문서의 의미특징을 이용한 주제 기반의 다중문서 요약

박선*, 안동언*, 김철원**
*전북대학교 전자정보공학부
**호남대학교 컴퓨터공학과
e-mail: *{sunbak, duan}@chonbuk.ac.kr, **cwkim@honam.sck.ac.kr

Topic-Based Multi-Document Summarization using Semantic Features of Documents

Sun Park*, Dong Un An*, Chul-Won Kim**
*Division of Electronic & Information Engineering, Chonbuk National University
**Dept of Computer Engineering, Honam University

요 약

인터넷의 발전은 대량의 정보를 양산하였고, 이러한 대량의 정보 집합 내에서는 비슷한 정보가 재활용 되거나 반복되는 정보중복문제를 가지고 있다. 중복되는 정보들로부터 사용자에게 원하는 정보를 신속히 검색할 수 있도록 하는 정보 요약에 대한 필요성은 점차 증가하고 있다. 본 논문은 비음수 행렬 인수분해(NMF, non-negative matrix factorization)에 의한 문서의 의미특징을 이용하여 주제기반의 다중문서를 요약하는 새로운 방법을 제안한다. 본 논문에서는 다중문서가 포함하고 있는 문서들 간의 고유구조를 문서요약에 이용하여서 요약의 질을 높일 수 있고, 주제와 문장 간의 유사성과 다양성 고려하여서 쉽게 과잉정보를 제거하여 문장을 요약할 수 있는 장점을 갖는다.

1. 서론

최근의 다중문서요약에 관한 관련연구는 다음과 같다. Goldstein외 저자들은 MMR(maximal marginal relevance)를 다중문서 요약에 적용한 통계적 방법을 제안하였다[1]. Hachey외 저자들은 대량의 문서로부터 질의 지향의 다중문서요약을 위하여 MMR과 LSA(Latent Semantic Analysis)를 이용한 방법을 제안하였다[2]. Sakurai와 Utsumi는 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[8]. 문서의 주제(topic)을 이용한 방법으로, Nomoto와 Matsumoto는 변형된 k-means를 이용하여 문서에서 다양한 주제를 찾은 후, 각 주제에 일치하는 문장을 선택하여 문서를 요약 하였다[6]. Harabagiu와 Lacatusu는 다중문서요약을 위하여 다섯 개의 주제를 이용한 문서요약방법에 대하여 비교 평가하였다[3]. Sassion은 주제기반의 다중문서요약 방법을 제안하였다. 제안된 방법은 문장을 제거하여 문서를 요약하는 방법으로 사용자가 지정한 비율까지 후보문장집합으로부터 문장을 제거하여 문서를 요약한다[9]

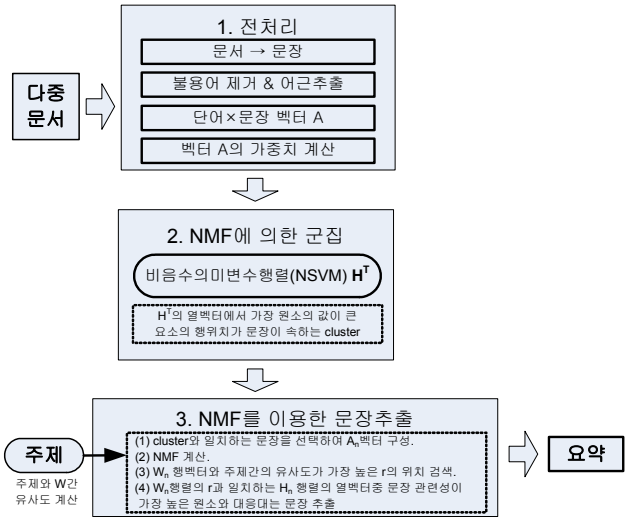
비음수 행렬 분해(NMF, Non-negative Matrix Factorization)는 Lee와 Seung이 제안한 방법으로 인간이 객체를 인식할 때 객체의 부분정보의 조합으로 인식하는 것에 착안하여, 객체정보를 기초특질(base feature)과 부호특질(encoding feature)로 나누어 부분정보(part-base)로 표현한다. 이러한 부분정보의 조합으로 전체 객체를 표현하는 방법은 대량의 정보를 효율적으로 표현 할 수 있는 방법이다[5].

본 논문은 비음수 행렬 분해에 의한 문서의 의미특징을 이용하여 주제기반의 다중문서를 요약하는 새로운 방법을 제안하였다. 제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 문서의 고유 구조가 반영된 비음수 의미 특징 행렬(NSFM, Non-negative Semantic Feature Matrix)을 이용하여 문장을 추출하기 때문에 요약의 질을 높일 수 있다. 둘째, 다중문서에 포함된 문장 간의 유사성과 다양성을 유도할 수 있도록 비음수 의미 변수 행렬(NSVM, Non-negative Semantic Variable Matrix)을 이용하여 쉽게 문서를 군집할 수 있다. 마지막으로 군집된 문서들로부터 쉽게 과잉정보(redundancy)를 제거하고, 가장 중요한 문장만을 간소화된 형식으로 사용자에게 제공할 수 있다.

2. 비음수 행렬 분해에 의한 다중 문서요약

본 장에서는 비음수 행렬 분해를 기반으로 문장을 추출하여 주제기반의 다중 문서를 요약 할 수 있는 방법을 제안한다. 효율적인 다중 문서요약을 위해서는 다음과 같

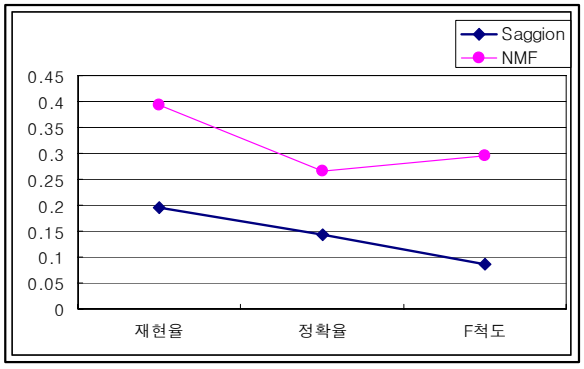
은 방법을 제안하였다. 첫째, 다중문서를 전 처리하여서 용어(term)과 문장(sentence)간의 행렬을 생성한다. 둘째, 비음수 의미 변수 행렬을 이용하여 문장을 군집하였다. 문장을 군집함으로써 유사한 문장들이 각각의 군집을 구성하여 중복 및 관련 없는 정보를 쉽게 구분 할 수 있으며, 문장 간의 다양성을 쉽게 보장하였다. 둘째, 주제와 군집의 비음수 의미 특징 행렬 간의 유사도를 계산하였다. 각각의 군집에서 주제와 유사도가 가장 높은 문장을 추출함으로써 주제와 관련된 정보를 최대한 반영하였다. 다음 그림1은 비음수 행렬 분해를 이용하여 다중문서 요약의 개요이다.



(그림1) 비음수 행렬 분해를 이용한 다중 문서요약

3. 실험 및 성능평가

본 논문에서 제안한 방법을 실험하기 위하여 '야후 코리아 뉴스'의 기사를 실험자료로 사용하였다[4]. 28개의 주제에 대하여 280건의 기사를 '야후 코리아 뉴스'에서 검색하였다. 성능 평가는 문선요약에서 주로 사용되는 정확률(Precision), 재현율(Recall), F-measure를 이용하였다[2, 7].



(그림 2) Saggion 방법과 제안방법의 비교 실험 결과

본 논문에서는 Saggion 방법[9]에 제안된 방법을 비교하였다. 다음 그림2는 실험 방법을 비교 평가한 결과이다. 실험에서 보듯이 제안된 방법이 더 좋은 성능을 보인다. 제안방법이 비음수 값과 부분정보를 이용하는 인간의 인식과정[5]과 유사한 과정으로 문서를 처리하기 때문이다.

4. 결론

본 논문은 다중 문서 요약을 위해서 비음수 의미 특징 행렬과 비음수 의미 변수 행렬을 이용하여 문장을 추출하는 새로운 방법을 제안하였다. 제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 문서의 고유 구조가 반영된 비음수 의미 특징 행렬을 이용하여 문장을 추출하기 때문에 요약의 질을 높일 수 있다. 둘째, 다중문서에 포함된 문장 간의 유사성과 다양성을 유도할 수 있도록 비음수 의미 변수 행렬을 이용하여 쉽게 문서를 군집할 수 있다. 마지막으로 군집된 문서들로부터 쉽게 과잉정보를 제거하고, 가장 중요한 문장만을 간소화된 형식으로 사용자에게 제공할 수 있다.

참고문헌

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

[2] 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)

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

[4] http://kr.news.yahoo.com/ (2009)

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

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

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

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

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