag{9}$$

여기서, S_{man}, S_{sum} 는 각각 사람과 제안된 방법에 의해 선택된 문장이다.

본 논문에서는 LSI 방법을 제안된 방법과 식(9)를 적용하여 실험평가 하였다[4]. LSI 와 제안방법에 k 의 값은 평가자에 의해 추출된 문서의 평균 개수인 3.8 을 기준하였다. 다음 <표 1>는 실험 방법을 비교평가한 결과이다.

<표 1> 각 실험 방법의 비교

구 분	LSI	NMF
정확률	0.445	0.700
재현율	0.340	0.650
F-measure	0.405	0.680

<표 1>의 F 값에 의하면 제안한 방법이 LSI 를 사용한 방법 보다는 약 0.275 성능이 우수함을 알 수 있다.

실험에서 보듯이 제안된 방법은 LSI 에 비하여 좋은 성능을 보인다. 제안방법이 비음수 값과 부분정보를 이용하는 인간의 인식과정[2]과 유사한 과정으로 문서를 처리하기 때문이다. 또한 2 장의 NMF 식 (4), (5), (6) 에서 보는 것과 같이 적은 비용을 통해 문장을 추출 할 수 있다.

4. 결론

본 논문은 문서요약을 위해 비음수 행렬 인수분해 (NMF, non-negative matrix factorization)로 문장을 추출하는 새로운 방법을 제안하였다. 제안된 방법은 다음과 같은 장점을 갖는다. 실험 결과 제안방법이 잠재의미분석방법에 비하여 더 좋은 문서요약 결과를 갖는다. 그리고 적은 계산비용을 통하여 쉽게 문장을 추출할 수 있다.

앞으로 제안 방법의 성능 향상을 위하여 다양한 종류의 가중치 및 전처리 방안에 대한 연구를 진행시켜야 하며, 문서의 크기에 따른 추출 문장 k 의 개수를 자동으로 선택할 수 있는 방법에 대한 연구가 진행 되어야 할 것이다.

참고문헌

[1] Chakrabarti, S. “Mining the Web : Discovering Knowledge from Hypertext Data”, Morgan Kaufmann (2003)

[2] Lee, D. D., Seung, H. S. “Learning the parts of objects by non-negative matrix factorization”, Nature (1999) 401:788-791,

[3] Lee, D. D., Seung, H. S. “Algorithms for non-negative matrix factorization”, In Advances in Neural Information Processing Systems, volume 13 (2001) 556-562

[4] Gong, Y., Liu, X. “Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis”, In proceeding of ACM SIGIR’01 (2001) 19-25

[5] Harabagiu, Sanda. “Finley Lacatusu. Topic Themes for Multi-Document Summarization”, In proceeding of ACM SIGIR’05 (2005) 202-209

[6] Marcu, D. “The automatic construction of large-scale corpora for summarization research”, In proceeding of ACM SIGIR’99 (1999) 137-144

[7] Nomoto, T. “Yuji Matsumoto. A New Approach to Unsupervised Text Summarization”, In proceeding of ACM SIGIR’01 (2001) 26-34

[8] Sum, J. T., Shen, D., Zeng, H. J. “Qiang Yang. Yuchang Lu. Zheng Chen. Web-Page Summarization Using Clickthrough Data”, In proceeding of ACM SIGIR’05 (2005) 194-201

[9] Frankes, W. B. Baeza-Yaes. R. “Information Retrieval : Data Structure & Algorithms”, Prentice-Hall (1992)

[10] Xu, W., Liu, X. and Gong, Y. “Document clustering based on non-negative matrix factorization”, In ACM SIGIR, Toronto, Canada (2003)

[11] Zha, Hongyuan. “Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering”, In proceeding of ACM SIGIR’02 (2002) 113-120

[12] http://kdd.ics.uci.edu/databases/reuters21578/reuters21-578.html (2005)

[13] Mani, I., Maybury, M. T. “Advances in Automatic Text”, The MIT Press (1999)

[14] Tombros. A., Sanderson. M. “Advantages of Query Biased Summaries in Information Retrival”, In the Proceeding of ACM SIGIR’98 (1998)