한국항행학회 논문지 제13권 제1호 2009년 2월

군집과 비음수 행렬 분해를 이용한 개인화된 문서 요약

Personalized Document Summarization Using NMF and Clustering

박 선*

Sun Park*

요 약

본 논문은 비음수 행렬 분해와 군집 방법을 이용하여 개인화된 문장을 추출하여 문서요약을 하는 새로운 방법을 제안한다. 제안된 방법은 검색 문서에 군집 방법을 이용하여 문서의 주제와 세부 주제를 반영한 문장을 추출하며, 비음수 행렬 분해에 의해 분해된 문서의 고유 의미 특징을 이용하여 사용자의 흥미가 잘 반영된 문장을 추출한다. 실험결과 제안방법이 유사도, 비음수행렬분해를 이용한 방법들에 비하여 좋은 성능을 보인다.

Abstract

We proposes a new method using the non-negative matrix factorization (NMF) and clustering method to extract the sentences for personalized document summarization. The proposed method uses clustering method for retrieving documents to extract sentences which are well reflected topics and sub-topics in document. Beside it can extract sentences with respect to query which are well reflected user interesting by using the inherent semantic features in document by NMF. The experimental results shows that the proposed method achieves better performance than other methods use the similarity and the NMF.

Key words : non-negative matrix factorization, personalized document summarization, cosine similarity

I. 서 론

인터넷 환경에서 끊임없는 정보의 증가는 사용자들에게 필요한 정보만을 검색할 수 있는 방법을 요구하고 있다. 이러한 효율적인 정보검색의 요구는 사용자에게 맞춤형 검색을 지원하도록 하는 개인화된 문서요약의 필요성을 점차 증가 시키고 있다.

문서 요약은 문서의 기본적인 내용을 유지하면서 문서의 량을 줄이는 작업이다[2]. 문서의 요약은 문서 내용전체를 요약하는 포괄적 문서 요약(generic document summary)과 사용자의 질의에 따라 질의에 관련 있는 내용만을 포함하는 질의 기반 문서 요약(query-based document summary)으로 나눌 수 있다[2].

개인화된 문서요약은 문서 자체의 정보를 요약하는 것 보다는 사용자의 흥미와 관련된 특별한 정보를 유지하면서 문서를 요약하는 작업이다[1].

개인화된 문서 요약의 최근 연구는 다음과 같다. Sanderson은 최고 문장 연산과 질의 확장을 이용하여 정밀한 사용자 직접 요약 방법을 제안하였다[3]. Tombros와 Sanderson은 제목 점수, 주제 점수, 용어의

* 호남대학교 컴퓨터공학과(Department of compuer Engineering, Honam University)
· 제1저자 (First Author) : 박 선
· 투고일자 : 2008년 12월 2일
· 심사(수정)일자 : 2008년 12월 3일 (수정일자 : 2009년 2월 16일)
· 게재일자 : 2009년 2월 28일

빈도 정보, 질의 점수 등을 조합하여서 사용자를 보조하는 요약문을 생성하는 방법을 제안하였다[4]. Varadarajan과 Hristidis는 질의와 가장 관련이 높은 단락과 의미 연관에 의한 복합 주제를 이용하여서 질의에 특화된 문서 요약 방법을 제안하였다[5]. Diaz와 Gervas는 개인화 문서요약 방법과 포괄적 문서요약 방법을 조합하여 개인화된 문서 요약 방법을 제안하였다. 이 방법은 용어의 위치와 주제 단어를 조합하는 포괄적 문서요약 방법을 사용하였으며, 개인화 문서 요약 방법은 사용자의 질의에 가장 접합한 문장을 추출하는 방법을 사용하였다[1]. Park외 저자들은 적합 척도와 비음수 행렬 분해(NMF, non-negative matrix factorization)를 이용한 자동 개인화된 문서 요약 방법을 제안하였다. 이 방법은 적합척도를 이용하는 포괄적 문서 요약 방법과 비음수 행렬 분해를 이용한 질의 기반의 문서 요약 방법을 조합하여 개인화된 문서를 요약한다[6].

본 논문은 비음수 행렬 분해(NMF, non-negative matrix factorization)에 의해한 질의 기반의 문서요약과 군집기반의 포괄적 문서요약을 조합하여 새로운 개인화된 문서 요약 방법을 제안하였다. 비음수 행렬 분해는 Lee와 Seung이 제안한 방법으로 자료를 비음수 의미 특징(NSF, non-negative semantic features)과 비음수 의미 변수(NSV, non-negative semantic variable)로 분해하는 알고리즘이다[7, 8].

제안된 방법은 다음과 같다. 요약할 문서를 문장으로 분해하고, 분해된 문장들은 벡터모델에 따라서 벡터로 표현한다. K-means 군집방법을 이용하여 문장을 군집하고, 군집과 군집에 포함된 문장 간의 유사도를 계산하여서 대표 문장을 선택하고, 이를 이용하여 포괄적 문서 요약을 한다. 비음수 행렬 분해를 이용하여 문장 벡터 행렬을 의미특징 행렬과 의미변수 행렬로 분해한다. 이를 이용하여서 질의와 유사도 값이 가장 높은 의미특징 열벡터를 선택하고, 이 의미 특징 열벡터에 대응되는 의미변수 행벡터를 선택한다. 의미 변수 행벡터에서 가장 큰 요소 값에 대응되는 문장을 선택하여 질의 기반의 문서요약을 한다. 선택된 포괄적 요약 문장과 질의 기반의 요약 문장을 조합하여 개인화된 문서요약을 한다.

제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 군집방법을 이용하여 문서의 중요 주제(major topic)와 세부 주제(sub topic)를 포함하는 포괄적 요약 문장을 추출한다. 비음수 행렬 분해에 분해된 의미 특징(semantic feature)들 문서의 고유 의미 구조(inherent semantic structure) [7]를 나타내기 때문에 사용자의 질의를 잘 반영하는 질의 기반의 문장을 추출한다. 포괄적 문서요약과 질의 기반의 문서요약 방법의 조합으로 개인화된 문서요약의 질을 높일 수 있다.

본 논문의 구성은 다음과 같다. 제2장에서는 비음수 행렬 분해 방법과 K-means군집방법을, 제3장은 제안한 요약방법을, 제4장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제5장에서 결론은 맺는다.

Ⅱ. 비음수 행렬 분해와 K-means 군집

비음수 행렬 분해(NMF, non-negative matrix factorization)는 주어진 비음수 행렬로부터 비음수의 인수를 찾는 행렬 분해 알고리즘이다[7, 8].

본 논문에서 행렬 X의 j번째 열벡터는 X*j로, i번째 행벡터는 Xi*로, i번째 행과 j번째 열의 원소는 Xij로 표시 한다.

비음수 행렬 분해 알고리즘은, 식(1)의 목표함수 J가 0에 가깝게 수렴 할 때까지 식(2)와 식(3)을 이용하여 행렬 W와 H의 값을 동시에 갱신한다.

$$J = \|A - WH\|^2 \tag{1}$$

식(1)의 목적은 행렬 A를 비음수 m×r 행렬 W와 비음수 r×n 행렬 H로 분해하는 것이다. 여기서, A는 m개의 용어와 n개의 문장으로 이루어진 m×n 행렬이고, r은 의미특징의 개수이다.

$$H_{ij} \leftarrow H_{ij} \frac{(W^T A)_{ij}}{(W^T W H)_{ij}} \tag{2}$$

$$W_{ij} \leftarrow W_{ji} \frac{(A H^T)_{ji}}{(W H H^T)_{ji}} \tag{3}$$

행렬 A의 j번째 열벡터 A*j는 행렬 W의 l번째 열벡터 W*l와 행렬 H의 요소 Hkj가 선형조합을 이루며

식(4)과 같다. 즉, l번째 의미 특징 벡터 W*l는 A*j의 문장 벡터 내에서의 가중치가 의미변수 Hlj이다.

$$A_{*j} = \sum_{l=1}^{r} H_{lj} W_{*l} \quad (4)$$

K-means 군집은 n개의 자료를 주어진 K개의 군집으로 묶는 알고리즘이다[9]. 본 논문에서는 문장을 군집하기 위하여 식(5)의 코사인 유사도를 이용한 거리 척도를 사용한다.

$$d(A_{*a}, A_{*b}) = 1 - csim(A_{*a}, A_{*b}) \quad (5)$$

$$csim(A_{*a}, A_{*b}) = \frac{\sum_{j=1}^{m} A_{ja} \times A_{jb}}{\sqrt{\sum_{j=1}^{m} A^{2_{ja}}} \times \sqrt{\sum_{j=1}^{m} A^{2_{jb}}}} \quad (6)$$

여기서, A*a와 A*b는 행렬 A의 a번째와 b번째 열벡터이다. 이 것 들은 비음수 값을 가지므로 0≦csim()≦1 이고, 따라서 0≦d()≦1이다.

Ⅲ. 개인화된 문서요약

본 장에서는 군집을 이용한 포괄적 문서 요약 방법과 비음수 행렬 분해를 이용한 질의 기반의 문서요약 방법을 조합하여서 개인화된 문서요약 방법을 제안한다. 제안 방법은 전처리 단계와 문장추출 단계, 개인화된 문서요약 단계로 이루어진다. 다음 장에서 세 단계에 대하여 자세히 기술한다.

3-1 전처리 단계

전처리 단계는 주어진 문서를 각각의 문장으로 분해 후, 불용어(stop-word) 제거, 어근(stemming)을 추출하며, 용어빈도 벡터를 생성한다[10, 11].

생성된 용어빈도 행렬 A는, i번째 문장 Ai는 용어빈도 벡터 A_i = [A_{1i}, $A2_i$, …, A_{ni}]T로 표현되고, 벡터 A_i 요인 A_{ji}는 i번째 문장에서 j번째 용어를 나타낸다.

3-2 문장 추출 단계

문장 추출 단계는 개인화된 문서 요약을 위하여서 문장을 추출하여 후보 문장 집합을 구성한다. 후보 문장 집합의 구성 방법은 군집 방법을 이용한 포괄적 문장과 비음수 행렬 분해를 이용한 질의 기반의 문장 추출로 구성된다.

군집 방법을 이용한 포괄적 문장 추출 방법은 다음과 같다. 추출할 문장의 개수 K를 설정하고, 전처리된 행렬 A에 식(5)를 이용한 K-means군집 방법으로 군집한다. 각각의 군집에서, 군집과 군집에 포함된 문장 간에 식(6)을 이용하여 코사인 유사도를 계산하고, 유사도가 가장 높은 문장을 추출하여서 유사도와 함께 후보문장 집합에 저장한다.

비음수 행렬 분해를 이용한 질의 기반의 문장 추출 방법[12, 13, 14]은 다음과 같다. 전처리된 행렬 A를 비음수 행렬 분해한다. 이 결과 얻어지는 비음수 의미 특징 행렬 W와 비음수 의미 변수 행렬 H는 다음 식(7)과 같다[7, 8]. 식(6)을 이용하여 비음수 의미 특징 열벡터와 질의 간의 유사도를 계산고, 유사도가 가장 높은 의미특징 열벡터를 선택하며, 선택된 의미 특징 열벡터와 대응 되는 의미 변수 행벡터를 선택한다. 선택된 의미 특징 열벡터에서 가장 큰 요소 값과 대응되는 문장을 추출하여 유사도와 함께 후보문장 집합에 저장한다.

$$A \approx WH \quad (7)$$

여기서, A는 n×m 행렬이고, n×r 행렬 W와 r×m 행렬 H는 행렬 A로부터 근사 값으로 분해 된 행렬이며, 각각 W = [W_1, W_2, ..., W_i], H = [H_1, H_2, ..., H_j]로 나타낸다.

3-3 개인화된 문서 요약 단계

개인화된 문서 요약 단계는 포괄적 문장 추출의 후보 문장과 질의 기반 문장 추출의 후보 문장을 조합하여서 문서를 요약한다. 개인화된 문서 요약은 식(8)을 이용하여서 계산된 개인화 점수(pscore) 값이 가장 높은 문장 순서로 추출하여 문서를 요약한다.

$$pscore_i = \frac{\alpha \cdot g_i + \beta \cdot q_i}{\alpha + \beta} \quad (8)$$

여기서, α와 β는 매개변수로, 각각 i번째 문장이 포괄적 후보문장 집합과 질의 기반 후보 집합에 있으면 1이고 없으면 0이며, g_i는 포괄적 후보 문장 집합에 저장된 i번째 문장의 유사도이고, q_i는 질의 기반 후보 문장 집합에 저장된 i번째 문장의 유사도이다.

Ⅳ. 실험 및 평가

본 장에서는 제안 방법을 실험하기 위하여 야후코리아 뉴스 에서 300건의 기사를 무작위로 선택하여 실험 자료로 사용하였다. 제안 방법을 비교하기 위하여 세 명의 평가자가 문서를 수동으로 요약하였다. 즉, 수동으로 요약한 요약문과 제안방법과 비교방법 간의 정확률, 재현율, F-measure를 비교 평가 하였다. 다음 표1은 평가 자료에 대한 특성을 나타낸다.

성능 평가는 문서요약에서 주로 사용되는 정확률(Precision), 재현율(Recall), F-measure를 이용하였다[10, 11]. 평가 척도는 다음 식(9)와 같다.

표 1.평가 자료의 특성
Table 1. Property of the test data set

문서의 속성	야후 코리아
문서의 수	300
30문장 이상인 문서의 수	58
문서당 평균 문장의 수	13
최소 문장의 수	5
최대 문장의 수	42

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

여기서 Sman, Ssum은 각각 사람과 제안된 방법에 의하여 선택된 문장이다.

실험은 네 가지 요약방법들에 대하여서 성능을 비교 평가한 것이다. 그림1은 야후 코리아 뉴스에 대한 각각의 평균 재현율, 평균 정확률, 평균 F-measure 에 대한 평가 결과이다. 여기서 UPS는 Diaz의 개인화된 요약 방법으로 개인화 요약에 기반한 유저 모델을 이용한 방법이다[1]. NMF는 박선등이 제안 방법으로 비음수 행렬 분해와 유사도를 이용한 질의 기반의 문서 요약방법이다[12, 13, 14]. PNMF는 Park등이 제안한 방법으로 적합척도에 의한 포괄적 문서 요약과 비음수 행렬 분해에 의한 질의 기반의 문서요약 방법을 조합하고, 이를 이용하여서 개인화된 문서를 요약 하는 방법이다[6]. CPNMF는 본 논문에서 제안된 방법이다.

그림1에서 보는 것과 같이 야후 코리아 뉴스를 이용한 평가 결과에서는 제안 방법인 CPNMF의 평균 재현율, 정확율, F-measure가 UPS에 비하여 11.6%, 11.8%, 11.0%가 높으며, NMF에 비해서는 22.1%, 20.3%, 14.2%가 높고, PNMF에 비해서는 5.2%, 6.3%, 3.5%가 높다.

성능 평가 결과 제안방법인 CPNMF가 가장 좋은 결과를 나타내며, 다음으로 PNMF, UPS 순으로 평가되었다. NMF의 성능이 가장 저조하였다.

이는 단순히 문서 내부의 고유 의미 특징을 이용한 NMF 방법 보다는 포괄적 문서요약 방법과 질의 기반의 문서 요약 방법을 이용하여 개인의 질의를 반영하면서 문서 전체의 주제를 반영하는 UPS 방법이 더 좋은 성능을 나타내는 것을 알 수 있다.

또한 UPS 방법 보다는 문서의 고유 구조에 사용자의 질의를 반영하는 PNMF 방법이 좀 더 의미 있는 요약문을 생성하는 것을 알 수 있다.

특히 제안 방법인 CPNMF 방법은 군집 방법에 문서의 중요 주제 및 세부 주제를 반영하면서, 문서 고유의 구조에 사용자의 흥미를 반영하여서 가장 의미 있는 문장이 추출되는 것을 알 수 있다.

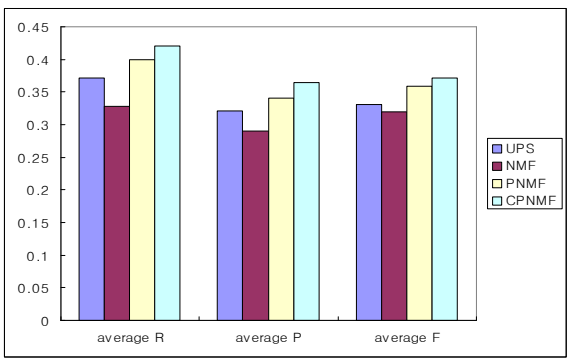

그림 1. 야후코리아 뉴스를 이용한 방법의 비교 결과
Fig. 1. Result of comparison methods

Ⅴ. 결 론

본 논문에서는 K-means 군집 방법을 이용하여 포괄적인 내용을 포함한 문장과 비음수 행렬 분해를 이용하여 사용자의 흥미를 반영한 문장을 추출하여서 개인화된 문서를 요약하는 방법을 제안하였다. 제안된 방법은 군집 방법을 이용하여서 문서에 포함된 중요 주제 및 세부 주제를 요약문에 반영 하였으며, 문서의 고유 특징을 나타나는 의미 특징에 사용자의 질의를 반영함으로써 사용자가 요구하는 주제에 적합한 문장을 추출한다. 또한, 포괄적 문장과 사용자의 요구 사항이 반영된 문장을 조합하여서 개인화된 문서를 요약함으로써 요약의 질을 높였다. 실험결과 이전에 제안된 개인화된 문서 요약 방법에 비하여 더 좋은 평가 결과를 보였다.

참 고 문 헌

[1] A., Diaz, P., Gservas, "User-model based personalized summarization", *Information Processing and Management,* 43, pp.1715-1734, 2007.

[2] I. Mani, M. T. Maybury, "dvances in Automatic Text," *The MIT Press*, 1999.

[3] M., Sanderson, "Accurate user directed summarization from existing tools", *In proceeding of the international conference on information and knowledge management,* pp.45-51, 1998.

[4] A., Tombros, M., Sanderson, "Advantages of Query Biased summaries in Information Retrieval", *In proceeding of ACM SIGIR*, 1998, pp.2-10.

[5] R., Varadarajan, V., Hristidis, "A System for Query Specific Document Summarization", *In proceeding of the CIKM,* pp.622-631, 2006.

[6] S. Park, J. W. Song, J. H. Lee, "Automatic Personalized Summarization using Non-negative Matrix Factorization and Relevance Measure" *In proceeding of IWSCA'08*, 2008.

[7] D. D. Lee, H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp.788-791, 1999.

[8] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization," *In Advances in Neural Information Processing Systems,* vol. 13, pp.556-562, 2001.

[9] J., Han, M., Kamber, "Data Mining Concepts and Techniques", *Morgan Kaufmann*, 2001.

[10] S. Chakrabarti, "Mining the Web : Discovering Knowledge from Hypertext Data," *Morgan Kaufmann*, 2003.

[11] W. B. Frankes, R. Baeza-Yaes, "Information Retrieval : Data Structure & Algorithms," *Prentice-Hall*, 1992.

[12] 박선, 이주홍, 안찬민, 박태수, 김재우, 김덕환, “비음수 행렬 인수분해를 이용한 일반적 문서 요약,” *제25회 한국정보처리학회 춘계학술발표대회 논문집,* 제13권, 제1호, 2006.

[13] 박선, “의미 특징 행렬과 의미 가변행렬을 이용한 질의 기반의 문서 요약”, *한국항행학회 논문지,* 제12권, 제4호, 2008.

[14] 박선, 이주홍, “비음수 행렬 분해와 K-means를 이용한 주제기반의 다중문서요약”, *한국정보과학회 논문지,* 제35권, 제4호, 2008.

박 선 (朴仙)

1996년 2월 : 전주대학교 전자계산학과(이학사)

2001년 8월 : 한남대학교 정보산업대학원 정보통신학과(공학석사)

2007년 8월 : 인하대학교 컴퓨터정보공학과 (공학박사)

2008~현재 : 호남대학교 컴퓨터공학과 전임강사

관심분야 : 정보검색, 데이터마이닝, 데이터베이스