비음수 행렬 인수분해와 NMF 군집방법을 이용한 다중문서요약

박선*, 이주홍**, 김철원*
*호남대학교 컴퓨터공학과
**인하대학교 컴퓨터정보공학과
e-mail : sunpark@honam.ac.kr, juhong@inha.ac.kr, cwkim@honam.ac.kr

Multi-document Summarization using Non-negative Matrix Factorization and NMF Clustering Method

Sun Park*, Ju-Hong Lee**, Chul-Won Kim*
*Dept. of Computer Engineering, Honam University
**Dept. of Computer Science and Information Engineering, Inha University

요 약

본 논문은 비음수 행렬 인수분해(NMF, non-negative matrix factorization)와 NMF 군집방법을 이용하여 다중문서를 요약하는 새로운 방법을 제안하였다. 본 논문에서 NMF 에 의해 계산된 의미 특징(semantic feature)은 문서의 고유 구조(inherent structure)를 반영하여 문장을 추출함으로써 요약의 질을 높일 수 있고, 의미 변수(semantic variable)를 이용한 문장의 군집은 문장 간의 유사성과 다양성 고려하여서 쉽게 과잉정보를 제거하여 문장을 요약할 수 있는 장점을 갖는다.

1. 서 론

문서 요약은 문서의 기본적인 내용을 유지하면서 문서의 량을 줄이는 작업으로 문서에서 가장 중요한 내용을 표현하는 것이다. 요약 방법의 적용 대상에 따라서 단일 문서와 다중 문서 요약으로 나눌 수 있다. 대부분의 단일 문서는 중요한 정보가 거의 문서 앞부분에 위치한다. 이런 이유 때문에 처음 문장(sentences)을 추출하여 문서를 요약하면 일반적으로 좋은 요약 결과를 얻을 수 있다. 대부분의 단일문서는 구조가 잠재적으로 일치됨을 알 수 있다. 그러나 다중문서의 요약은 단일문서 요약과는 반대로, 다른 구조 특성을 갖는 연관된 문서의 집합들에 의해서 요약된다. 이 때문에 단일문서에서 적용한 기술을 다중문서에는 적용 할 수 는 없다. 즉, 다중 문서 요약은 문서의 구조적 특성이 아닌, 문서 집단에서 공통적으로 포함하는 정보에 의해 요약된다. 이는 대부분의 다중 문서요약이 문서의 집합에서 유사한 문장이나 문단에 포함된 중요한 정보로 요약 됨을 알 수 있다. 원칙적으로 다중문서요약은 모든 문서에서 공통적으로 관련된 정보를 포함하면서 사용자의 질의에 직접 관련된 정보를 포함한다[1, 3, 4, 9, 11, 12].

비음수행인수분해(NMF; non-negative matrix factorization)는 Lee 와 Seung 이 제안한 방법으로 인간이 객체를 인식할 때 객체의 부분정보의 조합으로 인식하는 것에 착안하여, 객체정보를 기초특질(base feature)과 부호특질(encoding feature)로 나누어 부분정보(part-base)로 표현한다. 이러한 부분정보의 조합으로 전체 객체를 표현하는 방법은 대량의 정보를 효율적으로 표현 할 수 있는 방법이다. 현재 NMF 는 이미지 처리와 신호처리 분야에서 주로 응용되고 있으며, Xu 는 문서군집에 NMF 를 이용하였다[7, 8, 13].

본 논문은 NMF 를 이용하여 다중문서를 요약하는 새로운 방법을 제안하였다. 제안된 방법은 NMF 를 이용하여 문장을 군집하고, 군집된 문장집합을 NMF 를 이용하여 비음수 의미 특징 행렬(NSFM, non-negative semantic feature matrix)과 비음수 의미 변수 행렬

(NSVM, non-negative semantic variable matrix)로 분해한다. 다중문서요약을 위하여 NSFM 과 NSVM 로부터 공통된 내용과 내용상의 차이점을 유도하여 문서요약을 할 수 있는 알고리즘을 제안하였다.

본 논문의 구성은 다음과 같다. 제 2 장에서는 다중문서 요약에 관한 관련연구를, 제 3 장에서는 NMF 를 이용하여 다중문서 요약방법에 대하여 기술한다. 제 4 장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제 5 장에서 결론을 맺는다.

2. 관련연구

최근의 다중문서요약에 관한 관련연구는 다음과 같다. Goldstein 외 저자들은 MMR(maximal marginal relevance)를 다중문서 요약에 적용한 통계적 방법을 제안하였다[1]. Hachey 외 저자들은 대량의 문서로부터 질의 지향의 다중문서요약을 위하여 MMR 과 LSA(Latent Semantic Analysis)를 이용한 방법을 제안하였다[3].

문서의 주제(topic)을 이용한 방법으로, Nomoto 와 Matsumoto 는 변형된 k-means 를 이용하여 문서에서 다양한 주제를 찾은 후, 각 주제에 일치하는 문장을 선택하여 문서를 요약 하였다[9]. Harabagiu 와 Lacatusu 는 다중문서요약을 위하여 다섯 개의 주제를 이용한 문서요약방법에 대하여 비교 평가하였다[4]. Sassion 은 주제기반의 다중문서요약 방법을 제안하였다. 제안된 방법은 문장을 제거하여 문서를 요약하는 방법으로 사용자가 지정한 압축율까지 후보문장집합으로부터 문장을 제거하여 문서를 요약한다[12]

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[11].

3. 제안방법

본 장에서는 NMF 를 기반으로 문장을 추출하여 다중 문서요약을 할 수 있는 방법을 제안한다. 효율적인 다중 문서요약을 위해서는 다음과 같은 방법을 제안하였다. 첫째, NSVM 을 이용하여 문장을 군집하였다. 문장을 군집함으로써 유사한 문장들이 각각의 cluster 를 구성하여 중복 및 관련 없는 정보를 쉽게 구분 할 수 있으며, 문장간의 다양성을 쉽게 보장하였다. 둘째, 질의와 cluster 의 NSFM 간의 유사도를 계산하였다. 각 cluster 에서 질의와 유사도가 가장 높은 문장을 추출함으로써 사용자가 원하는 목적과 관련된 정보를 최대한 반영하였다.

제안 방법은 전처리 단계와 문서군집단계, 문장 추출에 의한 문서요약 단계로 이루어진다. 다음 장에서 세 단계에 대하여 자세히 기술한다.

3.1 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 계산하였다[10]. 벡터는 $T_i = [\ t_{1i}, t_{2i}, \dots, t_{ni}\ ]^T$ 는 i 번째 문장의 용어 빈도이다. 여기서 요소 t_{ji} 는 i 번째 절에서 출현한 j 번째 용어의 빈도이다. $\boldsymbol{A}$ 는 m 개의 용어와 n 개의 문장으로 이루어진 $m \times n$ 행렬이다. 요소 A_{ji} 는 i 번째 문장에서 j 번째 용어가 출현한 빈도의 가중치이다.

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

여기서 $L(j,i)$는 i 번째 문장에서 j 번째 용어를 위한 지역 가중치(local weight)이고, $G(j)$는 문서 전체에서 j 번째 용어를 위한 전역 가중치(global weight)로 다음과 같다.

$$G(j) = \log(N/n(j)) \tag{2}$$

여기서, N 은 전체문서에 포함된 문장의 총 개수고, $n(j)$는 j 번째 용어를 포함한 문장의 개수이다.

3.2 NMF 를 이용한 문장의 군집

문장의 군집화 단계는 NMF 에서 얻어진 NSVM 의 $\boldsymbol{H}^T$ 행렬을 이용하여 문장을 군집한다. 본 논문에서는 Xu 가 제안한 NMF 에 기반한 문서군집방법을 문장에 적용하였다[13]. 다음은 본 논문에 사용되는 NMF 이다[7, 8].

NMF 를 계산하는 방법은 다음과 같다. 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(3)와 같다.

$$A \approx WH \tag{3}$$

여기서 행렬 $\boldsymbol{A}$ 는 근사값을 가지는 $n \times r$ 행렬 $\mathbf{W}$

와 $r \times m$ 행렬 **H 로** 인수분해 된다. 여기서 r 은 일반적으로 n 이나 m 보다 작게 선택하여 행렬 $\boldsymbol{W}$ 나 행렬 $\boldsymbol{H}$ 가 행렬 $\boldsymbol{A}$ 보다 작게 한다.

목표함수가 만족되거나, 지정한 반복횟수가 만족할 때까지 식(4)와 (5)을 반복하여 갱신한다. $\boldsymbol{W}$ 와 $\boldsymbol{H}$ 행렬 값이 동시에 갱신 된다.

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T A)_{\alpha\mu}}{(W^T W H)_{\alpha\mu}} \tag{4}$$

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

NMF 를 이용하여 문장을 군집하는 방법[13]은 다음과 같다.

1. 다중문서 $\boldsymbol{D}$ 를 개개의 문장으로 분해하고, 군집할 cluster 의 개수 n 을 지정한다.
2. 각 문장으로부터 불용어 제거 및 어근 추출 후, 식(1)을 이용하여 용어 빈도 벡터의 가중치를 계산하여 용어-문장 행렬 $\boldsymbol{A}$ 를 구성한다.
3. 행렬 $\boldsymbol{A}$ 에 식 (4)와 (5)을 적용하여 식(2)와 같은 비음수 행렬 $\boldsymbol{W}, \boldsymbol{H}$ 로 인수분해 한다.
4. $\boldsymbol{H^T}$ 행렬을 이용하여 각 문장에 속하는 cluster 를 결정하다.

위의 4 번째 단계에서 $\boldsymbol{H^T}$ 행렬은 $\boldsymbol{H}$ 행렬의 전치행렬이다. 여기서 문장이 속할 cluster 를 결정하는 방법은 $\boldsymbol{H^T}$ 행렬의 임의의 열 벡터에서 가장 큰 요소의 값을 가지는 행의 위치가 그 임의의 열 벡터와 대응하는 문장이 속하는 cluster 이다[13].

3.3 NMF 를 이용한 문장 추출

문장 추출 단계는 3.2 장에서 얻어진 n 개의 cluster 를 이용하여 문장을 추출한다. 각각의 cluster 를 다시 NMF 하고, 행렬 $\boldsymbol{V_n}$ 와 질의간의 유사도가 가장 높은 행 벡터와 일치하는 행렬 $\boldsymbol{H_n}$ 의 열 벡터를 찾은 후, 열 벡터에서 가장 문장 관련도가 높은 행과 일치하는 문장을 추출한다.

의미 특징 벡터와 질의 간의 유사도를 구하는 식은 (6)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 ($w_{iq} \geq 0$). m 은 r 열 벡터의 요소들의 수로, 벡터 $\vec{q}$ = (w_{1q}, w_{2q}, …, w_{mq})로 나타낸다[2, 7, 14].

$$sim(w_j, q) = \frac{\vec{w}_j \cdot \vec{q}}{|\vec{w}_j| \times |\vec{q}|} \tag{6}$$

문장의 관련도(*RS*, relevance of a sentence)를 다음과 같이 정의 하였다.

$$RS_j = \sum_{i=1}^{r} H_{ij} \cdot weight(H_{i\bullet}) \tag{7}$$

$$weight(H_{i\bullet}) = \frac{\sum_{q=1}^{n} H_{iq}}{\sum_{p=1}^{r}\sum_{q=1}^{n} H_{pq}} \tag{8}$$

여기서 가중치 ($H_{i\bullet}$) 는 모든 의미 특징에서 i 번째의 의미 특징($W_{i\bullet}$)과 관련된 관계를 의미하고, 문장의 관련성은 의미 특징에 의해서 표현되는 중요한 주제(major topic)가 문장에 얼마나 반영되는가를 의미한다

NMF 를 이용한 문장 추출 방법은 다음과 같다.

1. n 개의 cluster 부터 각각 행렬 $\boldsymbol{A_n}$ 을 구성한다.
2. 행렬 $\boldsymbol{A_n}$ 에 식(4)과 식(5)를 적용하여 식(2)과 같은 비음수 행렬 $\underline{\boldsymbol{W_n}}$, $\mathbf{H_n}$ 으로 인수분해 한다.
3. 식(6)을 이용하여 행렬 $\boldsymbol{W_n}$ 의 열 벡터들과 질의간 유사도를 계산하여 가장 유사도가 높은 p 번째 열 벡터 $W_{\cdot p}$ 를 찾는다.
4. 식(7)와 식(8)을 이용하여 $\boldsymbol{H_n}$ 에 대한 문장 관련도를 계산한다.
5. 행렬 $\boldsymbol{H_n}$ 에서 p 번째 행에 포함된 행 벡터 $H_{p\cdot}$ 에서 가장 큰 요소 값을 가진 q 열과 같은 열에 있는 행렬 $\boldsymbol{A_n}$ 의 문장 벡터 $A_{\cdot q}$ 에 대응되는 문장을 선택한다.
6. 만약 미리 정의된 cluster 의 n 만큼 문장이 선택되면 알고리즘을 종료하고, 그렇지 않으면 2 단계로 간다.

4. 성능평가

본 논문에서는 한글 문서에 대한 용어를 추출하기 위하여 한글 언어 분석기인 HAM(hangul analysis module)을 이용하였다[8]. 본 논문에서 '야후코리아 뉴스'[7] 에서 240 건의 기사를 실험 자료로 사용하였다. 제안방법을 비교하기 위하여 두명의 평가자에 의해 수동으로 요약 되었다. 평가자에 의하여 Yahoo Korea 는 평균 3.8 문장이 선택되어 요약되었다.

성능 평가는 문서요약에서 주로 사용되는 정확률(*P*, precision), 재현율(*R*, recall), F-measure(*F*)를 이용하였다

[2, 14]. 평가척도는 다음 식 (9)이다.

$$R = \frac{|S_{man} \cap S_{sum}|}{|S_{man}|}, \quad P = \frac{|S_{man} \cap S_{sum}|}{|S_{sum}|}, \quad F = \frac{2RP}{R+P} \quad (9)$$

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

본 논문에서는 Saggion 방법[12]에 제안된 방법을 비교하였고, 식(9)을 적용하여 실험평가 하였다. 다음 <표 2>는 실험 방법을 비교 평가한 결과이다.

<표 2> 각 실험 방법의 비교 결과

Test data	Saggion			제안방법		
	R	**P**	**F**	**R**	**P**	**F**
Yahoo Korea	0.19	0.14	0.16	0.39	0.27	0.32

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

5. 결론

본 논문은 다중 문서 요약을 위해서 비음수 의미 특징 행렬(NSFM)과 비음수 의미 변수 행렬(NSVM)를 이용하여 문장을 추출하는 새로운 방법을 제안하였다. 제안된 방법은 다음과 같은 장점을 갖는다. 문서의 고유 구조가 반영된 NSFM 을 이용하여 문장을 추출하기 때문에 요약의 질을 높일 수고, 다중문서에 포함된 문장간의 유사성과 다양성을 유도할 수 있도록 NSVM 을 이용하여 쉽게 문서를 군집할 수 있다. 마지막으로 군집된 문서들로부터 쉽게 과잉정보(redundancy)를 제거하고, 가장 중요한 문장만을 간소화된 형식으로 사용자에게 제공할 수 있다.

6. 참고문헌

[1] Goldstein. J., Mittal. V., Carbonell. J., Callan. J, Creating and Evaluating Multi-Document Sentence Extract Summaries. The Proceeding of CIKM (2000)

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

[3] Hachey. B., Murray. G., Reitter. D, The Embra System at DUC 2005: Query-oriented Multi-document Summarization with a Very Large Latent Semantic Space, In Proceedings of the DUC'05, (2005)

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

[5] http://kr.news.yahoo.com/ (2005)

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

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

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

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

[10] Ricardo, B. Y., Berthier, R. N.: Moden Information Retrieval, ACM Press (1999)

[11] Sakurai, T., Utsumi, A.: Query-based Multidocument Summarization for Information Retrieval. The Proceeding of NTCIR-4 (2004)

[12] Sassion. H.: Topic-based Summarization at DUC 2005. In Proceedings of the Document Understanding Conference 2005 (DUC'05), (2005)

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