멀티모달 방법론과 텍스트 마이닝 기반의 뉴스 비디오 마이닝

(A News Video Mining based on Multi-modal Approach and Text Mining)

이 한 성 † 임 영 희 †† 유 재 학 †† 오 승 근 ††† 박 대 희 ††
(Hansung Lee) (Younghee Im) (Jaehak Yu) (Seunggeun Oh) (Daihee Park)

요 약 정보·통신기술이 발전함에 따라 멀티미디어 데이터를 포함하는 디지털 기록물의 양은 기하급수적으로 증가하고 있다. 특히 뉴스 비디오는 시대상을 반영하는 풍부한 정보를 내포하고 있으므로, 이를 효과적으로 관리하고 분석하기 위한 뉴스 비디오 데이터베이스 및 뉴스 비디오 마이닝은 광범위하게 연구되어왔다. 그러나 현재까지의 뉴스 비디오 관련 연구들은 뉴스 기사에 대한 브라우징, 검색, 요약에 치중되어 있으며, 뉴스 비디오에 내재되어 있는 풍부한 잠재적 지식을 탐사하는 고수준의 의미 분석 단계에는 이르지 못하고 있다. 본 논문에서는 뉴스 비디오 클립과 스크립트를 동시에 이용하는, 멀티모달 방법론과 텍스트 마이닝 기반의 뉴스 비디오 마이닝 시스템을 제안한다. 제안된 시스템은 텍스트 마이닝의 군집분석을 통해 뉴스 기사들을 자동 분류하고, 분류 결과에 대해 기간별 군집 추이그래프, 군집성장도 분석 및 네트워크 분석을 수행함으로써, 뉴스 비디오의 기사별 주제와 관련한 다각적 분석을 수행한다. 제안된 시스템의 타당성 검증을 위하여 "2007년 제2차 남북 정상회담" 관련 뉴스 비디오를 대상으로 뉴스 비디오 분석을 수행하였다.

키워드 : 뉴스 비디오, 비디오 마이닝, 유사도 행렬, 계층적 군집화, 네트워크 분석

Abstract With rapid growth of information and computer communication technologies, the numbers of digital documents including multimedia data have been recently exploded. In particular, news video database and news video mining have became the subject of extensive research, to develop effective and efficient tools for manipulation and analysis of news videos, because of their information richness. However, many research focus on browsing, retrieval and summarization of news videos. Up to date, it is a relatively early state to discover and to analyse the plentiful latent semantic knowledge from news videos. In this paper, we propose the news video mining system based on multi-modal approach and text mining, which uses the visual-textual information of news video clips and their scripts. The proposed system systematically constructs a taxonomy of news video stories in automatic manner with hierarchical clustering algorithm which is one of text mining methods. Then, it multilaterally analyzes the topics of news video stories by means of time-cluster trend graph, weighted cluster growth index, and network analysis. To clarify the validity of our approach, we analyzed the news videos on "The Second Summit of South and North Korea in 2007".

Key words : news video, video mining, similarity matrix, hierarchical clustering, network analysis

· 본 연구는 교육과학기술부와 한국산업기술재단의 지역혁신인력양성사업으로 수행된 연구결과임

† 정 회 원 : 한국전자통신연구원 휴먼인식기술연구팀 선임연구원
mohan@etri.re.kr
†† 정 회 원 : 고려대학교 세종캠퍼스 컴퓨터정보학과 교수
yheeim@korea.ac.kr
dbzzang@korea.ac.kr
dhpark@korea.ac.kr
(Corresponding author임)
††† 학생회원 : 고려대학교 전산학과
gmo85@korea.ac.kr

논문접수 : 2009년 11월 12일
심사완료 : 2010년 4월 28일

정보과학회논문지: 데이타베이스 제37권 제3호(2010.6)

1. 서 론

정보·통신 기술이 발전함에 따라 멀티미디어 데이터를 포함한 다양한 디지털 기록물들이 급증하고 있다. 따라서 사용자가 방대한 양의 디지털 기록물들로부터 적합한 지식 정보를 검색, 분류, 분석하는 것은 점점 더 어려워지고 있다. 결국, 대용량의 디지털 기록물들로부터 의미 있는 지식을 추출하기 위한 지식탐사 기법에 대한 요구가 증가하고 있다. 특히 뉴스 비디오는 정치, 경제, 사회, 문화 등 시대상을 반영하는 풍부한 정보를 내포하고 있으므로, 이를 효과적으로 관리하고 분석하기 위한 뉴스 비디오 데이터베이스 및 뉴스 비디오 마이닝 기법에 대한 연구가 요구된다[1,2].

문헌 조사에 따르면, 대부분의 뉴스 비디오 관련 연구들은 뉴스 비디오의 브라우징(browsing)과 검색을 위한 전처리 과정에 해당하는 뉴스 비디오 파싱 및 색인과 같은 내부구조 분석(intra-structure analysis)에 집중되어 있다[1-9]. 뉴스 비디오 파싱은 뉴스 비디오를 효과적으로 색인화하기 위한 방법론으로서, 크게 뉴스 비디오 클립을 샷 단위로 분할하는 샷 경계 탐지(shot boundary detection)[1-3], 분할된 샷들 중에서 앵커가 포함된 샷을 찾기 위한 앵커 샷 탐지(anchor shot detection)[4-6], 뉴스 비디오 클립을 기사 단위로 분할하는 뉴스 기사 경계 탐지(news story boundary detection)[7-9]로 구성된다. 반면 뉴스 기사간의 관계 분석(inter-structure analysis)에 대한 연구로는 뉴스 비디오의 토픽 스레딩(topic-threading)[10] 및 오토 다큐멘팅(auto-documenting)[11] 등의 연구가 진행 중이다. 뉴스 비디오의 토픽 스레딩은 뉴스 비디오를 미리 정의되어 있는 주제로 분류한 후, 사용자로 하여금 관심 있는 주제에 대해 뉴스 기사들을 브라우징 및 검색할 수 있도록 지원하는 방법이다[10]. 오토 다큐멘팅은 뉴스 기사들을 주제별로 분류하고, 각 주제에 대하여 토픽 트리(topic tree)를 생성함으로써 뉴스 기사에 대한 요약 및 뉴스 기사의 전개 상황을 자동으로 생성하는 방법이다[11]. 결국, 현재까지의 뉴스 비디오 관련 연구들은 뉴스 기사에 대한 브라우징, 검색, 요약에 치중되어 있으며, 뉴스 비디오에 내재되어 있는 풍부한 잠재적 지식을 탐사하는 고수준의 의미 분석 단계에는 이르지 못하고 있다.

한편, 뉴스 비디오에서의 의미론적 격차(semantic gap)에 대한 해결 방법으로써, 동영상의 저수준 특징 정보와 뉴스 비디오의 구조적 성격상 함께 제공되는 오디오 혹은 자막 및 스크립트(script)와 같은 텍스트 기반의 고수준 의미 정보를 동시에 활용하는 멀티모달 방법론(multi-modal approach)[10-13]에 관한 연구가 진행 중이다. I. Ide 등은 뉴스 비디오의 자막 및 스크립트로부터 단어들을 추출하여 미리 정의되어 있는 주제로 분류한 후, 사용자가 원하는 주제에 관한 뉴스 기사들을 브라우징 및 검색할 수 있도록 지원하였다[10]. J. Kim 등은 자막 정보를 이용하여 뉴스 비디오에서 의미 있는(semantically meaningful) 장면(highlight)을 추출하고, 오디오 정보를 이용하여 자막과 뉴스 비디오의 시간정렬을 수행함으로써 뉴스 비디오를 효율적으로 요약하고 색인할 수 있는 방법을 제안하였다[12]. J. Shen 등은 비디오의 비주얼 정보와 오디오 정보를 서브 스페이스(subspace)로 매핑 하여 의미 있는 이벤트를 탐지하는 방법론을 제시하였다[13]. 그러나 이러한 연구방법들도 멀티모달 방법론에 의한 의미론적 격차의 중개자(bridge)로써의 역할과는 달리 뉴스 기사에 대한 브라우징, 검색, 요약등과 같은 기능만을 수행하고 있다.

반면, 문서 정보에 대한 검색 및 분석을 지원하는 텍스트 마이닝 분야에 대한 연구는 이미 상당한 수준에 올라와 있으며, 다양한 디지털 기록매체를 효율적으로 분류 및 분석하기 위하여 전산학과 문헌정보학의 학제간 연구가 활발히 진행되고 있다. 특히 지적구조 분석에 텍스트 마이닝의 군집화 기법을 적용하려는 시도가 진행되고 있다[14-17]. 디지털 기록물들에 내재되어 있는 서지 정보들을 분석하여 새로운 지식을 도출해내는 계량정보 분석 방법에서는 관련 문서를 체계적으로 분류 및 분석하여 원하는 정보를 쉽게 탐색하거나, 텍스트 마이닝의 한 기법인 관련용어간의 유사도 정도에 따라 문서를 군집화 하는 방법이 주로 사용되고 있다. 텍스트 마이닝은 자연어로 구성된 비구조적인 텍스트 안에서 패턴 또는 관계를 추출하여 지식을 발견하는 과정으로서, 주로 텍스트의 자동 분류작업이나 텍스트로부터 새로운 지식을 생성하는 작업에 활용된다[18].

본 논문에서는 뉴스 비디오의 고수준 분석을 위하여 뉴스 비디오 클립(news video clip)과 스크립트를 동시에 이용하는 멀티모달 방법론을 취하고자 한다. 단, 전술한 기존의 방법론과는 달리 이미 기술적으로 성숙한 텍스트 마이닝 기법을 스크립트에 적용함으로써, 뉴스 비디오 분석의 기술적 한계를 간접적으로 극복하고자 한다. 즉, 뉴스 비디오 마이닝의 전처리과정으로서, 뉴스 기사별 대표 영상(key-frame) 추출과 해당 스크립트의 형태소 분석 작업을 통하여 영상 정보와 텍스트 정보를 모두 이용한 뉴스 기사 간 유사도 행렬을 생성한다. 생성된 유사도 행렬을 기반으로 텍스트 마이닝의 계층적 군집분석을 수행하여 뉴스 기사들을 자동 분류하고, 군집분석 결과에 대한 기간별 군집 추이그래프와 군집성장도 분석을 수행하여 뉴스 기사를 다각적으로 분석한다. 또한 군집 네트워크 분석을 통하여 군집간의 상관관

계 분석을 수행한다. 결국, 뉴스 비디오를 기사 주제별로 분류하고, 뉴스 기사들 간의 연관성 및 해당 뉴스 기사의 전개 상황, 추이 분석 등을 수행한다. 제안하는 방법론의 타당성 검증을 위해 "2007년 제2차 남북 정상회담" 관련 뉴스 기사들을 대상으로 뉴스 비디오 분석을 수행한다. 본 논문에서 제안하는 방법론은 멀티모달 방법론과 텍스트 마이닝을 뉴스 비디오에 적용함으로써, 기존의 뉴스 비디오 분석 기술력의 한계점을 극복할 수 있는 의미 있는 연구라고 할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제안하는 뉴스 비디오 마이닝 시스템의 개괄적인 구조에 대해 기술하고, 뉴스 비디오 마이닝을 위한 전처리 과정과 계층적 군집화 알고리즘을 기반으로 하는 뉴스 비디오 분석 단계에 대해 설명한다. 3장에서는 사례 연구를 통한 뉴스 비디오 분석을 기술한다. 마지막으로 4장에서는 결론 및 향후 연구 방향을 논한다.

2. 뉴스 비디오 마이닝 시스템

본 논문의 뉴스 비디오 마이닝 시스템은 뉴스 비디오 클립과 스크립트를 모두 이용하여 관련 뉴스 기사들에 대한 다각적이고 심층적인 분석을 수행한다. 제안하는 뉴스 비디오 마이닝 시스템은 크게 전처리 과정과 데이터 분석 과정으로 구성된다. 전처리 과정은 뉴스 비디오를 기사 단위로 나누는 뉴스 비디오 기사 분할 과정과 분할된 뉴스 기사 각각에 대하여 대표 영상 및 스크립트를 추출하고 코클러스터링(co-clustering)을 이용하여 뉴스 기사들을 주제별로 분류하는 과정으로 구성된다. 심층적인 분석이 요구되는 주제가 선택되면, 코클러스터링을 통해 해당 주제로 분류된 기사들에 대해 계층적 군집 분석을 수행하고, 군집 분석 결과에 대한 다각적 분석을 수행한다. 계층적 군집 분석 수행 시, 전처리 과정에서 추출된 영상 정보와 텍스트 정보를 모두 이용함으로써 보다 정확한 분석을 수행하고자 한다. 본 논문에서 제안하는 뉴스 비디오 마이닝의 전체 시스템 구조는 그림 1과 같다. 본 논문에서는 그림 1의 아래 부분에 사각형으로 표시된 데이터분석 모듈을 중점적으로 설명하고자 한다.

2.1 전처리 과정

뉴스 비디오 마이닝 시스템의 전처리 과정은 본 연구의 선행 연구 결과물[1,2]로써, 본 논문에서는 전처리 과정을 개괄적으로 요약하여 서술한다.

2.1.1 뉴스 비디오 기사 분할 과정

뉴스 비디오를 효과적으로 분류하고 분석하기 위한 가장 중요한 과정은 뉴스 비디오를 기사 단위로 나누는 뉴스 비디오 기사 분할이다. 뉴스 비디오 기사 분할은 일반적으로 샷 경계 탐지, 앵커 샷 탐지, 그리고 뉴스 기사 경계 탐지의 세 단계로 구성 된다[1-9]. 뉴스 비디오 기사 분할의 첫 단계는 뉴스 비디오의 샷과 샷 사이의 장면 변환의 경계를 탐지하여 비디오를 샷들로 분할하는 샷 경계 탐지이다. 본 논문에서는 뉴스 비디오 분할에 적합하도록 최적화된 샷 경계 탐지 알고리즘을 이용하고자 한다[2]. 본 논문에서 사용된 샷 경계 탐지 알고리즘은 SVD(singular value decomposition)를 기반으로 점증적 클러스터링 알고리즘인 ART(adaptive reso-

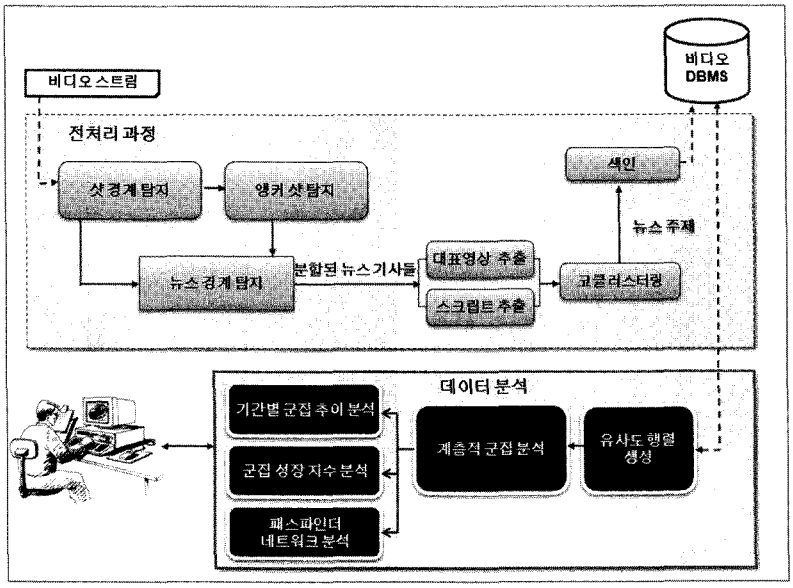

그림 1 뉴스 비디오 마이닝 시스템의 구조도

nance theory)에 커널 대치(kernel substitution) 혹은 커널 트릭(kernel trick)이라 불리는 기법을 적용한 알고리즘으로 다음과 같은 특징을 갖는다[2]. 1) 급격한 장면 변환과 점진적 장면 변환을 하나의 알고리즘으로 탐지하여 한 번의 데이터 탐색으로 샷 분할을 수행한다; 2) 뉴스 비디오 샷 경계 탐지의 재현율을 높임으로써 앵커 샷 탐지 단계의 입력으로 사용되는 데이터의 오류를 최소화한다; 3) 분할된 샷들을 정적 샷과 동적 샷으로 분류함으로써 앵커 샷 탐지 단계의 탐색 공간을 축소한다; 4) SVD를 통해 뉴스 비디오를 구성하는 연속적인 프레임에서 잡음과 아주 작은 변화를 제거하여 분류 성능을 높일 뿐만 아니라 커널 방법을 도입함으로써 서로 가까이 매핑되어 있는 샷들을 보다 효과적으로 탐지한다.

뉴스 비디오 기사 분할의 두 번째 단계는 분할된 샷들 중 앵커가 존재하는 샷을 찾는 앵커 샷 탐지 단계이다. 본 논문에서는 다단계 구조를 가지고 있는 MASD (multi-phase anchor shot detection)[6]을 이용하고자 한다. MASD는 4개의 구성요소로 구성되어 있다. 1) 얼굴 영역의 후보군을 찾기 위한 피부색 탐지기; 2) 얼굴을 포함하는 영상을 찾기 위한 얼굴 탐지기; 3) 검색된 대표 영상의 벡터 표현을 위한 NMF(non-negative matrix factorization); 4) 최종적으로 앵커 샷을 탐지하기 위한 단일 클래스 SVM 모듈들. MASD는 피부색 탐지 및 얼굴 탐지기를 통해 앵커 샷 탐지를 위한 탐색 공간을 대폭 축소하여 고속의 탐지가 가능할 뿐만 아니라 SVM을 통해 높은 탐지율을 보인다.

뉴스 비디오 기사 분할의 마지막 단계인 뉴스 기사 경계 탐지를 위해서 본 논문에서는 비교적 간단한 휴리스틱(heuristic)을 이용하고자 한다[9]. 전형적으로 뉴스 기사는 앵커 샷으로 시작을 하고 인터뷰와 기자의 취재 내용을 포함하는 뉴스 기사의 본 내용이 방영된다. 따라서 앵커 샷과 다음 앵커 샷 전까지가 하나의 뉴스 기사를 이룬다. 본 논문에서는 앞서 설명한 뉴스 비디오의 구조적 특징을 이용하여 앵커 샷들을 기준으로 뉴스 기사 경계를 탐지한다.

2.1.2 코클러스터링을 이용한 뉴스 주제별 분류과정

제안하는 뉴스 비디오 마이닝 시스템의 전처리를 위한 두 번째 단계는 분할된 뉴스 기사를 주제별로 분류하는 과정이다. 본 논문에서는 뉴스 비디오의 영상정보와 텍스트 정보를 대상으로 코클러스터링을 수행함으로써 분할된 뉴스 기사를 주제별로 분류하고자 한다[11]. 뉴스 기사별 스크립트로부터 형태소 분석을 통하여 단어들을 추출하고, 동영상 클립으로부터 대표 영상을 추출한다. 뉴스 기사별로 추출된 단어와 대표 영상을 이용하여 이분 그래프(bipartite graph) 모델을 생성한다.

뉴스 기사들의 집합을 $S=\{s_1,s_2,...,s_l\}$, 단어들의 집합을 $T=\{t_1,t_2,...,t_m\}$, 대표 영상의 집합을 $V=\{v_1,v_2,...,v_n\}$ 라고 정의하면, 이분 그래프 모델은 다음과 같이 정의된다.

$$G=(S,P,E) \quad (1)$$

여기서 $P=T\cup V$ 이며 다음의 두 조건 $T\cap V=\varnothing$, $P\cap S=\varnothing$ 을 만족한다. E는 뉴스 기사 S와 텍스트 및 영상 정보 P 사이의 방향성 없는 연결을 의미한다.

식 (1)의 이분 그래프 모델은 텍스트-뉴스 기사 행렬 A_1 및 영상-뉴스 기사행렬 A_2의 결합 행렬 A로 표현 표현할 수 있다.

$$A=\begin{bmatrix}A_1\\A_2\end{bmatrix} \quad (2)$$

여기서 A_1은 $A_1(i,j)=tf_{ij}\times idf_i$을 이용하여 계산하며, A_2는 $A_2(i,j)=kf_{ij}\times\log_2(N/sf_i)$을 이용하여 계산한다. tf_{ij}는 뉴스 기사 j에서 발생된 단어 i의 빈도이며, idf_i는 단어 i의 역문서 빈도(inverse document frequency)이다. kf_{ij}는 뉴스 기사 j에서 발생된 대표 영상 i의 빈도이며, N은 전체 뉴스 기사의 개수이다. sf_i는 대표 영상 i의 빈도를 의미한다.

뉴스 기사들을 주제별로 분류하기 위해서는 식 (2)의 이분 그래프를 그래프 분할 알고리즘에 의해 여러 개의 부분 그래프(sub-graph)로 분할한다. 본 논문에서는 [11]에서 제안된 코클러스터링 알고리즘을 이용하여 뉴스 기사들을 주제별로 분류하였다.

2.2 데이터 분석 과정

2.2.1 유사도 행렬

뉴스 비디오의 기사들을 대상으로 계층적 군집 분석을 수행하기 위해서는 뉴스 기사 간 유사도 행렬을 정의하여야 한다. 본 논문에서는 뉴스 비디오의 계층적 군집 분석을 위하여 뉴스 기사별 비디오 클립과 스크립트를 동시에 사용한다. 본 논문에서는 X. Wu 등[11]이 소개한 뉴스 비디오의 영상 정보와 텍스트 정보를 모두 이용한 유사도 함수를 이용하여 유사도 행렬을 정의한다. 계층적 군집 분석의 입력으로 사용될 뉴스 기사 간 유사도 행렬을 생성하기 위하여, 형태소 분석을 통하여 실험 데이터 집합에 포함되어 있는 스크립트로부터 단어를 추출하고, 동영상 클립으로부터 대표 영상을 추출한다. 추출된 단어와 대표 영상을 대상으로 식 (3)과 식 (4)의 tf-idf 함수를 이용하여 단어별 가중치 및 대표 영상의 가중치를 계산한다.

$$w_k(S_i)=\frac{tf_{w_k,S_i}}{tf_{w_k,S_i}+0.5+(1.5*\frac{len(S_i)}{awl})}*\frac{\log\frac{g+0.5}{sf_{wk}}}{\log(g+1.0)} \quad (3)$$

$$v_l(S_i)=\begin{cases}\dfrac{tf_{v_l,S_i}}{tf_{v_l,S_i}+0.5+(1.5*\dfrac{kflen(S_i)}{akfl})}*\dfrac{\log\dfrac{g+0.5}{sf_{vl}}}{\log(g+1.0)} & \text{if } v_l \notin NDK \\ \alpha\dfrac{tf_{v_l,S_i}}{tf_{v_l,S_i}+0.5+(1.5*\dfrac{kflen(S_i)}{akfl})}*\dfrac{\log\dfrac{g+0.5}{sf_{vl}}}{\log(g+1.0)} & \text{if } v_l \in NDK\end{cases} \quad (4)$$

여기서 $w_k(S_i)$는 뉴스 기사 S_i에 속해 있는 단어 w_k의 가중치를 의미하며, $v_l(S_i)$는 뉴스 기사 S_i에 속해 있는 대표 영상 v_l의 가중치를 의미한다. tf_{w_k,S_i}는 뉴스 기사 S_i에 속해 있는 단어 w_k의 빈도수이며, tf_{v_l,S_i}는 뉴스 기사 S_i에 속해있는 대표 영상 v_l의 빈도수 이다. aul는 뉴스 기사들에서 나타는 단어수의 평균을 의미하며, $akfl$는 뉴스 기사들에 나타나는 대표 영상수의 평균이다. sf_{wk}는 단어 w_k을 포함하고 있는 뉴스 기사의 수이며, sf_{vl}는 대표 영상 v_l을 포함하고 있는 뉴스 기사의 수를 의미한다. $len(S_i)$은 뉴스 기사 S_i에 속해 있는 단어수이며, $kflen(S_i)$은 뉴스 기사 S_i에 속해 있는 대표 영상의 수이다. g는 전체 뉴스 기사의 수를 의미하며, α는 중복되지 않는 대표 영상의 가중치 값을 의미한다.

뉴스 기사별 단어 가중치 및 대표 영상 가중치를 입력으로 식 (5)의 코사인 유사도 함수를 이용하여 뉴스 기사 간 유사도 행렬을 생성한다.

$$SIM(S_i,S_j)=\frac{\sum_{k=1}^{m}w_k(S_i)w_k(S_j)+\sum_{l=1}^{n}v_l(S_i)v_l(S_j)}{\sqrt{\sum_{k=1}^{m}w_k(S_i)^2\sum_{k=1}^{m}w_k(S_j)^2}\sqrt{\sum_{l=1}^{n}v_l(S_i)^2v_l(S_j)^2}} \quad (5)$$

여기서 m은 전체 단어의 개수이며, n은 전체 대표 영상의 개수를 의미한다.

2.2.2 계층적 군집화 알고리즘

본 논문에서는 계층적 군집화를 통한 뉴스 기사들의 분류 및 분석을 위하여 평균 연결법(average linkage)을 이용한 계층적 군집화 방법을 사용한다. 계층적 군집화란 처음 각 대상이 독립군집으로 출발하여, 유사도가 가장 큰 대상끼리 군집을 생성하고, 생성된 군집간의 비교를 통하여 상위 개념의 새로운 군집을 생성하는 작업을 반복하여 최종적으로 하나의 군집으로 묶는 방법이다. 본 논문에서 사용한 계층적 군집화 방법은 평균 연결법에 근간을 두고 있다. 평균 연결법이란 군집간의 거리를 계산할 때, 한 군집의 모든 구성원들과 다른 군집의 모든 구성원들 간의 거리의 평균을 기준으로 하는 방법이다[18]. 제일 근접한 U,V 개체를 한 군집으로 묶은 뒤 (U,V) 군집과 W군집과의 유사도는 다음과 같이 정의된다.

$$S_{(U,V)W}=\frac{\Sigma\Sigma s_{ik}}{N_{(U,V)}N_W} \quad (6)$$

여기서 s_{ik}는 군집 (U,V)의 i번째 개체와 군집 W의 k번째 개체간의 유사도를 의미한다. $N_{(U,V)}$와 N_W는 각각 군집 (U,V)의 개체 수 및 군집 W의 개체수를 의미한다.

형성된 군집과 다른 군집의 거리를 계산할 때 최장거리를 기준으로 하는 완전연결법은 군집내의 아웃라이어(outlier)가 최장거리 계산에 사용될 경우, 군집화의 결과가 잘못될 가능성이 있다. 따라서 본 논문에서는 각 군집에 포함된 모든 구성원들의 값을 사용하는 평균연결법[18]을 사용하였다.

2.2.3 시기별 군집 성장도

본 논문에서는 각 군집 소속 뉴스 기사의 방송 시기별 분포를 확인하기 위하여 J. Lee 등[14]이 제안한 시기별 분석을 수행하였다. 전체 기간을 1기와 2기로 나누고, 식 (7)의 군집 성장 지수(CGI: cluster growth index)를 계산한다. 군집 성장 지수는 1기와 비교하여 2기에 뉴스 기사가 증가하여 군집이 성장한 정도를 상대적 비율로 비교하기 위한 지표이다.

$$CGI=\frac{\text{2기 뉴스 기사 수}-\text{1기 뉴스 기사 수}}{\text{2기 뉴스 기사 수}+\text{1기 뉴스 기사 수}} \quad (7)$$

군집 성장 지수는 상대적 군집 성장도를 나타내는 지표로서, 절대적 뉴스 기사의 증가 정도를 반영하기 어렵다. 따라서 군집 성장 지수에 각 군집의 성장 규모를 반영하는 지수로 다음과 같은 가중 군집 성장 지수(WCGI: Weighted CGI)를 산출할 수 있다.

$$WCGI=|\text{2기 뉴스 기사 수}-\text{1기 뉴스 기사 수}|\times CGI \quad (8)$$

군집 성장 지수가 상대적 비율을 이용한 지표인데 반해, 가중 군집 성장 지수는 절대적인 뉴스 기사의 증가를 반영하는 지표가 된다. 즉, 가중 군집 성장 지수는 증가한 뉴스 기사 수에 비례하여 값이 커진다. 따라서 가중군집 성장 지수가 보다 분별력 있는 지표이며, 본 논문에서는 가중 군집 성장 지수를 사용하여 군집의 시기별 분석을 수행한다.

2.2.4 패스파인더 네트워크

본 논문에서는 군집간의 상관관계 및 연관성을 확인하기 위하여 패스파인더 네트워크 스케일링(pathfinder network scaling)을 이용한 네트워크 분석(network analysis)을 수행한다. 패스파인더 네트워크는 가중치가 있는 모든 링크가 생성된 상태에서 삼각 부등식(triangle inequality)을 위반하는 경로를 제거함으로써 생성되는 네트워크이다. 이를 생성하는 알고리즘을 패스파인더 네트워크 알고리즘이라 하며, 때로는 다변량 분석 기법의 일종으로 간주하여 패스파인더 네트워크 스

케일링이라 부르기도 한다[19,20].

삼각 부등식 위반 여부를 결정하기 위해서는 두 가지 파라미터인 q와 r이 필요하다. 파라미터 q는 노드사이의 경로거리를 산출하는데 고려하는 최대 링크의 수를 뜻한다. n개의 노드를 고려할 경우, 파라미터 q는 2에서 n-1까지로 설정된다. 파라미터 r은 민스코프스키 거리 공식의 제곱수로서, 두 노드 n_i와 n_j사이의 특정 경로를 구성하는 여러 링크들이 가지고 있는 가중치를 거리 $W(n_i,n_j)$에 반영한 것이다.

$$W(n_i,n_j) \le (\sum_{l=1}^{k-1} W^r(n_l,n_{l+1}))^{1/r} \quad \forall k=2,3,...,q \quad (9)$$

식 (9)에서 r이 1이면 각 링크 가중치의 합이 그대로 경로의 거리가 되고, r이 무한대이면 경로를 구성하는 링크의 가중치 중 최대값이 경로의 거리가 된다. r이 커질수록 경로의 길이가 짧아지므로 남는 링크의 수는 줄어든다.

3. 사례 연구: 뉴스 비디오의 계층적 군집화 분석 및 의미 해석

본 논문에서 제안하는 계층적 군집화를 이용한 뉴스 비디오 마이닝의 타당성을 검증하기 위하여, 2007년 8월부터 2007년 10월 사이의 KBS와 MBC 뉴스 비디오 중, "2007년 제2차 남북 정상회담"을 주제로 갖는 뉴스 기사들에 대해 마이닝 분석을 수행하였다. 분석 대상이 되는 뉴스 비디오는 2007년 8월 6일부터 남북 정상 회담 직후인 2007년 10월 6일까지의 뉴스 중 "남북 정상회담" 관련 기사들이다. 실험에 사용된 데이터 집합은 KBS 85개 MBC 95개 총 180개의 뉴스 기사로 구성이 되어있으며, 해당 뉴스 기사의 비디오 클립과 스크립트를 모두 사용하였다. 데이터 집합에 대한 설명을 표 1에 정리하였다.

계층적 군집화를 통한 "2007년 제2차 남북 정상회담"에 관한 뉴스 기사들의 분류 및 분석을 위하여 평균 연결법을 이용한 계층적 군집화 방법을 사용하였다. 실험 데이터 집합에 대한 계층적 군집화 결과는 다음의 그림 2와 같다. 계층적 군집화를 통하여 생성된 덴드로그램(dendrogram) 분석 시, 각 군집의 이름은 해당 군집에 소속되어 있는 뉴스기사들의 제목을 고려하여 본 논문의 저자들이 결정하였다.

표 1 실험 데이터 집합

	설 명		
실험 데이터 주제	2007년 제2차 남북 정상회담		
방송 일자	2007년 8월 8일, 9일, 10일 2007년 10월 1일, 2일, 3일, 4일, 5일, 6일		
데이터 개수	KBS	MBC	합계
	85	95	180

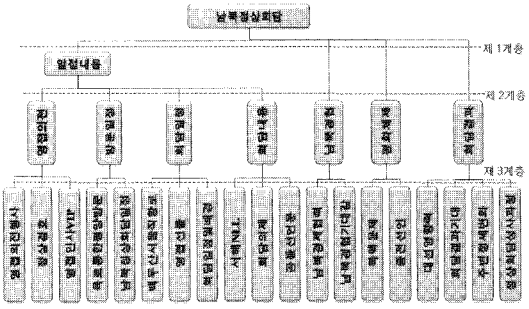

그림 2 평균연결 기반 군집분석을 통한 "2007년 제2차 남북 정상회담" 관련 기사들의 계층적 분류

계층적 군집화 결과, 그림 2와 같이 "2007년 제2차 남북 정상회담" 관련 뉴스 기사들에 대한 계층적 분류(taxonomy)를 구축할 수 있었다. 뉴스 기사들은 크게 '회담 일정 및 회담 내용', '남북경협', '평화체제', '회담 결과'의 4개의 군집으로 나뉘며, '회담 일정 및 회담 내용'은 다시 '영접 의전', '방문 일정', '회담 일정', '회담 내용'으로 세분화되었다. 최하위 계층인 제3계층은 총 19개의 세부 군집으로 구성되었다.

3.1 기간별 군집 성향 분석

본 절에서는 기간별 군집 성향 분석을 통하여, "2007년 제2차 남북 정상 회담"에 관한 TV 뉴스 기사의 변화를 알아보고자 한다. 방영된 뉴스의 개괄적인 동향 파악을 위하여, 기간을 남북 정상회담 이전, 남북 정상회담 기간, 남북 정상회담 이후로 나누었으며 군집화 결과의 제 1계층에서의 기간별 군집 성향 분석을 수행하였다. 그림 3에 의하면, 남북 정상회담이 시작되기 전에는 회담 결과에 대한 기대감이 매우 높았으나(C1), 회담이 진행되면서 실질적으로 해결해야 될 많은 문제점들이 도출됨으로 인해서 점점 기대감이 사라지고 있으며, 남북 간의 당면한 문제인 평화 체제 구축(C2)에 관련된 기사가 남북 정상회담의 시작을 기점으로 늘어나고 있음을 보여준다.

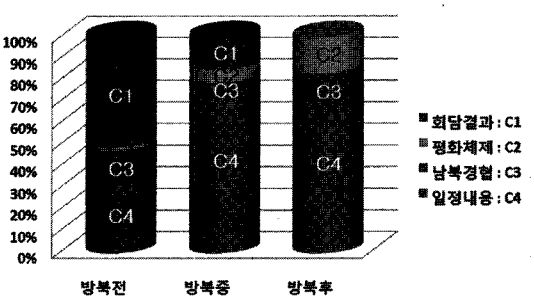

그림 3 제 1계층의 기간별 군집 추이 그래프

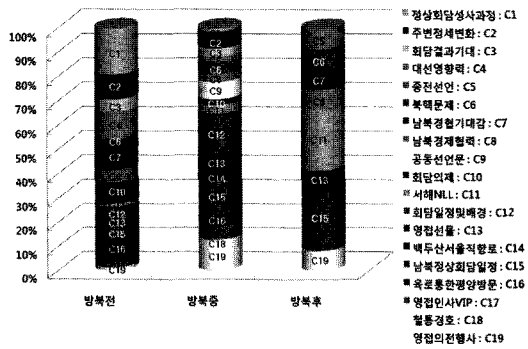

그림 4 제3계층의 기간별 군집 추이 그래프

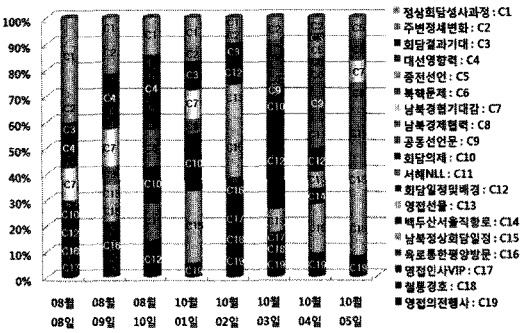

그림 5 제3계층의 일자별 군집 추이 그래프

남북 정상회담에 대한 보다 세밀한 뉴스 기사의 동향을 파악하기 위하여, 군집 결과의 제3계층에서의 기간별 군집 추이 분석을 수행하였다. 그림 4에 의하면 방북 전에는 남북 정상회담이 2007년도 대통령 선거에 미칠 영향에 대한 우려의 목소리를 담은 기사들(C4)이 많이 방영되었으나, 남북 정상회담 시작을 기점으로 뉴스 방영이 급격히 감소하였음을 알 수 있다. 또한, 한반도 주변국가의 정세 변환(C2)과 정상 회담의 결과에 대한 기대(C3)에 관한 뉴스 기사는 남북 정상회담이 끝남과 동시에 더 이상 방영되지 않고 있다. 반면, 종전 선언(C5), 북핵 문제(C6) 및 서해 NLL문제(C11) 관련 기사는 남북 정상회담이 시작되면서 보다 많은 뉴스 기사들이 발생하고 있음을 보여준다. 특히 서해 NLL문제와 관련된 기사들은 남북 정상회담 이후 폭발적으로 증가하였음을 보여준다. 이는 남북 간의 첨예한 문제로서, 방북 전에는 남북 정상회담에 부정적 영향을 줄 수 있는 기사들이 많이 방영되지 않은 반면, 방북 후에는 남북 간 협력을 위하여 해결해야 할 문제들에 대한 뉴스들이 더 많이 생산되고 있음을 보여준다.

남북 정상회담 관련 뉴스 기사를 보다 세밀한 분석을 위하여, 군집 결과를 제3계층에서 일자별 군집 추이 분석을 수행하였다. 그림 5에 의하면 남북 정상회담이 한반도 주변 국가 및 정세에 끼칠 영향에 대한 기사(C2)는 그 방송 분포는 상대적으로 많지 않으나 여러 날에 걸쳐 방송되었음을 보여준다. 이는 한반도 주변국가가 남북 정상회담에 미치는 영향력이 상당히 큼을 보여준다. 또한 남북 경제 협력과 관련 있는 기사(C7, C8) 역시 여러 날에 걸쳐 비중 있게 다루어지고 있음을 보여준다. 이는 제2차 남북 정상회담의 여러 의제 중 남북 경제 협력이 가장 중요하고 중심된 주제임을 반영한다. 반면 종전 선언(C5)은 남북 정상회담에서 서로의 합의를 실행하기 위한 조건으로 논의되어 10월 4일 이후부터 비중 있게 방영되었다.

3.2 가중 군집 성장 지수 분석

본 절에서는 가중 군집 성장 지수를 통하여 각 군집의 성장 정도를 분석하였다. 본 연구에서는 방북 이전을 제 1기로, 방북 이후를 제2기로 하여 가중 군집 성장 지수를 계산하였다.

남북 정상회담 관련 기사의 계층 군집화 결과(그림 2)를 기준으로 하여 가중 군집 성장 지수 분석을 수행하였으며, 제3계층에서의 가중 군집 성장 지수 그래프는 그림 6에 제시하였다. 분석결과, 남북 정상회담이 진행됨에 따라 '남북 경협 기대감'에 대한 뉴스 기사가 급격히 감소하고 있음을 보여준다. 이는 남북 정상회담 기간 동안 남북 협력 체제를 이루기 위한 실무적인 문제점들이 들어나고 실질적인 성과가 많이 도출되지 못했음을 시사한다. 반면 '서해 NLL', '북핵문제' 및 '종전선언'과 같이 남북이 실질적으로 당면한 문제를 부각시키는 뉴스 기사들은 늘어났음을 알 수 있다.

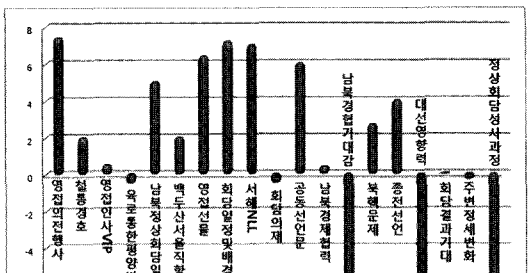

그림 6 제3계층에서의 가중 군집 성장 지수 그래프

3.3 패스파인더 네트워크 분석

남북 정상회담관련 뉴스 기사의 군집 분석 후 각 군집들 간의 상관관계를 알아보기 위하여 패스파인더 네

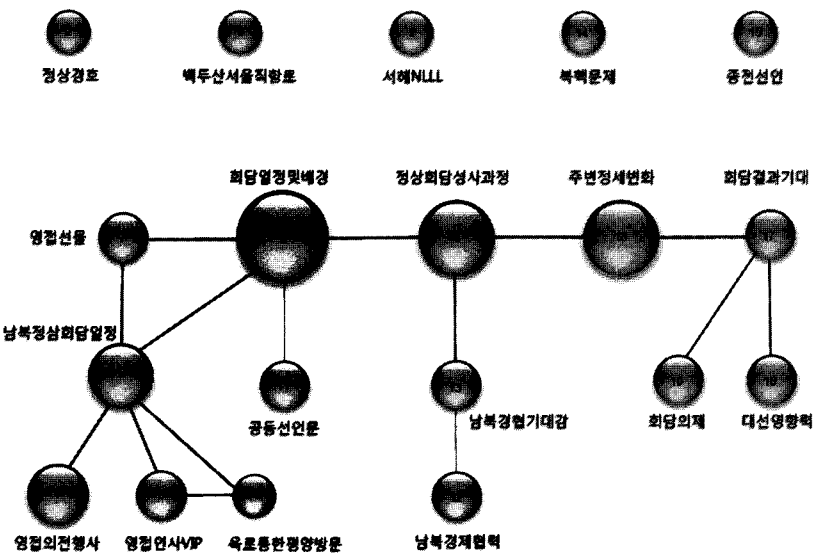

그림 7 "제2차 남북 정상회담" 관련 뉴스 기사의 패스파인더 네트워크 분석 결과

트워크 분석을 수행하였으며 그 결과를 그림 7에 제시하였다. 각 노드의 숫자는 그림 2의 제3계층에서의 군집 번호이며, 각 노드를 연결하는 링크의 굵기는 각 노드간의 관련 정도를 나타낸다. 본 분석을 통하여, 각 군집간의 상관관계를 확인할 수 있다. 먼저 '남북정상회담일정'은 '영접선물', '영접의전행사', '영접인사 VIP', '육로통한 평양방문' 그리고 '회담일정 및 배경'과 밀접한 관계를 가지고 있으며, '정상회담 성사과정'으로 관계를 확장하고 있다. '정상회담 성사과정'은 '남북경협 기대감'과 비교적 밀접한 관계를 가지고 있으며, '남북경제협력'으로 관계를 확장하고 있다. 마지막으로 '회담결과기대'는 '주변정세 변화'와 밀접한 관계를 보이며, '대선 영향력' 및 '회담의제'와도 비교적 밀접한 관계가 형성되어 있음을 알 수 있다. 또한, '서해 NLL', '북핵문제', '종전선언' 등은 그 자체로 중요한 이슈이며 다른 군집과는 별도로 진행된 주제임을 알 수 있다.

4. 결 론

본 논문에서는 뉴스 비디오 클립과 스크립트를 동시에 이용하는 멀티모달 방법론 및 스크립트에 대한 텍스트 마이닝을 적용하여 뉴스 비디오의 고수준 의미 분석을 수행하였다. 즉, 텍스트 마이닝 및 지적구조 분석에서 널리 사용되고 있는 계층적 군집화 기법을 뉴스 비디오 분석에 적용하여 뉴스 기사들을 자동 분류하였으며, 군집 분석 결과를 토대로 기간별 군집 추이 그래프, 군집 성장도 분석기법과 패스파인더 네트워크 분석기법을 통하여 뉴스 기사들의 다각적 분석을 수행하였다. 또한 본 논문에서 제안하는 방법론의 타당성을 보이기 위하여 "2007년 제2차 남북 정상회담" 관련 뉴스 기사들을 대상으로 뉴스 비디오 분석을 수행하였다.

뉴스 비디오 분석 결과 "2007년 제2차 남북 정상회담" 관련 뉴스 기사들에 대한 계층적 분류를 구축할 수 있었다. 뉴스 기사들은 크게 '회담 일정 및 회담 내용', '남북 경협', '평화체제', '회담결과'로 구분될 수 있었으며, 최하위 계층인 제3계층은 총 19개의 세부 군집으로 구성된 계층적 구조를 파악할 수 있었다. 또한, 계층별 기간-군집 추이 그래프, 가중 군집 성장 지수 분석 및 패스파인더 네트워크 분석을 통하여 기간별 남북 정상 회담 관련 기사의 동향과 각 군집의 성장패턴을 분석하였을 뿐만 아니라 각 군집들 간의 상호 연관관계를 분석하였다. 그 결과 "2007년 제2차 남북 정상회담" 관련 뉴스 기사들의 분류와 시간의 변화에 따른 뉴스 기사 내용의 변화 등과 같은 잠재적 지식을 도출할 수 있었다.

향후 연구과제로서는, 기존의 데이터 마이닝 및 텍스트 마이닝에서 사용되고 있는 다양한 방법론들을 뉴스 비디오를 포함하는 멀티미디어 마이닝으로 확장하는 방법에 대한 연구를 수행할 계획이며, 또한 멀티미디어 자체의 특성을 최대한 고려한 특화된 멀티미디어 마이닝 알고리즘의 개발에 대한 연구를 수행할 계획이다.

참 고 문 헌

[1] H. Lee, Y. Im, D. Park, S. Lee, "News Video Shot Boundary Detection using Singular Value Decomposition and Incremental Clustering," *Journal of KIISE: Software and Applications*, vol.36, no.2, pp.169-177, Feb. 2009. (in Korean)

[2] H. Lee, J. Yu, Y. Im, J. Gil, D. Park, "A Unified Scheme of Shot Boundary Detection and Anchor Shot Detection in News Video Story Parsing," *Multimed Tools Appl.*, 2010. DOI 10.1007/s11042-010-0462-x (online published)

[3] Z. Cernekova, I. Pitas, C. Nikou, "Information Theory-Based Shot Cut/Fade Detection and Video Summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol.16, no.1, pp.82-91, 2006.

[4] X. Luan, Y. Xie, L. Wu, J. Wen, S. Lao, "AnchorClu: An anchorperson shot detection method based on clustering," *Proc. of Int. Conf. on Parallel and Distributed Computing, Appl. and Technol.*, pp.840-844, 2005.

[5] X. Gao, J. Li, B. Yang, "A Graph-Theoretical Clustering based Anchorperson Shot Detection for News Video Indexing," *Proc. of Int. Conf. on Computational Intelligence and Multimedia Appl.*, pp.108-113, 2003.

[6] H. Lee, Y. Im, J. Park, D. Park, "A New Anchor Shot Detection System for News Video Indexing," *Journal of KIIS*, vol.18, no.1, pp.133-138, Feb. 2008.

[7] C. Ko, W. Xie, "News Video Segmentation and Categorization Techniques for Content-Demand Browsing," *Proc. of Cong. on Image and Sig. Proc.*, vol.2, pp.530-534, 2008.

[8] Y. Fang, X. Zhai, J. Fan, "News Video Story Segmentation," *Proc. of Int. Conf. on Multimedia Modeling*, pp.397-400, 2006.

[9] H. Lee, *A Data Cube System for The Semantic Analysis of News Video*, Ph.D. Dissertation, Korea University, Korea, December 2007.

[10] I. Ide, H. Mo, N. Katayama, S. Satoh, "Topic Threading for Structuring a Large-scale News Video Archive," *Proc. of Int. Conf. on Image and Video Retrieval, LNCS*, vol.3115, pp.123-131, 2004.

[11] X. Wu, C.-W. Ngo, Q. Li, "Threading and Auto-documenting News Videos: a promising solution to rapidly browse news topics," *IEEE Sig. Proc. Magazine*, vol.23, issue. 2, pp.59-68, 2006.

[12] J. Kim, H. Chang, Y. Kim, K. Kang, M. Kim, J. Kim, H. Kim, "Multimodal Approach for Summarizing and Indexing News Video," *ETRI Journal*, vol.24, no.1, pp.1-11, Feb. 2002.

[13] J. Shen, D. Tao, X. Li, "Modality Mixture Projections for Semantic Video Event Detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol.18, no.11, Nov. 2008.

[14] J. Lee, J. Moon, H. Kim, "Examining the Intellectual Structure of Records Management & Archival Science in Korea with Text Mining," *Journal of Korean Society for Library and Information Science*, vol.41, no.1, pp.345-369, 2007. (in Korean)

[15] L. Egghe, "Expansion of the field of informetrics: the second special issue," *Information Processing and Management*, vol.42, no.6, pp.1405-1407, 2006.

[16] P. Losiewicz, D. Oard, R. Kostoff, "Textual data mining to support science and technology management," *Journal of Intelligent Information Systems*, vol.15, no.2, pp.99-119, 2000.

[17] P. Glenisson, W. Glanzel, O. Persson, "Combining Full-text Analysis and Bibliometric indicators," *Scientometrics*, vol.63, no.1, pp.163-180, 2005.

[18] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2nd edn, pp.408-418, 2007.

[19] R. W. Schvaneveldt, *Pathfinder Associative Networks: Studies in Knowledge Organization*, Norwood, NJ: Ablex, 1990.

[20] C. Chen, "Generalized Similarity Analysis and Pathfinder Network Scaling," *Interacting with Computers*, vol.10, no.2, pp.107-128, 1998.

이 한 성

1996년 고려대학교 전산학과(학사). 1996년~1999년 (주)대우엔지니어링. 2002년 고려대학교 전산학과(석사). 2008년 고려대학교 전산학과(박사). 2006년 3월~2007년 2월 고려대학교 컴퓨터정보학과 초빙전임강사. 2008년 9월~2009년 2월 고려대학교 BK21 연구교수. 2009년 11월~현재 한국전자통신연구원. 관심분야는 휴먼인식, 멀티미디어마이닝, 기계학습, 지능 데이터베이스

임 영 희

1994년 고려대학교 전산학과(학사). 1996년 고려대학교 전산학과(석사). 2001년 고려대학교 전산학과(박사). 2001년~2003년 대전대학교 컴퓨터정보통신공학부 강의전담교수. 2003년~현재 고려대학교 컴퓨터정보학과 강사. 관심분야는 인공지능, 정보검색, 텍스트마이닝, 데이터마이닝

유 재 학

2001년 건국대학교 전산과학과(학사). 2003년 고려대학교 전산학과(석사). 2010년 고려대학교 전산학과(박사). 2006년 3월~2008년 2월 고려대학교 컴퓨터정보학과 초빙전임강사. 2008년 3월~현재 고려대학교 컴퓨터정보학과 강사. 관심분야는 데이터마이닝, 기계학습, 네트워크 마이닝, 침입탐지

오 승 근

2008년 고려대학교 컴퓨터정보학과(학사) 2010년 고려대학교 전산학과(석사). 2010년~현재 고려대학교 전산학과 박사과정 관심분야는 데이터마이닝, 패턴인식, 비디오 이벤트 인식

박 대 희

1982년 고려대학교 수학과(학사). 1984년 고려대학교 수학과(석사). 1989년 플로리다 주립대학 전산학과(석사). 1992년 플로리다 주립대학 전산학과(박사). 1993년~현재 고려대학교 컴퓨터정보학과 교수. 관심분야는 지능 데이터베이스, 데이터마이닝, 인공지능, 퍼지이론