한국항행학회 논문지 제12권 제4호 2008년 8월

의미 특징 행렬과 의미 가변행렬을 이용한 질의 기반의 문서 요약

Query-Based Summarization using Semantic Feature Matrix and Semantic Variable Matrix

박 선*

Sun Park*

요 약

본 논문은 의미특징행렬(semantic feature matrix)과 의미변수행렬(semantic variable matrix)을 이용하는 질의 기반의 새로운 문서를 요약방법을 제안한다. 제안된 방법은 비지도 학습 방법으로 질의와 문장 간에 사전 학습이 필요 없고, 의미 특징(semantic feature)과 의미변수(semantic variable)를 이용하여 질의에 적합한 하위 주제를 잘 반영하여서 정확한 문서를 요약 할 수 있다. 이것은 비음수 행렬 분해가 주제들로 구성된 문서의 내부구조를 나타내는 의미특징을 자연스럽게 추출할 수 있기 때문이다. 실험결과 제안방법이 다른 방법에 비하여 좋은 성능을 보인다.

Abstract

This paper proposes a new query-based document summarization method using the semantic feature matrix and the semantic variable matrix. The proposed method doesn't need the training phase using training data comprising queries and query specific documents. And it exactly summarizes documents for the given query by using semantic features and semantic variables that is better at identifying sub-topics of document. Because the NMF have a great power to naturally extract semantic features representing the inherent structure of a document. The experimental results show that the proposed method achieves better performance than other methods.

Key words : Query based document summarization, Non-negative matrix factorization, Semantic feature matrix, Semantic variable matrix, Uunsupervised method

I. 서 론

인터넷 상의 문서 검색이나 개인화된 정보처리의 중요성이 증가하고 있다. 이 때문에 특정 사용자의 관심사항에 중점을 두는 문서요약의 필요성이 점점 증가 하고 있다. 문서 요약은 문서의 기본적인 내용을 유지하면서 문서의 량을 줄이는 작업이다. 문서의 요약은 제시되는 방법에 따서 문서 내용전체를 요약하는 포괄적 요약(generic summary)과 사용자의 질의

* 호남대학교 컴퓨터공학과(Department of compuer Engineering, Honam University)
· 제1저자 (First Author) : 박 선
· 투고일자 : 2008년 8월 4일
· 심사(수정)일자 : 2008년 8월 5일 (수정일자 : 2008년 8월 22일)
· 게재일자 : 2008년 8월 30일

에 따라 질의에 관련 있는 내용만을 포함하는 질의 기반 요약(query based summary)으로 나눌 수 있다[1].

본 논문은 NMF로 계산한 비음수 의미 특징 행렬과 질의간의 유사도를 이용하여 문장을 추출하여 문서를 요약하는 새로운 방법을 제안하였다. 비음수 행렬 인수분해는 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[2].

제안된 방법은 다음과 같다. 문서로부터 문장을 분해하고, 분해된 문장들은 벡터모델에 따라서 벡터로 표시된다[1] 이 문장벡터들은 NMF를 이용하여 비음수 의미특징 행렬(non-negative semantic feature matrix) 과 비음수 의미변수 행렬(non-negative semantic variable matirx)로 분해되고, 이 행렬들은 문서 내에서 문장들의 의미특징 벡터에 가중치인 의미변수를 곱한 값의 선형 합으로 표시된다. 즉, 의미특징 벡터는 문장의 내부 특징을 나타내며, 의미변수는 문장 내에서의 의미특징의 중요도를 나타낸다. 주제와 유사도 값이 가장 큰 의미특징 열벡터를 선택하고, 이 의미특징 열벡터에 대응되는 의미변수 행 벡터를 선택한다. 선택한 의미변수 행벡터의 문장의 적합도를 계산하여서 가장 큰 적합도에 대응되는 문장을 추출한다. 미리 설정된 요약문장의 개수가 만족할 때까지 질의와 다음으로 유사도가 높은 의미특징으로부터 문장을 추출한다.

제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 비지도학습에 의한 문서요약으로 사전 질의와 문서간의 학습이 필요 없다. 둘째, 의미 특질들이 문서내의 하위주제를 잘 파악하기 때문에 질의에 적합한 문장을 추출하여 문서요약의 질을 높일 수 있다. 셋째, 의미변수의 문장의 적합도를 계산하여 의미특징의 값은 높으나 전체 문서들에서 별로 중요하게 나타나지 않는 문장들이 선택되는 오류를 피할 수 있다. 마지막으로 단일문서뿐만 아니라 다중문서에서도 쉽게 문서요약을 할 수 있다.

본 논문의 구성은 다음과 같다. 제2장은 관련연구를, 제3장에서는 제안한 문서요약방법을, 제4장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제5장에서는 결론을 맺는다.

Ⅱ. 관련연구

질의 기반의 문서요약에 대한 최근 연구는 다음과 같다. 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 요약한다. 이들의 방법은 요약되는 문장이 장문 일 때에는 효과적이나 요약되는 문장이 단문 일 때는 좋은 성능을 보장하지 못한다[3].

Sassion은 주제기반의 다중문서요약 방법을 제안하였다. 이들의 제안방법은 질의와 문장 간의 유사도와 n-gram유사도를 이용하여 후보문장집합을 생성하고, 사용자가 지정한 압축율까지 후보문장집합으로부터 문장을 제거하여 문서를 요약한다[4].

Park이외의 저자들은 NMF를 이용한 질의기반의 문서요약 방법을 제안하였다. 이들의 제안방법은 질의와 의미특징간의 유사도를 이용하여 문서의 내부구조를 반영한 요약문을 생성하나, 의미특징의 값이 높으나 실제로 전체문장에서 중요하지 않은 문장이 선택될 수도 있는 단점을 가진다[5].

Ⅲ. 질의기반의 문서요약

본 장에서는 NMF를 이용하여 문장을 추출하는 질의 기반의 문서요약방법을 제안한다. 제안 방법은 전처리 단계와 문서 추출에 의한 문서요약 단계로 이루어진다. 다음 장에서는 NMF를 이용하여 의미특징 행렬과 의미변수 행렬을 계산하는 방법을 설명하고, 제안된 문서요약의 전처리 단계와 문서추출단계에 대하여 자세히 기술 한다. 다음 (그림1)은 제안된 질의기반 문서요약 방법의 개요이다.

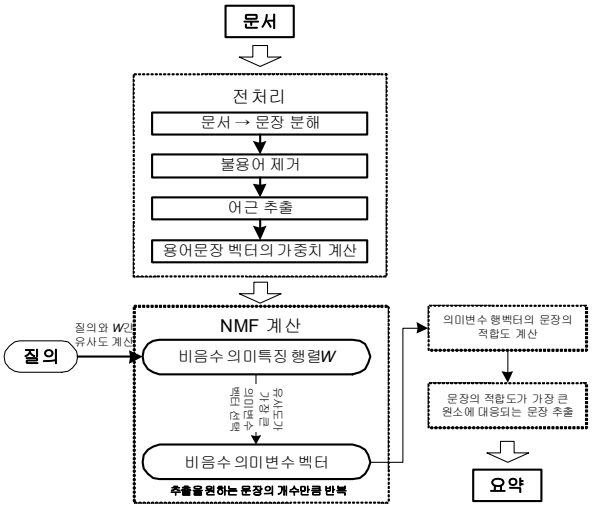

그림 1. 의미특징과 의미변수를 이용한 질의기반 문서요약 방법

Fig. 1. Query-based document summarization method using semantic feature and semantic variable

3-1 NMF

NMF는 주어진 비음수 행렬로부터 비음수의 인수를 찾는 행렬분해 알고리즘이다[2]. 비음수 행렬 분해 알고리즘은 식(1)의 목표함수 J 가 0에 가깝게 수렴 할 때까지 식(2)를 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.

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T V)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}} \ ,$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(VH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \tag{2}$$

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

행렬 A의 j번째 열벡터 A*j는 행렬 W의 l번째 열벡터 W*l와 행렬 H의 요소 hkj가 선형조합을 이루며 식(3)과 같다.

$$A_{*j} = \sum_{l=1}^{r} h_{kj} W_{*l} \tag{3}$$

3-2 전처리 단계

전처리 단계는 주어진 문서를 각각의 문장으로 분해 후, 불용어 제거, 어근추출, 가중치 계산으로 이루어진다[6]. 이후 용어빈도 벡터를 생성하고 식(4)를 이용하여 가중치를 계산한다[7].

벡터 $T_i = [\ t_{1i}, t_{2i}, \cdots, t_{ni}\ ]^T$는 i번째 문장의 용어빈도이다. 여기서 요소 t_{ji}는 i번째 문장에서 출현한 j번째 용어의 빈도이다. i번째 문장 A_i는 가중치가 부여된 용어빈도 벡터 $A_i = [\ a_{1i}, a_{2i}, \cdots, a_{ni}\ ]^T$로 표현되고, 벡터 A_i 요인 a_{ji}는 식(4)와 같이 정의된다.

$$a_{ji} = L(t_{ji}) \cdot G(t_{ji}) \tag{4}$$

여기서 $L(t_{ji})$는 i번째 문장에서 j번째 용어를 위한 지역 가중치(local weight)이고, $G(t_{ji})$는 문서 전체에서 j번째 용어를 위한 전역 가중치(global weight)로 다음 식(5)와 같이 정의된다.

$$L(i) = tf(i), \ G(i) = \log(N/n(i)) \tag{5}$$

여기서, $tf(i)$는 문장에서 i번째 용어가 출현한 빈도, N은 문서에서 문장의 총 개수이다. $n(i)$는 i번째 용어를 포함한 문장의 개수이다.

3-3 문서요약 단계

문서요약 단계는 비음수 의미 특징 행렬과 질의 사이의 유사도를 계산하여 유사도가 가장 높은 문장을 추출한다. 여기서 사용되는 식은 다음과 같다. 문서에서 총 m개의 용어와 n개의 문장으로 이루어진 m×n 행렬 A는 A = [A1, A2, ···, An]로 나타내며, 각 행 벡터 Ai는 i번째 문장의 가중치 용어빈도 벡터

이다.

주어진 행렬 A를 비음수 행렬 인수분해 하면 다음 식(6)과 같다[2].

$$A \approx WH \tag{6}$$

여기서, A는 n×m 행렬이고, n×r 행렬 W와 r×m 행렬 H는 행렬 A로부터 근사값으로 분해 된 행렬이며, 각각 W = [w1, w2, ..., wi], H = [h1, h2, ..., hj]로 나타낸다.

의미 특징 벡터 와 질의 와의 유사도를 구하는 식은 (7)과 같다.

$$sim(\vec{w},\vec{q}) = \frac{\vec{w}\cdot\vec{q}}{|\vec{w}|\times|\vec{q}|} \tag{7}$$

여기서 wij 는 j번째 r계수에서의 i번째 의미 특징인 요소이고 (wij ≥ 0), wiq 는 i번째 의미 특징 요소와 일치하는 q번째 질의의 용어이다 (wiq ≥ 0). m 은 r열 벡터의 요소들의 수로, 벡터 $\vec{q}$ = (w1q, w2q, ⋯, wmq)로 나타낸다[1].

문장의 적합도를 식 (8)과 같이 정의한다. 식(9)의 weight (Hi*)는 모든 의미특징들 중에서 i번째 의미특징 (W*i) 의 적합도와의 관계를 의미하고, 식(8)의 문장의 적합도는 의미특징에 의해서 표현되는 주요주제가 문장에 얼마나 반영되는가를 의미한다. 즉, 식(9)를 사용함으로써 의미특징의 값은 높으나 전체 문서들에서 별로 중요하게 나타나지 않는 문장들이 선택되는 오류를 피할 수 있다.

RS (relevance of a sentence) j =

$$\sum_{i=1}^{r} H_{ij} \cdot weight(H_{i*}) \tag{8}$$

$$weight(H_{i*}) = \frac{\sum_{q=1}^{n} H_{iq}}{\sum_{p=1}^{r}\sum_{q=1}^{n} H_{pq}} \tag{9}$$

비음수 의미 특징 행렬을 이용한 문서요약 방법은 다음과 같다.

1. 문서를 개개의 문장으로 분해하고, 분해 할 계수 r의 수를 지정한다.
2. 각각의 문장에 대한 불용어 제거 및 어근추출 후, 식(4)를 이용하여 용어빈도 벡터의 가중치를 계산하여 용어문장 가중치 행렬 A를 구성한다.
3. 행렬 A에 식(1)과 식(2)를 적용하여 식(6)과 같은 비음수 행렬 W, H로 분해한다.
4. 식(7)을 이용하여 행렬 W의 열 벡터들과 질의 간의 유사도를 계산하여 가장 유사도가 높은 p번째 열 벡터 W*p를 찾는다.
5. 행렬 H에서 p번째 행에 포함된 행 벡터 Hp*에서 문장의 적합도를 계산하고, 적합도가 가장 큰 요소 값을 가진 q열과 같은 열에 있는 행렬 A의 문장 벡터 A*q에 대응되는 문장을 선택한다.

위의 4번째 단계에서 질의와 유사도가 가장 높은 열 벡터 W*p는 질의와 연관이 있는 가장 중요한 특징들의 집합이다.

Ⅳ. 실험 및 평가

본 논문에서 제안한 방법의 성능을 평가하기 위하여 DUC2005의 평가 자료로 실험하였다. DUC은 전문가들이 작성한 이상적인 요약문과 제안된 시스템이 만든 요약문을 비교하여 각 시스템의 성능을 평가하는 국제회의다. DUC 2005의 평가 자료는 50개의 주제와 주제에 관련된 25-50개의 관련된 문서로 구성되어 있다[8].

실험에 대한 평가방법은 DUC에서 평가 방법으로 사용되고 있는 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)를 이용하였다[9]. ROUGE는 ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, ROUGE-SU 등의 자동평가 방법을 포함하고 있다[9].

실험은 네 가지 요약방법들에 대하여서 성능을 비교 평가한 것이다. 그림 2, 3, 4는 각각 평균 재현율, 평균 정확률, 평균 f-measure 에 대한 평가 결과이다. 여기서 LSA는 Gong[10]의 방법으로 LSA를 이용한 문서요약 방법이고, N-gram은 Saggion[4]이 제안한 방법으로 문장간의 유사도와 n-gram 유사도를 이용

한 문서요약 방법이다. NMF는 이전에 저자들이 제안한 방법으로 NMF의 의미특징을 이용한 문서요약 방법이다 [5]. RS-NMF는 본 논문에서 제안한 방법으로 NMF방법에 문장의 적합도를 계산하는 방법이다. 다음은 제안방법의 성능을 분석한 것이다.

그림 2에서 보는 것과 같이 제안 방법(RS-NMF)의 평균 재현율이 LSA에 비하여 10.5%가 높으며, N-gram에 비해서는 40.7%가 높고, NMF에 비해서는 6.5%가 높다. 그림 3에서 보는 것과 같이 제안 방법의 평균 정확율이 LSA에 비하여 23.8%가 높으며, N-gram에 비해서는 9.8%가 높고, NMF에 비해서는 6.6%가 높다. 그림 4에서 보는 것과 같이 제안 방법의 f-measure가 LSA에 비하여 22.0%가 높으며, N-gram에 비해서는 14.8%가 높고, NMF에 비해서는 4.9%가 높다.

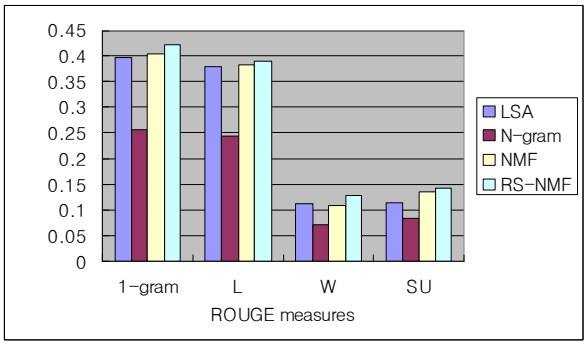

그림 2. 비교방법들의 재현율 평가 결과
Fig. 2. Result of recall of comparison methods

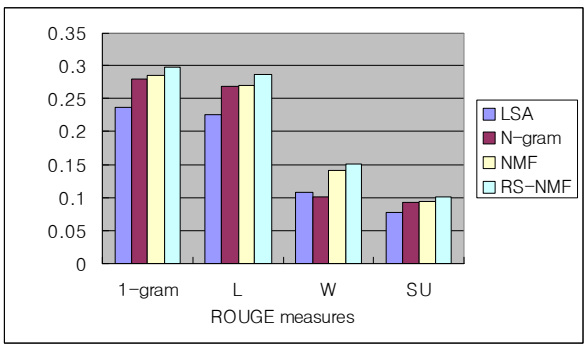

그림 3. 비교 방법들의 정확률 평가 결과
Fig. 3. Result of precision of comparison methods

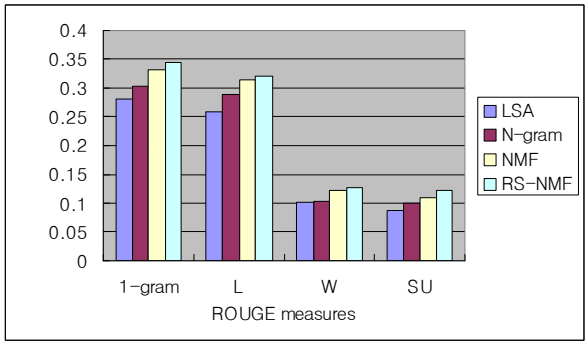

그림 4. 비교방법들의 f-measure 평가 결과
Fig. 4. Result of f-measure of comparison methods

실험결과 제안방법인 RS-NMF가장 좋은 결과를 나타내며, 그 다음으로 NMF가 좋은 결과를 나타낸다. 평균 재현율 평가에서는 N-gram 방법이 가장 성능이 좋지 않으며, 평균 정확율과 f-measure 평가에서는 LSA가 가장 성능이 저조하다. 재현율 평가에서 LSA가 N-gram보다 좋은 것은 LSA의 문서의 잠재구조가 N-gram의 문장내의 연속적인 패턴보다 참조요약문에 더 적합한 것으로 보이며, 정확율과 f-measure에서는 LSA의 잠재구조보다 문장내의 용어들 간의 연속적인 관계가 더욱 정확한 후보 요약문 생성에 영향을 미치는 것을 알 수 있다. 제안방법인 RS-NMF가 NMF의 질의와 관련된 유사도뿐만 아니라 문서 내부의 고유 특성을 반영하는 방법에 비하여 문서의 고유구조에 문장의 적합도를 계산하여 문서를 요약한다. 이 때문에 의미특징의 값은 높지만 실제로 중요하지 않는 문장이 추출되는 것을 최소화 시켜서 비교방법들에 비하여 더욱 의미 있는 문장을 추출한다.

Ⅴ. 결 론

본 논문은 비음수 의미 특징 행렬을 이용하여 질의 기반의 문서를 요약하는 새로운 방법을 제안하였다. 질의기반의 문서요약에서는 질의와 관련성이 높으면서 문서의 공통적인 정보를 포함하는 것이 중요하다. 우리는 제안방법에서 문서의 내부구조를 나타내는 의미특징과 질의와의 유사도를 이용하여 질의와 관련성이 높으면서도 문서의 내부특징을 반영하여 의미 있는 문장을 추출하였다. 제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 사전 질의와 문서간

의 학습이 필요 없이 문서의 내부특징과 질의 만을 이용하여 문장을 추출한다. 둘째, 의미 특질과 의미변수를 이용하여 의미 있는 문장을 추출함으로써 문서요약의 질을 높일 수 있다. 마지막으로 의미변수에 대한 문장의 적합도를 계산하여 의미특징의 값이 크나 실제 중요하지 않은 문장이 추출되는 것을 최소화 시킨다.

앞으로 제안 방법의 성능 향상을 위하여 다양한 종류의 가중치 및 전처리 방안에 대한 연구를 진행시켜야 하며, 문서요약의 성능 향상을 위하여 비음수 행렬 W, H간의 관계에 대한 깊이 있는 연구가 진행되어야 할 것이다.

참 고 문 헌

[1] Mani, I., "Automatic Summarization", *John Benjamins Publishing Company*, 2001.

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

[3] Sakurai, T., Utsumi, A., "Query-based Multidocument Summarization for Information Retrieval", *The Proceeding of NTCIR'04*, 2004.

[4] Sassion. H. "Topic-based Summarization at DUC 2005", *In Proceedings of the DUC'05*, 2005.

[5] Park, S., Lee, J. H., Ahn, C. M., Hong, J. S., Chun, S. J.: Query Based Summarization using Non-negative Matrix Factorization. In proceeding of KES, 84-89, 2006.

[6] Frankes, W. B., Ricardo, B. Y., "Information Retrieval : Data Structure & Algorithms", *Prentice-Hall,* 1992.

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

[8] Hoa., H., D.: Overview of DUC 2005. In Proceedings of the DUC, 2005.

[9] Chin-Yew, L.: ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL, 2004.

[10] Gong, Y., Liu, X., "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis", *In Proceeding of ACM SIGIR'05,* pp.19-25, 2001.

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