비음수 의미 가변 행렬을 기반으로 한 자동 포괄적 문서 요약*

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

Automatic Generic Summarization Based on Non-negative Semantic Variable Matrix

Sun Park[O], 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

인터넷의 급속한 확산과 대량 정보의 이동은 문서의 요약을 더욱 필요로 하고 있다. 본 논문은 비음수 행렬 인수분해로(NMF, non-negative matrix factorization) 얻어진 비음수 의미 가변 행렬(NSVM, non-negative semantic variable matrix)을 이용하여 자동으로 포괄적 문서요약 하는 새로운 방법을 제안하였다. 제안된 방법은 인간의 인식 과정과 유사한 비음수 제약을 사용한다. 이 결과 잠재의미색인에 비해 더욱 의미 있는 문장을 선택하여 문서를 요약할 수 있다. 또한, 비지도 학습에 의한 문서요약으로 사전 전문가에 의한 학습문장이 필요 없으며, 적은 계산비용을 통하여 쉽게 문장을 추출할 수 있는 장점을 갖는다.

1. 서 론

문서 요약은 문서의 기본적인 내용을 유지하면서 문서의 양을 줄이는 작업이다[11]. 문서의 요약은 제시된 방법에 따라서 문서 내용전체를 요약하는 포괄적 요약(Generic Summary)와 사용자의 질의에 따라 질의와 관련 있는 내용만을 포함하는 질의 중심 요약(Query-focused Summary)으로 나눌 수 있다[10, 13].

현재의 자동 문서요약 방법은 지도 학습(supervised approaches)과 비지도 학습(unsupervised approaches) 방법이 있다. 비지도 학습 방법은 사람에 의해 만들어진 요약이 없이 관련 매개변수를 결정할 수 있는데 비하여, 지도 학습 방법은 요약 방법에 사용되는 매개변수나 특징 등을 추출하거나, 사람이 만든 요약문을 필요로 한다.

지도학습방법에 의한 최근의 문서요약은 다음과 같다. Chuang과 Yang은 부분 문장을 추출하여 자동으로 문서를 요약하는 방법을 제안하였다. 이들의 방법은 특징벡터를 기반으로 중요한 문장의 구획을 추출하기 위해서 요약기(summarizer)를 훈련하는 비지도 학습 알고리즘을 이용하였다[3]. Amini와 Gallinari는 준지도 학습 방법(semi-supervised learning)의 알고리즘을 이용한 문서 요약 방법을 제안하였다. 이들의 방법은 대량의 레이블이 없는 자료를 처리하기 위해 소량의 레이블 자료를 이용한다[1]. 지도학습 방법은 훈련 문서의 집합과 연관된 추출 요약문을 요구하다. 또한 이 방법들은 요약 시스템의 훈련을 위해서는 많은 양의 문서에 레이블링을 해야 하기 때문에 시간이 많이 소되고, 많은 응용분야에 적합하지 않다.

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

비지도 학습 방법의 문서요약은 다음과 같다. 잠재의미분석(LSA, latent semantic Analysis)을 이용한 방법으로, Gong과 Liu는 문서를 요약하였고[5], Sum과 Shen은 웹 페이지를 요약하였다[13]. 잠재의미분석은 문장을 선택하기 위하여 복합(multiple) singular vector의 구성요소(component)를 이용한다. Singular vector에 일치하는 singular value는 양수와 음수의 구성요소를 갖고, 이러한 의미가 작은 singular vector 구소요소 값에 의해서 추출 문장의 순위가 구성 될 수 있다[14]. 문서의 주제(topic)를 이용한 방법으로, Nomoto와 Matsumoto는 변형된 k-means를 이용하여 문서에서 다양한 주제를 찾은 후, 각 주제에 일치하는 문장을 선택하여 문서를 요약 하였다[12]. Zha는 문장과 용어(terms)로부터 saliency scores를 계산하고, 이를 이용하여 문장들을 주제그룹(topical groups)들로 군집하여 문서를 요약하였다[14]. Harabagiu와 Lacatusu는 다중문서(multi-document) 요약을 위하여 다섯 개의 주제를 이용한 문서요약방법에 대하여 비교 평가하였다[6]. 이들 방법은 먼저 주제를 추출 한 다음 문서를 요약하기 때문에 계산비용이 많이 드는 단점이 있다.

비음수 행렬 인수분해(NMF, non-negative matrix factorization)는 Lee와 Seung이 제안한 방법으로 다변량 자료를 유용하게 분해하는 알고리즘이다[8,9].

본 논문은 비음수 행렬 인수분해를 이용하여 문장을 추출하여 문서를 요약하는 새로운 방법을 제안하였다. 제안된 방법은 다음과 같은 장점을 갖는다. 첫째, NMF에 의해 찾아지는 의미 특징(semantic feature)들이 비 음수 값을 갖기 때문에 잠재의미분석에 비해 의미 있는

문서요약 결과를 갖는다. 둘째, 적은 계산비용으로 쉽게 문장을 추출할 수 있다.
본 논문의 구성은 다음과 같다. 제2장에서는 제안한 문서요약방법을, 제3장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제4장에서 결론은 맺는다.

2. 포괄적 요약의 생성

문서는 여러 개의 주제로 구성된다. 어떤 주제들은 다수의 문서나, 문서의 주요 내용을 표현할 수 있다. 다른 주제는 단지 주요 주제를 보충하는데 언급하거나, 전체 내용을 좀더 완벽하게 만든다. 효율적인 포괄적 문서요약은 가능한 문서의 중요한 주제를 포함하면서, 동시에 중복을 최소화 시키는 것이다[5].
본 장에서는 NMF를 기반으로 문장을 추출하여 포괄적 문서요약을 할 수 있는 방법을 제안한다. 제안 방법은 전처리 단계와 문서요약 단계로 이루어진다. 다음 세부 장에서 두 단계에 대하여 자세히 기술한다.

2.1 전처리

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 계산하였다[9]. 벡터는 $T_i = [\ t_{1i},\ t_{2i},\ \cdots,\ t_{ni}\ ]^T$는 i번째 문장의 term-frequency 이다. 여기서 요소 t_{ji}는 i번째 절에서 출현한 j번째 term의 빈도이다. A는 m개의 term과 n개의 sentence로 이루어진 $m \times n$ 행렬이다. 요소 A_{ji}는 i번째 문장에서 j번째 term이 출현한 빈도의 가중치이다.

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

여기서 $L(j,i)$는 i번째 문장에서 j번째 term을 위한 지역 가중치(local weight)이고, $G(j,i)$는 문서 전체에서 j번째 term을 위한 전역 가중치(global weight)로 다음의 가중치를 사용하였다.

$L(i) = 0.5 + 0.5 * (tf(i)/tf(max))$, $G(i) = \log(N/n(i))$.
여기서, $tf(i)$는 문장에서 i번째 term이 출현한 빈도, $tf(max)$는 문장에서 가장 큰 발생 빈도를 가진 term, N은 문서에서 문장의 총 개수이다. $n(i)$는 i번째 term을 포함한 문장의 개수이다.

2.2 비음수 의미 가변 행렬에 의한 문서요약

NMF를 적용한 문서요약 단계는 다음과 같다. 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.

$$A \approx WH \quad (2)$$

본 논문에서 제안한 방법을 실험하기 위하여 Reuters-21578 [15] 컬렉션과 야후코리아 뉴스[16] 에서 200건의 기사를 무작위로 선택하여 테스트 자료로 여기서 행렬 A는 근사값을 가지는 $n \times r$ 행렬 W와 $r \times m$ 행렬 H로 인수분해 된다. 여기서 r은 일반적으로 n이나 m보다 작게 선택하여 행렬 W나 행렬 H가 행렬 A보다 작게 한다. NMF 다음으로 $\|A - WH\|^2$ 가 수렴 할 때까지 식(4)을 이용하며, W와 H 행렬 값이 동시에 갱신 된다.

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T WH)_{a\mu}},\ W_{ia} \leftarrow W_{ia} \frac{(VH^T)_{ia}}{(WHH^T)_{ia}} \quad (4)$$

행렬 A의 j번째 열벡터 $A_{\cdot j}$는 행렬 W의 l번째 열벡터 $W_{\cdot l}$와 행렬 H의 요소 h_{lj}가 선형조합(linear combination)을 이루며 식(5)과 같다.

$$A_{\cdot j} = \sum_{l=1}^{r} h_{lj} W_{\cdot l} \quad (5)$$

비음수 행렬 W와 H에 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 NMF에 의해 의미적으로 관련 있는 용어로 군집화된다. 의미적으로 관련된 군집이 의미 특징으로 결합하여, 문맥상에서 동음이의어를 구별하는데 NMF를 사용한다[8].

비음수 행렬 인수분해를 이용한 문서요약 방법은 다음과 같다.

1. 문서 D를 개개의 문장(sentences)으로 분해하고, 추출문장의 개수 k를 지정한다.
2. 각각의 문장에 대한 불용어 제거 및 어근추출 후, 식(1)을 이용하여 term-frequency 벡터의 가중치 계산, term-sentence 행렬 A를 구성한다.
3. 행렬 A에 식(3)과 식(4)를 적용하여 식(2)과 같은 비음수 행렬 W, H로 인수분해 한다.
4. 행렬 H에서 p번째 행에 포함된 행 벡터 $H_{p\cdot}$의 요소의 합 $\sum_{i=1}^{n} H_{pi}$ 을 각각 행 벡터 별로 계산한다. 행 벡터의 요소의 합 값이 큰 순서로 k개의 행 벡터 $H_{p\cdot}$를 선택한다.
5. 선택된 k개의 행 벡터 각각에서, 행에서 가장 큰 요소 값을 가진 q열과 같은 열에 있는 행렬 A의 문장 벡터 $A_{\cdot q}$에 대응되는 문장을 선택한다.

3. 성능평가

본 논문에서는 영어문서의 term추출을 위하여 Rijsbergen의 불용어 목록을 이용한 불용어 제거 및 Porter 스테밍 알고리즘을 이용하여 스테밍 하였다[4]. 영어문서의 불용어 제거와 스테밍 과정과는 달리, 한글문서에서 term을 추출하는 방법은 어렵다. 본 논문에서는 한글 문서에 대한 term을 추출하기 위하여 한글 언어 분석기인 HAM(hangul analysis module)을 이용하였다. 이것은 형태소 분석에의한 자동 색인기능을 지원하다[9].

제안하는 방법을 비교하기 위하여 명의 평가자에 의해 수동으로 요약 되었다. 평가자에 의하여 Retuers는 평균 2.67 문장을 Yahoo Korea는 평균 2.08 문장이 선택되었다. 다음 <표 1> 은 평가자료에 대한 특성을 나타낸다.

[표 1] 평가자료에 대한 특성표

문서속성	Reuters-21578	Yahoo Korea News
문서수	100	100
문장이 10개이상인 문서	35	32
평균문장수	10.09	10.1
최소문장수	3	3
최대문장수	40	36

성능 평가는 문선요약에서 주로 사용되는 정확률(Precision), 재현율(Recall), F-measure를 이용하였다[4]. 평가척도는 다음 식 (6)이다.

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

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

본 논문에서는 LSA 방법을 제안된 방법과 식(6)을 적용하여 실험평가 하였다[5]. 다음 <표 2>는 실험 방법을 비교 평가한 결과이다.

<표 2> 각 실험 방법의 비교 결과

[표 2] 평가결과

Test data	LSA			NMF		
	R	P	F	R	P	F
Reuter-21578	0.49	0.23	0.30	0.52	0.24	0.31
Yahoo Korea	0.48	0.21	0.27	0.53	0.20	0.28

실험에서 보듯이 제안된 방법은 LSA에 비하여 좋은 성능을 보인다. 제안방법이 비음수 값과 부분정보를 이용하는 인간의 인식과정[8]과 유사한 과정으로 문서를 처리하기 때문이다.

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. 적은 계산비용을 통하여 쉽게 문장을 추출할 수 있다.

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

5. 참고문헌

[1] Amini, M. R., Gallinari, P.: The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization. In Proceeding of ACM SIGIR' 02, 105-112 (2002)

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

[3] Chuang, W. T., Yang, J.: Extracting Sentence Segments for Text Summarization: A Machine Learning Approach. In Proceeding of ACM SIGIR' 00, 152-159 (2000)

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

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

[6] Harabagiu, Sanda.: Finley Lacatusu. Topic Themes for Multi-Document Summarization. In proceeding of ACM SIGIR' 05 (2005) 202-209

[7] Kang, S. S.: Information Retrieval and Morpheme Analysis. HongReung Science Publishing Co. (2002)

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

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

[10] Marcu, D.: The automatic construction of large-scale corpora for summarization research. In proceeding of ACM SIGIR' 99 (1999) 137-144

[11] Mani, I., Maybury, M. T.: Advances in Automatic Text. The MIT Press (1999)

[12] Nomoto, T.: Yuji Matsumoto. A New Approach to Unsupervised Text Summarization. In proceeding of ACM SIGIR' 01 (2001) 26-34

[13] 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

[14] Zha, Hongyuan.: Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. In proceeding of ACM SIGIR' 02 (2002)

[15]http://kdd.ics.uci.edu/database/reuter21578/reuters21578.html (2006)

[16] http://kr.news.yahoo.com/ (2006)