감정요소를 사용한 정보검색에 관한 연구

김 명 관† · 박 영 택††

요 약

감정요소를 사용한 정보검색시스템은 감정에 기반 한 정보검색을 수행하기 위하여 감정시소러스를 구성하였으며 이를 사용한 감정요소추출기를 구현하였다. 감정요소추출기는 기본 5가지 감정 요소를 해당 문서에서 추출하여 문서를 벡터화시킨다. 벡터화시킨 문서들은 k-nearest neighbor, 단순 베이지안 및 상관계수기법을 사용한 2단계 투표방식을 통해 학습하고 분류하였다. 실험결과 분류 방식과 K-means를 이용한 클러스터링에서 감정요소에 기반 한 방식이 더 우수하다는 결과와 5,000 단어 미만의 문서 검색에 감정기반 검색이 유리하다는 것을 보였다.

A Study of using Emotional Features for Information Retrieval Systems

Myung-Gwan Kim† · Young-Tack Park††

ABSTRACT

In this paper, we propose a novel approach to employ emotional features to document retrieval systems. Five emotional features, such as HAPPY, SAD, ANGRY, FEAR, and DISGUST, have been used to represent Korean document. Users are allowed to use these features for retrieving their documents. Next, retrieved documents are learned by classification methods like cohesion factor, naive Bayesian, and, k-nearest neighbor approaches. In order to combine various approaches, voting method has been used. In addition, k-means clustering has been used for our experimentation. The performance of our approach proved to be better in accuracy than other methods, and be better in short texts rather than large documents.

키워드 : 에이전트(Agent), 감정정보(Emotional Information), 정보검색, 기계학습(Machine Learning), 인공지능(AI)

1. 서 론

국내에 인터넷 상용서비스가 시작된 이후로 현재 약 3000만 명이 넘는 사용자들이 있으며 매년 2배 가까이 그 수가 늘어나고 있다. 이제 인터넷은 연구원이나 대학원생의 전유물은 아니며 일반인들이 TV를 통해서 까지 사용하는 대중화된 매체가 되어가고 있다. 2002년 한국리서치 미디어인덱스의 2,350명을 상대로 얻은 조사결과[18]에 따르면 인터넷 사용이유에 따른 분포는 다음과 같다. 복수선택을 하였을 때 전자메일을 목적으로 사용하는 경우가 72.2%, 재미 64.4%, 파일다운하기 51.1%, 물건/서비스 정보가 48.0%로 나타나고 있다.

감정표현은 위 항목 중 두 번째 위치를 차지하는 재미와 밀접한 관계를 갖는 언어 도구이다. 현재 우리말에 대한 감정표현시소러스에 대한 연구는 매우 미약한 실정이다. 다음 21세기에는 인터넷을 이용한 수많은 오락 서비스가 이루어질 것이며 이를 위해 감정표현을 지원하는 시소러스(Thesaurus)의 개발과 우리말 감정표현검색기의 필요성이 대두될 것이다.

본 시스템은 로젯의 시소러스와 워드넷을 기반으로 구성된 시스템으로서 5가지 감정표현인 행복, 슬픔, 노여움, 공포, 혐오를 영어시소러스로 구축하고 대상 문서에서 이들 요소들을 추출하며, 사용자가 키워드와 함께 요구하는 감정요소 정도값에 의해 검색을 하는 시스템이다. 예를 들어 대다수의 사람들은 비디오 테이프 대여를 위해 감독이나 배우의 이름 또는 영화 내용 중의 키워드를 사용하지는 않는다. 대신에 "웃긴 영화 주세요" 또는 "최근에 나온 슬픈 영화 있나요" 등으로 요구할 것이다. 따라서 본 시스템은 다음과 같은 질의가 가능하다. "무서운 이야기가 있는 사이트는?", "슬픈 영화를 찾아 주세요", "2003년에 있었던 즐거운 사건들은?"

본 연구의 목표는 한국어 감정기반 검색시스템을 구축하는 것이다. 이를 위해 감정 단어들을 색인하고 검색하는 기술이 필요하다. 그 다음 한국어에 대한 감정표현 단어시소러스를 구축하는 것이다. 정보검색분야의 요구사항에 의해 언론연구원 등에서 한국어시소러스 구축에 대한 많은 시도

※ 본 연구는 숭실대학교의 연구지원 정책에 따라 지원을 받은 연구입니다.
† 정 회 원 : 서울보건대학 전산정보처리과 교수
†† 정 회 원 : 숭실대학교 컴퓨터학부 교수
논문접수 : 2003년 7월 23일, 심사완료 : 2003년 9월 22일

가 있었지만 전문분야 단어들에 대한 것이지 감정표현 단어들에 대한 사례는 없었다. 추출된 감정요소들은 대상 문서들의 벡터 값으로 저장되며 검색시스템은 키워드와 감정 벡터값의 요구로 이를 검색하게 된다.

벡터화된 문서값들은 상관계수, 단순 베이지안, k-Nearest Neighbor 기법에 의해서 분류되고 학습된다. 본 시스템에서는 3가지 기법을 이용하여 2단계 투표방식을 사용, 분류의 정확도와 안정도를 높였으며 감정요소 기반 분류가 기존에 단어들의 빈도만을 사용한 방식에 비하여 더 우수하다는 것을 보였다. 또한 실험 결과 문서의 크기에 따라 벡터 값의 표준편차가 다르게 나타났으며 웹과 같은 크기가 작은 문서일수록 감정 벡터에 의한 분류가 더 효과적임을 보였다.

다음 실험으로 K-means 방법을 사용하여 클러스터링을 해보았으며, 결과 감정요소를 기반으로 한 문서 클러스터링이 기존의 방식보다 더 좋은 결과를 나타내었다.

2장에서는 감정정보처리와 관련된 연구들에 대하여 다루었다. 3장에서는 감정에이전트를 사용한 인터넷 검색시스템의 구성과 개요에 대해서 설명하였다. 4장에서는 일반적인 키워드 자동색인기법과 이와 결합된 감정요소검색시스템에 대하여 다루었다. 5장에서는 한국어감정성분추출과 관련하여 인터넷에서 대상 문서들을 모아오는 에이전트 프로그램과 한국어 감정시소러스구축 및 감정요소추출 등을 다룬다. 6장에서는 분류를 위한 상관계수, 단순 베이지안, k-Nearest Neighbor 기법과 클러스터링 성능 비교를 위한 K-means 알고리즘에 대하여 다루었다. 7장에서는 인터넷의 13개의 국내 신문과 월간 잡지 사이트 및 DVD 판매 사이트의 영화 줄거리 요약 등에서 가져온 1,000개 문서에 대한 감정요소를 추출하여 실험을 통해 각 방법들에 의한 분류, 이들 문서를 통해 얻은 문서의 크기와 감정 벡터값 사이의 관계 및 K-means 방식을 사용한 클러스터링 성능 비교 등을 살펴보았다. 8장은 결론 및 향후 연구방향에 대하여 논하였다.

2. 관련 연구들

정보처리에 있어서 감정에 대한 연구는 인공지능 분야의 중요한 과제였다. 대표적인 감정처리를 포함한 시스템은 1981년 Colby의 PARRY[5], 1983년 Dyer의 BORIS[6]와 1987년 OdEd[7], 1991년 Reeves의 THUNDER[13], 1987년에 발표된 ACRES, 1992년 CMU의 Oz, 1997년 Wright[17]의 감정에이전트(Emotional Agents) 등이 있다.

PARRY는 '공포', '분노', '불신' 세 가지 감정 요소를 지원하는 시스템으로서 영어와 비슷한 입력형식을 가졌다. 주로 정신병 환자와의 대화를 목적으로 설계되었으며 악의 적인 질문이 발견되면 물리적인 공포와 심리적인 분노로부터 불신에 관한 변수 값을 변경하여서 대응하는 결과 값을 출력하는 구조였다. BORIS는 유명한 인공지능 사례인 '이혼'에 관한 에피소드를 인식하는 심층인식(In depth understanding) 소프트웨어이다. 문맥 이해를 목적으로 설계되었으며 감정은 이야기 분석에 단서로서 사용된다. 감정을 위하여 6개의 슬롯을 두어 변화를 처리하였다. 단순히 문맥 이해의 관점으로 감정요소를 다루었으므로 원칙, 목표, 성향 등의 구별을 갖지 않았다. 즉, 죄수가 탈옥 후 목표는 달성했지만 원칙과의 괴리 때문에 괴로워하는 감정의 변화를 표현할 수 없었다.

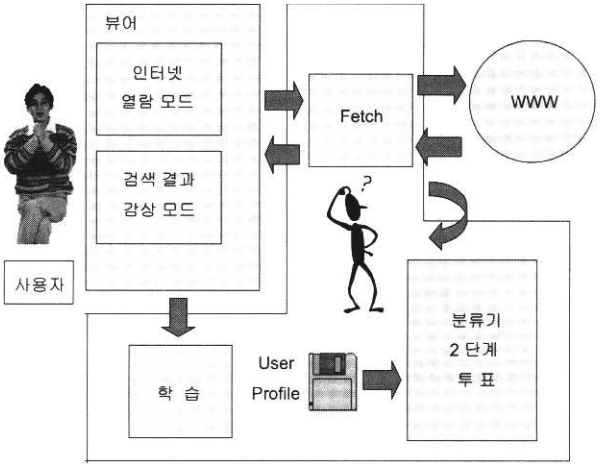
역시 Dyer의 OdEd는 신문의 사설을 이해하기 위한 시스템으로서 Shank의 CD(Conceptual Dependency) 이론을 적용하여 구현하였다. 지식베이스를 가지고 일종의 목표 기반 추론을 수행하였다. 즉, 실망을 만나면 지금의 목표는 취소하고 정지되어 있는 목표들 중에 다른 것을 선택하게 된다. 1991년에 나온 Reeves의 THUNDER는 역시 문서 구조를 이해하기 위한 시스템으로서 윤리 패턴을 가지고 이야기의 진행을 표현하고자 하였다.

Oz는 CMU에서 개발된 가상현실과 대화식 소설을 작성하기 위한 시스템이다. 이 시스템에서 생성된 가상 행위자들은 서로 경쟁적 행위를 하며 과정을 통해 환경설정을 변화시킨다. 각각이 목표 지향적인(Goal Directed Behavior) 행위를 하여 적절한 감정 변수를 갖는다. 이 시스템은 OCC (Ortony, Collins, Clore)들의 모델을 기반으로 하고 있다. 각 감정 요소들이 발생하는 기저를 다음과 같이 표현하고 있다. 즉 "JOY"는 목표가 성공했을 때 발생하는 감정요소이며 "HOPE"는 목표가 성공할 상황이 될 때 발생한다. 반면 "FEAR"는 목표가 실패할 가능성이 높을 때 발생한다.

그러나 Oz는 각 감정 요소의 정도를 표현할 수 없었으며 같은 타입의 감정들 사이의 차이를 보여주지 못한다. 위와 같이 살펴본 감정 요소를 사용한 기존의 시스템들은 주로 추론과 문장 이해를 목표로 구현된 것이다. 그러나 문장 이해는 대단위의 기반 지식을 갖추어야 얼마만큼의 결과를 보여줄 수 있다. 따라서 제한된 응용에서의 성과와는 달리 일반적인 범주에의 적용은 어려움을 갖는다. 본 논문에서 제안하는 시스템은 "전체 문장을 대표하는 감정값은 해당 감정 단어의 빈도에 비례한다"는 정의에서 출발한다. 정보검색에서의 자동 색인 원리와 일치한다[20]. 따라서 본 시스템은 각 감정 요소들의 정도를 부여할 수 있으며 5가지 (행복, 슬픔, 노여움, 공포, 혐오 : OCC 분류) 감정 요소들의 조합에 의해 해당 문서의 감정 대표 값을 산출하게 된다. 기존의 시스템들과 달리 본 논문에서 제안하는 감정기반분류 및 검색을 위한 감정 에이전트는 일반적인 인터넷 환경의 웹 문서 및 뉴스, 소설, 영화 등의 사이트분류 및 검색에 응용할 수가 있다.

3. 감정에이전트를 사용한 인터넷정보검색시스템

감정에이전트를 사용한 인터넷정보검색기의 구현이 본 연구의 목표이다. 이 시스템은 (그림 1)과 같이 구성된다. 각 구성 요소들의 기능은 다음과 같다.

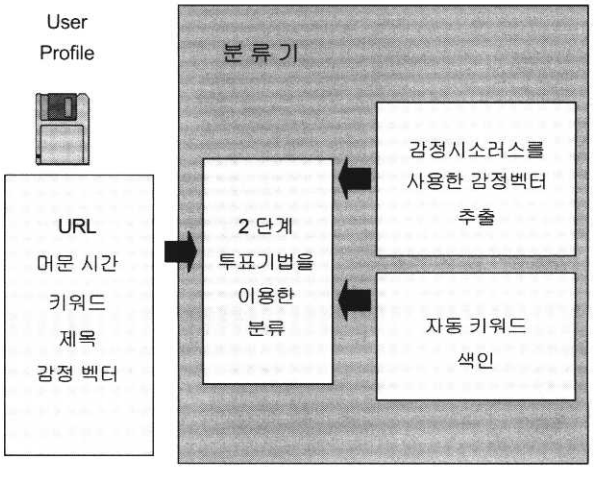


(그림 1) 인터넷정보검색을 위한 감정에이전트

- 뷰어모듈 : 뷰어는 사용자가 인터넷 검색을 하기 위해 보여지는 화면 인터페이스이다. 일반적인 웹 서핑을 위한 열람 모드와 에이전트가 사용자의 감정요구에 따라 웹에서 모아온 분류결과 문서들을 감상하는 감상 모드로 나누어진다.
- 학습모듈 : 에이전트가 모아온 분류의 문서들은 사용자의 평가에 의해 감정 벡터의 값이 조정된다. 이와 같은 사용자요구에 따른 되먹임(Feed Back)을 기존의 학습결과에 저장하는 모듈이다.
- 분류모듈 : 이 모듈은 웹으로부터 가져온 문서들의 감정벡터들을 이미 학습된 분류에 따라 2단계 투표기법으로 분류하는 모듈이다.
- Fetch 모듈 : 정기적으로 웹에서 사용자의 프로파일(방문 URL, 머문 시간, 키워드, 페이지의 제목, 당시의 감정값)에 맞는 문서들을 모아온다.

이를 구현하기 위해서 기존의 자동색인 및 검색시스템과의 통합적인 구성이 요구되어 진다. 즉 외부에서 들어오는 Web 및 각종 문서들은 감정요소 추출기에서 5가지 감정요소가 추출되어서 해당 문서의 감정벡터를 구성한다. 감정요소추출과정은 해당 문서의 형태소분석을 통하여 감정시소러스에 등록된 단어들의 감정요소들의 수를 누적하여 얻어진다. 이렇게 얻어진 감정벡터는 계산의 편리를 위하여 정규화 되며 상관계수, 단순 베이지안, k-Nearest Neighbor 기법들을 이용하여 분류된다[1]. 또한 해당 문서는 형태소분석을 통해 얻은 명사정보에 대해 다빈도, 저빈도 단어제거, 기능어제거 등의 과정 후에 키워드를 얻어낸다. 역화일 구성을 통해서 자동색인을 구현하였다. 이렇게 얻어진 감정요소 벡터화일과 키워드자동색인파일은 1차와 2차 검색 과정에 사용된다. 사용자는 본인이 찾고자 하는 문서의 감정요소값을 제시하고 검색기는 이 사용자 요구 감정요소값과 가장 유사한 문서를 상관계수, 단순 베이지안, k-Nearest Neighbor 기법을 이용하여 값을 구하고 세 가지 방법을 투표하여 2가지 방식 이상의 분류에 해당하는 문서를 유사문서로 분류한다. 이렇게 얻어진 유사감정 문서들 중에서 사용자가 입력한 키워드의 중요치(weight)에 따른 2차 검색을 실시하게 된다.

이때 사용자 인터페이스에서 감정요소는 그래픽입력기로, 키워드는 문자상자로 입력하게 된다. 감정에이전트의 핵심이 되는 웹 문서에서의 감정벡터추출과 이를 각 기본 감정요소(행복, 슬픔, 노여움, 공포, 혐오)로 분류하는 분류기는 그림 2와 같다. 분류기의 구성 내용은 감정시소러스를 사용하여 웹 문서에서 벡터를 추출하는 소 모듈과 일반적인 키워드색인 소 모듈, 상관계수, 단순 베이지안, k-Nearest Neighbor 기법을 사용하여 이들 중 다수결의 결과로 분류하는 2단계투표기법분류기 등이 포함된다. 에이전트는 분류된 문서 중에서 사용자의 프로파일을 참조하여 가장 적합한 문서를 찾고 이를 값이 적합성이 높은 순서로 배열하고 사용자에게 감상시킨다.



(그림 2) 문서 감정벡터 분류기(Classifier)

분류기의 구성 내용인 감정시소러스를 4장에서 다루며 한국어 감정요소추출기는 5장에서 2단계 투표분류기법은 6장에서 설명한다.

4. 한국어 감정시소러스 구축

본 연구의 첫 목표는 감정표현 단어들에 대한 시소러스

를 구축하는 것이다. 정보검색분야의 요구사항에 의해 언론연구원 등에서 한국어시소러스 구축에 대한 많은 시도가 있었지만 전문분야 단어들에 대한 것이지 감정표현단어들에 대한 시소러스 구축은 없었다. 감정단어들에 대한 연구의 대표적인 결과물은 영국에서 나온 로젯(Roget)의 시소러스[14]와 미국 프린스틴 대학의 의드넷[11]을 들 수 있다. 로젯의 시소러스는 클레스 6에 감정(AFFECTION)을 일반적인 감정 단어, 개인적인, 동정, 도덕적인 것, 종교적인 감정 단어들로 분류하고 각각 단어 군들의 유사어들을 280개의 군으로 나누어 정리해 놓고 있다. 1976년 밀러와 존슨에 의해 만들어진 워드넷(WordNet)은 심리학과 언어공학이 만들어낸 결과물이다. 1993년 새로 구성된 결과물에 의하면 95,600개의 단어구성으로 이루어져 있으며 51,500개의 단순 단어와 44,100개의 연어로 이루어져 있다. 70,100개의 단어의미가 있으며 유사어가 있다. 이 워드넷은 5개의 범주로 구성되어있는데 명사, 동사, 부사, 형용사, 기능어로 되어있다.

예를 들어 "04787051 12 n 01 sentimentality 0 002 @ 04786928 n 0000 ~ 04787172 n 0000 | … sentiment mawkishness"는 워드넷의 각 단어가 상위어, 하위어, 유사어 등에 대한 연결을 네트웍으로 구성하고 있음을 나타낸다. 즉 04787051은 단어 sentimentality의 고유번호이며 @은 상위어가 04786928이며 …은 하위어가 04787172라는 의미이다. 또한 맨 뒤에는 유사어들이 작성되어있다.

이미 발표된 ECRAS에서는 감정 단어들을 행복, 슬픔, 노여움, 공포, 혐오(Happy, Sad, Angry, Fear, Disgust) 등 5가지 범주로 나누고 각 범주의 단어들을 워드넷과 유사한 망 형태로 구성하였다[21]. 또한 각 감정성분들은 로젯시소러스의 유사어 분류를 사용하였다. 구성한 결과 행복은 1,700단어, 슬픔은 1,300 단어, 노여움은 1,400, 공포는 1,300 혐오는 1,500개로 구축했으며. 이와 같은 기본 단어군과 워드넷의 관련 단어들을 연결하여 감정 성분 단어들의 시소러스를 구축하였다. 특히 위의 워드넷을 이용하여 로젯의 시소러스에서 가져온 분류 단어들에 유사어와 상위어 하위어 등을 본 시스템의 시소러스로 구성하였다. 감정 범주 분류기는 그림 2와 같은 모습으로 구현된다.

위의 내용 중 단어 옆에 있는 숫자는 상위어 및 하위어를 표시한다. 따라서 시소러스에서의 위치에 따라 누적값이 차등으로 적용된다. 기본적으로는 로젯의 시소러스를 사용하였으며 워드넷의 상위어 하위어 부분을 일부 포함 시켰다. 또한 감정성분 추출 방법은 다음과 같다.

① 입력된 문서에서 하나의 단어를 추출한다.
(Poter의 알고리즘을 사용하여 복수형, 진행형 등을 표준형으로 바꾼다.)
② 불용어인 경우 이를 제거하고(불용어 사전으로) 아니면 시소러스를 검색한다.
③ 시소러스를 사용하여 해당 감정성분 값을 누적한다.
④ 값을 계산할 때 시소러스의 상위어 하위어 유사어 등에 값을 차등화 하여 누적한다.
⑤ 만들어진 5개의 벡터 값을 정규화 한다.

한국어 감정 시소러스는 ECRAS의 영어감정시소러스를 기반으로 구축하였다. 이를 위해서 영한 사전 20,000 단어를 ACCESS 데이터베이스로 구성하였다. 사전은 영어, 한국어, 품사 필드로 이루어져 있으며 영어 감정 단어를 영한 사전에서 검색하여 해당하는 한국어 및 품사로 구성하였다.

결과로서 한국어 행복 1,054 단어, 슬픔 745 단어, 노여움 571 단어, 공포 575 단어, 혐오 720 단어를 구축할 수 있었다.

5. 한국어 감정 성분 추출

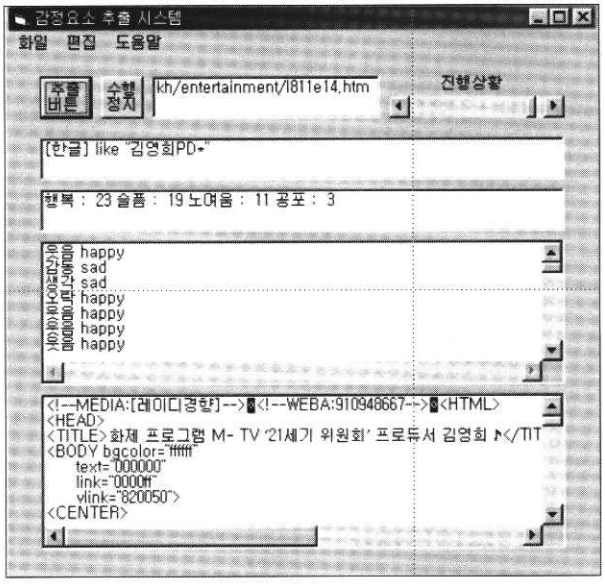
한국어 감정요소를 추출하기 위해서는 대상 문서로부터 각종 형태소사전들을 사용하여 주요 단어들을 추출하게 된다. 사용하는 사전들은 다음과 같다. 첫 번째 명사사전은 약 90,000개의 단순명사와 지명, 인명, 시사 용어들로 구성된다. 주로 신문의 사설과 기사, 백과사전, 소설, 사회과학 서적 등에서 인용되었다 두 번째 조사사전은 약 500개로 구성되어 있다. 각종 어미변화에 따른 변화형과 격조사들로 구성되며 제일 긴 조사 어미를 대상 어절에 먼저 시도하고 1음절 조사, 어미는 애매성을 제거하기 위하여 명사, 복합명사 처리를 시도한 후에 검사하게 된다. 세 번째로 불용어 사전이다. 약 10,000개의 단어들로 이루어져 있는 중심어가 될 수 없는 어절들의 모임이다. 각종 소설 및 사회서적을 원본으로 색인어가 될 만한 어절들을 제거하면서 얻어졌다. 여기서 각 문서로부터 감정 시소러스를 사용하여 5개 성분의 벡터(행복, 슬픔, 노여움, 공포, 혐오)를 추출하며 문서 j의 추출된 성분 벡터는 다음과 같이 표현한다.

$$EVj = \langle V1j, V2j, V3j, V4j, V5j \rangle$$
$$Vij = \log(Fij) / \log(Fwj)$$
or
$$Vij = Fij / Fwj$$

위에서 Vij는 문서 j에서의 i번째 감정 성분을 나타내며 Fij는 i번째 감정 단어가 j문서에서 나타난 빈도이며 Fwj는 j문서에 있는 모든 단어의 개수가 된다. Vij는 0에서 1 사이의 값을 갖는다. 키워드 자동 색인은 기존 Salton[15] 등의 일반적인 방법을 사용하였다.

(그림 3)은 URL을 입력으로 해서 가져온 HTML 문서를 키워드 추출하여 이의 감정값을 시소러스로부터 얻어 누적

하는 수행을 보여주고 있다. 위의 예에서는 21세기 위원회라는 TV 프로의 프로듀서를 인터뷰한 내용이며 전체적으로는 즐거운 이야기가 기본을 이루고 있음을 알 수 있다.



(그림 3) 감정 요소 추출기의 화면

6. 실험에 사용하는 방법론들

본 연구에서는 감정기반 정보검색에 대하여 분류(Classification), 클러스터링 관점에서 기존의 단어의 빈도에 기반한 정보검색방법론과의 성능에 대한 비교를 실시하였다. 이를 위해 사용한 각각의 방법론은 다음과 같다.

6.1 분류 기법

본 실험에서 사용한 분류기법(classification method)은 상관계수법, 베이지안, k-Nearest Neighbor 등 3가지이다. 상관계수는 통계학에서 사용하는 두 표본 사이의 상관도를 구하는 방법이다. 상관계수(Correlation)는 다음과 같이 구한다. 여기에서 Data(j, i)는 실험 문서의 감정 시소러스에 있는 단어들의 분포를 docu_freq(i)는 학습되어 있는 실험 문서들의 감정단어 분포를 나타낸다.

```
For i = 0 To max_arry - 1
  sum_xy = sum_xy + docu_freq(i) * posi(i)
  sum_x = sum_x + docu_freq(i)
  sum_y = sum_y + posi(i)
  exp_x = exp_x + docu_freq(i) ^ 2
  exp_y = exp_y + posi(i) ^ 2
Next i
  Crr = ((max_arry × sum_xy) - (sum_x * sum_y))
        / (Sqr(max_arry*exp_x-sum_x ^ 2)
        × Sqr(max_arry × exp_y - sum_y ^ 2))
```

베이지안은 대표적인 사전확률과 사후확률을 이용한 분류기법으로서 본 연구에서는 단순 베이지안 분류기법(naive bayesian classification method)을 사용하였다. 이 방식은 실험문서에 있는 감정단어 각각에다가 학습된 해당 단어의 확률들을 계속 곱해나가는 것이다.

3번째 방법은 k-nearest neighbor 방식으로서 instant 기반 및 후 수행(Lazy) 학습의 대표적인 알고리즘이다.[12] 이 방법은 학습할 대상 개체들을 n 차원 공간 속의 한 점으로 인식하고 벡터화 된 좌표를 학습 시 단순하게 기억장치에 저장만 한다. 클래스를 모르는 질의어가 올 때 이를 n 차원상의 벡터로 표현하고 점 사이의 거리 구하는 방식으로 가장 가까운 k 개의 점들을 구하여 이들 점들이 가지고 있는 클래스 정보로 질의어의 클래스를 결정하는 기법이다. 이 경우 질의어 개체가 m 개의 성분을 가진다면 다음과 같이 표현된다.

$$<a1(x), a2(x), a3(x), \cdots, am(x)>$$

두 점들 사이의 거리는 다음과 같다.

$$d(xi, xj) = \sqrt{\sum_{r=1}^{m} (a_r(xi) - a_r(xj))^2}$$

위와 같은 거리 구하는 방식으로 질의어 x_q에 대한 분류값을 학습된 개체 중 가까운 거리의 x1, x2, ···, x_k 를 대상으로 다음과 같이 구한다.

$$f(x_q) = \underset{v \in V}{\operatorname{argmax}} \sum_{i=1}^{k} \delta(v, f(x_i))$$

이 k-NN 기법은 단순하면서도 동작 과정을 쉽게 이해할 수 있고, 학습시간이 빠르며(ID3의 100배 정도) 높은 정확성(accuracy)으로 널리 사용되고 있다. 다음과 같이 구현하였다. scale1과 scale2는 학습된 데이터와 실험 데이터의 규모를 맞추어주는 변수이다.

```
For i = 1 To totalData
k_near = 0
  For j = 1 To Field_Count
    k_near = k_near + (docu_freq(j) - Data(j, i)) ^ 2
  Next j
  k_near = Sqr(k_near)
Next i
```

2단계 분류기법은 위와 같은 각각의 분류결과들을 다수결로 최종분류 결과를 정하는 방법이다[23].

6.2 K-means 클러스터링

클러스터링의 K-means 방법은 실제로 가장 보편적으로 이용되는 방법이다. 사전에 결정된 군집수 k에 기초하여 전체 데이터를 상대적으로 유사한 k개의 군집으로 구분하는 방법이다. 설명을 간단히 하기 위해 2차원의 다이어그램을 이용하여 설명한다(실제로는 많은 차원의 환경에서 이루어진다).

- 첫번째 단계에서 군집의 수 k를 정한다.
- 두번째 단계에서 초기 K개 군집의 중심을 선택한다.
- 세번째 단계에서 각 관찰치를 그 중심과 가장 가까운 거리에 있는 군집에 할당한다.
- 네번째 각 군집별로 그에 속하는 관찰치를 이용해 새로운 중심을 계산한다.

위의 과정을 기존의 중심과 새로운 중심의 차이가 없을 때까지 반복한다.

K-means method는 상업용 software tool과 함께 계산방식에 많은 변동을 가졌다. 즉, 초기의 시드들을 선택하는데 있어서 또 다음 중심 값을 계산하는데 있어서 다른 방법이 적용되거나 군집과 관련한 레코드들의 거리를 계산하기보다는 확률밀도를 이용하기도 한다.

K-means 알고리즘의 단계를 살펴보면, 데이터베이스의 레코드들은 일단 공간에서 점으로 맵핑이 되어야 한다. 하지만 우리가 마케팅, 영업 등에서 접하는 데이터 베이스들은 그렇지 않다. 그래서 우리는 레코드를 공간의 점으로 취급하기 위해서 모든 필드를 수치형 변수로 바꾼다. 만약 두 점들이 가까이 근접해 있다면, 우리는 그 점들을 데이터베이스에서 같은 레코드라고 표현할 수 있다. 그러나 이러한 접근은 여러 형태의 변수 타입들이 벡터의 요소로 표현되기 어렵다라는 문제점을 가지고 있다[22].

7. 실험 및 고찰

본 논문에서 제안한 시스템의 실제응용을 위해 기존 DVD 판매사이트에서 영화소개를 하는 줄거리요약(Synopsis) 물들을 사용하여 실험환경을 구축하였다. 각 각의 내용은 사이트에서 희극과 공포로 이미 분류되어 있는 것으로서 여러 사이트에서 약 1,000여 개를 구할 수 있었다.

7.1 분류(Classification)

첫 번째 실험은 이들을 각각 희극 500여개, 공포물 500여개를 가지고 실시하였다. 실험결과는 <표 1>에 나타내고 있으며 첫 번째 열은 학습한 감정요소기반분류의 인식률이며 두 번째 열은 일반적인 단어의 빈도에 기반 한 학습으로 분류한 결과이며 세 번째는 두 경우 인식률의 평균을 제시하였다. 결과로 Nearest Neighbor 기법이 가장 우수한 것으로 나타났으며 감정요소기반 분류가 더욱 우수 한 것으로 나타났다.

7.2 문서의 크기와의 관계

두 번째 실험은 각종 사이트 및 CD-ROM에서 수집한 다양한 실험문서들의 감정요소벡터를 구한 후에 감정요소벡터들의 문서 크기에 따른 표준편차를 구해 보았다.

〈표 2〉 영문과 한글 실험자료들로부터 추출된 감정요소들의 표준편차 평균

	10,000 단어 이상 문서의 감정요소표준편차평균	5,000 단어 미만 문서의 감정요소표준편차평균
ECRAS	6.20	10.12
본 시스템	5.47	9.32
문서 종류	영화대본, 소설원고, 세익스피어의 희곡	홈페이지, 신문기사, 방송대본, 사설, Web 게시판, 잡지

표준편차가 클수록 사용자의 감정요소에 따른 검색에 애매성이 적어질 것이다. 영어 시스템인 ECRAS에서의 실험결과는 5,000 단어 미만의 짧고 간략한 신문기사(1,000개) 등은 변별력 즉 표준편차의 크기가 크게 나타났으며(표준편차 평균 : 10.12) 상대적으로 큰 10,000 단어 이상의 영화대본(100개)이나 세익스피어의 희곡 등(18개)은 감정 요소들 간의 차이가 크지 않게 나타났었다(표준편차 평균 : 6.20).

한국어를 사용한 실험 결과는 5,000 단어 미만인 경우 표준편차평균이 9.32이며 10,000 단어 이상의 문서에서는 5.47 정도로 나타났다. 이는 영문과 마찬가지로 문서가 커질수록 각 감정요소들의 편차가 줄어드는 현상을 말해주고 있다. 따라서 감정요소에 의한 문서 분류 및 검색은 한국어에 있어서도 인터넷의 웹 등 온라인 문서의 주류를 이루는 게시

〈표 1〉 각 기법들의 정확도

	감정요소 기반 (학습된 데이터)	감정요소 기반 (학습 안된 데이터)	일반 단어 빈도 기반 (학습된 데이터)	일반 단어 빈도 기반 (학습 안된 데이터)	평 균
상관 계수법	45%	40%	30%	54%	42.3%
Naive Bayesian	82%	78%	72%	66%	74.5%
Nearest Neighbor	85%	85%	80%	65%	78.8%
평 균	70.7%	67.7%	60.7%	61.7%	

판 및 신문 기사나 잡지의 기고 내용 등 가볍게 읽을 수 있는 문서들이 적합한 것으로 보인다.

7.3 클러스터링

세 번째 실험은 K-means 방법을 사용한 클러스터링이다. 즉 희극과 공포 물이므로 클러스터링 크기 K = 2로 놓고 감정요소기반 자료와 일반적인 단어 빈도를 기반으로 한 자료를 가지고 클러스터링을 수행하였을 때 목표한 희극, 공포의 클러스터링 결과가 감정기반 접근방식이 더 우수함을 보였다.

〈표 3〉 K-means 기법을 사용한 클러스터링

클러스터링 기법	감정요소 기반	일반 단어 빈도 기반
K-means	75%	63%

8. 결론 및 향후 연구 방향

본 논문에서는 한국어 감정성분을 해당 문서로부터 추출하여 감정성분을 기반으로 한 감정에이전트를 사용한 인터넷 정보검색 시스템을 제안하였다. 이미 구축된 영문시스템인 ECRAS를 기반으로[21] 한국어감정 시소러스를 구성하였다.

본 연구의 기반은 정보검색 이론과 마찬가지로 "전체문장을 대표하는 감정값은 해당 감정단어의 빈도에 비례한다"는 정의에서 출발한다. 행복, 슬픔, 노여움, 공포, 혐오 등의 기본 감정요소들의 조합에 의해 해당 문서의 감정 대표값을 산출하게 된다. 기존의 감정정보처리시스템들과 달리 본 논문에서 제안하는 감정에이전트는 일반적인 인터넷 환경의 웹 문서 및 뉴스, 소설, 영화 등의 사이트분류 및 검색에 응용할 수가 있다. 5가지 감정 성분의 추출로 나온 결과는 정도 값으로 대상 문서의 대표감정값을 알 수 있다. 사용자는 이를 키워드와 자신이 찾고자 하는 감정요소 값을 질의어로 정보검색을 할 수 있다.

또한 감정기반분류를 위하여 강력한 분류기법인 단순 베이지안, k-Nearest Neighbor 기법 등의 결과를 투표를 통해 결정하는 2 단계 투표기법을 제안하였으며 실험을 통해 기존 단어빈도기반분류방식보다 안정성과 정확도가 우수함을 보였다. 또한 영문과 같이 한국어 문서에서도 각 감정요소들의 표준편차가 대상 문서의 크기에 반비례한다는 것을 보였다. 세 번째 실험을 통해 K-means를 사용한 클러스터링에도 우수한 효과가 있음을 보였다. 비주얼 베이직으로 시스템을 구현하였으며 ActiveX Control 및 ACCESS 데이터베이스를 사용하였다. 인터넷상의 각종 신문 기사들과 월간 여성 잡지 기사들 및 CD로 출판된 동아일보 사설, 각종 영화대본 및 DVD 판매사이트의 영화 줄거리요약 등을 사용하였다. 앞으로 한국어감정시소러스를 더욱 보강하는 일과 편리한 사용자 인터페이스의 개발 및 응용분야를 개발하는 노력이 필요하다.

참 고 문 헌

[1] Aha, D. W., "Instance-Based Learning Methods," Machine Learning, 6, pp.37-66, 1991.

[2] Alpaydim, E., "Voting Over Condensed Nearest Neighbors," Bogazici Univ, 1995.

[3] Alpaydim, E., "GAL : Networks that Grow When They Learn and Shrink When They Forget," International Computer Science Institute, Berkeley : CA, TR-91-032, 1996.

[4] Clark, M. S., "Affect and Cognition," LEA Publishers, 1982.

[5] Colby, M., "Modeling a paranoid mind," The Behavioral and Brain Sciences, 4(4), pp.515-560, 1981.

[6] Dyer, M. G., "In depth understanding," MIT Press, 1983.

[7] Dyer, M. G., "Emotions and their computations," Cognition and Emotion, 1(3), pp.323-347, 1987.

[8] Elliot, C. D., "A Process model of emotions in a multi-agent system," Ph.D thesis, north-west Univ, 1992.

[9] Fisher, D., "Knowledge Acquisition via Incremental Conceptual Clustering," Machine Learning, 2, pp.139-172, 1987.

[10] Hart, P. E., "The Condensed Nearest Neighbor Rule," IEEE Transaction on Information Theory, 14, pp.515-516, 1968.

[11] Miller, G. A., "WordNet : An On-line Lexical Data Base," Hillsdale, 1993.

[12] Mitchell, T., "Machine Learning," McGraw-Hill, 1997.

[13] Reeves, J. F., "Computational morality : A process model of belief conflict and resolution for story understanding," Technical Report UCLA-AI-91-05, UCLA AI Lab, 1991.

[14] Roget, P. M., "Roget's Thesaurus," Gramercy Books, 1979.

[15] Salton, G., "Automatic Text Processing," Addison Wesley, 1989.

[16] Sestito, S., "Automated Knowledge Acquisition," Prentice Hall, 1994.

[17] Wright, I. P., "Emotional Agents," Ph. D. thesis, Univ. of Birmingham, 1997.

[18] 한국리서치 미디어인덱스, "2002년 1R 조사결과", http://ad.joins.com/trend/internet-6.asp, 2002.

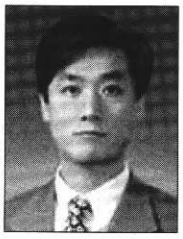
[19] 유상진 외 3인, "현대 통계학", 법영사, 1997.

[20] 정영미, "정보 검색론", 정음사, pp.181-208, 1986.

[21] 김명관, 박영택, "감정기반 정보 검색시스템에 관한 연구", 한국문헌정보학회, 제32권 제4호, 1998.

[22] "Automatic Cluster Detection," http://home.pusan.ac.kr/~pnustat/info/DataMinig/2-3.htm, 2003.

[23] 김명관, "2단계 분류기법을 이용한 영상 분류기 개발", 컴퓨터산업교육학회논문지, Vol.3, No.5, May, 2002.



김 명 관

e-mail : binsum@shjc.ac.kr

1985년 숭실대학교 전자계산학과(학사)

1987년 숭실대학교 대학원 전자계산학과 (석사)

1989년~1993년 한국전자통신연구원 인공지능연구실 연구원

1993년~현재 서울보건대학 전산정보처리과 교수

관심분야 : 정보검색, 자연어처리, 에이전트 등

박 영 택

e-mail : park@computing.soongsil.ac.kr

1978년 서울대학교 전자공학과(학사)

1980년 KAIST 전자계산학과(석사)

1992년 Illinois at Urbana-Champaign 컴퓨터과학과 박사

1981년~현재 숭실대학교 컴퓨터학부 교수

관심분야 : 지능에이전트, 웹에이전트, 모바일에이전트, 기계학습 등