Cognitive Mapping을 이용한 Information Filtering System

김진화*, 이승훈**, 변현수**

* 서강대학교 경영학과
서울 마포구 신수동 1번지, 121-742
Tel: +82-2-705-8860, Fax: +82-2-705-8519, E-mail: jinhwakim@sogang.ac.kr
** * 서강대학교 경영학과
서울 마포구 신수동 1번지, 121-742
Tel: +82-2-705-8860, Fax: +82-2-705-8519, E-mail: au4240, elbim@sogang.ac.kr

Abstract

대량의 정보를 취급하는 현대사회에서는 획득하는 정보를 모두 처리할 수 없어 이용자의 요구에 맞는 정보를 얻기 위해 정보 필터링 시스템을 사용한다. 그러나 정보 필터링 시스템이 이용자의 의도와 다르게 정보를 분류하거나 이용자의 다양한 요구를 반영하지 못할 때는 정보 필터링을 사용하지 않은 경우보다 못할 수 있다. 본 연구에서는 정보필터링의 정확도를 향상시키기 위해 인간의 뇌에서의 정보처리 과정을 시뮬레이션하는 인지적 매핑의 정보 필터링 시스템을 제안하고자 한다. 본 연구에서 제안하는 시스템은 특정 단어 또는 패턴만을 이용하여 필터링하는 기존 시스템과는 달리 단어의 존재, 단어와 단어의 연결이 창출하는 의미와 단어의 가중치를 종합하여 정보를 필터링하는 점에서 의의가 있다.

Keyword

Information Filtering, Filtering Agent, Contextual Filtering, Brain Mapping, Cognitive Mapping

1. 서론

현대 사회는 정보화 사회로 사회 전체가 정보가치의 창출에 주력하고 있다. 사회 중심기반을 정보화 사회로 변화시킨 정보통신기술의 발전을 통해 정보를 효과적으로 창출, 처리, 관리, 통제, 저장하고 이를 빠르고 용이하게 전달할 뿐 아니라 확대된 지적 능력들간의 소통과 연계를 통해 연결된 정보의 양과 가치는 무한하게 증폭되고 있다. 과거와 달리 현대 사회는 개인이 처리해야 하는 정보의 양도 기하급수적으로 증가하고 있다. 이러한 환경하에서 중요한 것은 필요한 정보를 신속하게 획득하는 것과 필요하지 않는 정보를 취하지 않는 것이다. 필요한 정보의 선별을 위해 정보기술을 이용하여 꼭 필요한 정보를 걸러내는 작업의 중요성이 증가하고 있다. 이와 같이 필요한 정보를 얻기 위해 정보를 선별, 배포하는 다양한 과정인 정보 필터링은 방대한 양의 정보가 중단 없이 생산되고 정보의 증가율이 높은 분야인 인터넷, 특히 유즈넷(Usenet) 분야에서 활발히 적용되고 있다[8]. 정보 필터링의 적용영역은 유즈넷 뉴스, 전자우편뿐만 아니라 정보의 증가율과 유입량이 많은 분야에 적용할 수 있다.

본 연구에서는 향후 급속한 정보의 증가가 예상되는 전자저널, 전자문헌, 영상이나 동영상 등의 멀티미디어 분야에 적용할 수 있으며 이용자의 요구에 맞는 정보를 획득할 수 있도록 하는 새로운 정보 필터링 시스템을 제안하고자 한다. 기존의 정보 필터링은 이용자의 정보요구(User Profile)에 따라 유동적인 정보원으로부터 적합한 정보를 선별하여 제공하는 것에 초점을 두었으나 본 연구에서는 인간이 정보를 획득하고 획득한 정보를 뇌에서 처리하는 과정을 필

터링에 적용한 정보 필터링 시스템을 제안하고자 한다. 인간은 정보처리과정에 있어 시각 등의 감각기관을 통해 획득한 정보를 직렬적으로 처리하는 것이 아니라 개인이 학습하고 경험한 것에 기초하여 획득한 정보를 취사선택을 하므로 일괄적인 방법으로 필요/불필요의 기준을 수립하여 정보를 필터링하는 경우 필요한 정보를 차단하게 되는 경우가 발생할 수 있다.

2. 이론적 고찰

2.1 정보 필터링과 필터링 에이전트

네트워크 기술의 발달로 정보를 쉽게 공유할 수 있으며 정보의 종류가 다양해지고 정보의 양도 증가하여 필요로 하는 정보를 찾기 위한 시간과 노력은 증가하게 된다. 이러한 상태를 Information Overload라 한다. 이와 같이 정보 필터링은 인터넷의 발달과 더불어 발전하기 시작한 네트웍 정보의 자동화된 전달에 따른 정보탐색의 한 분야로 정의할 수 있으며 시간적으로 생성되는 대량의 정보를 이용자의 정보요구를 만족시킬 수 있도록 중요도에 따라 정렬하여 정보원을 제공하는 것을 의미한다[15]. 또한 정보 필터링은 비정형 또는 반정형 데이터를 이용하는 시스템이며 대용량의 유동적인 문자정보를 다루는 시스템이라 할 수 있다[3]. 정보 필터링과 정보검색의 차이는 정보검색이 불특정 검색요구에 대하여 특정의 정보를 찾아내는 것이며 정보 필터링은 이용자의 특별한 정보요구에 대응하여 유동적인 정보원으로부터 적합한 정보를 선별 제공하는 것을 특징으로 한다.

정보 필터링의 적용 영역으로는 유즈넷 뉴스분야가 시초를 이루나 전자우편과 웹에도 파급이 되어 기하급수적이며 시시각각으로 정보의 유입되고 원래부터 정보가 전자적으로 이루어져 시스템을 설계하고 실험할 수 있는 환경이 될 수 있으며[15] 이용자의 입장에서 정보의 형태보다는 내용을 중심으로 하여 적용영역을 일반적인 뉴스, 서지정보, 학술저널 목차, 원문 정보 등으로 확장할 수 있다.

Information Overload를 해결하기 위해 사용하고 있는 정보 필터링 시스템은 다음과 같은 문제점을 가지고 있다. 사용자의 관심을 반영한 사용자 요구(user profile)를 바탕으로 필터링을 수행하기 위해서는 많은 학습시간을 필요로 하며, 사용자 프로파일이 사용자별로 개인화되어 있지 않을 경우 효율적인 필터링을 기대하기 어렵다. 또한 새로운 사용자나 새로운 분야의 정보를 요구할 경우 사용자의 선호를 반영한 사용자 프로파일이 구축되어 있지 않을 경우 처음부터 학습을 새로 시작해야 하는 문제가 발생한다[21].

정보 필터링 에이전트는 인터넷으로부터 이용자 프로파일을 구성하여 이용자의 관심이 있는 정보를 검색하여 이용자에게 제공해 주는 에이전트다[7]. 정보 필터링 에이전트가 정보검색의 부속물로 정의되기도 하는데 사용자의 정보요구로부터 저장된 추가의 정보를 사용하여 불필요한 문서를 제거하고 검색된 문서 집합을 정제하기 때문이다. 정보필터링 에이전트는 웹 검색 에이전트와 같이 대량의 온라인 정보로 인한 Information Overload를 다루지만 낮은 정확률의 문제점을 보완하기 위해 사용자 정보 프로파일을 다루는 점에서 차이가 있다. 웹 검색 에이전트가 사용자가 관심을 갖는 특정 웹 사이트를 찾는데 유용한 반면 정보 필터링 에이전트는 정보를 다양한 근원지로부터 모은 후 사용자 개인의 선호도에 기반하여 여과된 정보를 사용자에게 제공한다[1]. 그러나 정보 필터링 에이전트 또한 사용자에게 제시하는 모든 텍스트를 다루는데 많은 시간과 노력이 필요하다는 문제점이 있다.

2.2 스팸메일 필터링(Spam mail Filtering)

스팸메일이란 발신자가 다른 목적으로 수신자의 동의 없이 전자메시지를 발송하거나 아무 관계가 없는 수신자에게 발송된 유익하지도, 원하지도 않은 전자메시지를 말한다. 스팸메일은 전자우편 이용자에게는 많은 시간과 비용을 낭비하게 하며 웹 메일 서비스

업자에게는 인터넷 체증 가중 및 통신 저하 등 유무형의 피해를 야기하고 있다. 최근 들어 바이러스 유포, 사용자 정보 해킹 등의 목적을 가진 스팸메일의 대량 유포로 사용자 정보를 해킹하거나 시스템에 바이러스를 감염시키는 등의 반 사회적인 문제점은 간과할 수 없게 되었다.

일반인들에게 이메일은 편지나 전화를 대체하는 도구로 활용되나 기업환경에 있어서 이메일은 업무 환경의 시간적, 공간적 문제를 극복할 수 있게 하며 사내 커뮤니케이션을 원활하게 하는 장점을 제공하며 회사 내 업무처리과정을 전체적으로 변화시키는 중요한 수단으로 성장하였다[19] 그러나 스팸메일의 등장은 그 반대급부가 크게 나타나고 있다. 스팸메일은 개인에게는 정신적, 물리적 스트레스를 증가시키고 업무진행을 지연시키며 스팸메일 발송으로 메일서버에 부하를 주며 공공자원인 네트워크 자원을 독점하는 결과를 가져오고 있다[12]. 따라서 수신자에게 스팸메일이 도착하기 이전에 이메일을 필터링할 수 있는 방법에 대해서 많은 관심이 기울여지고 있다.

정옥란 등[11]은 사용자의 메일 처리과정을 일정기간 관찰하여 각각 개인에 맞는 규칙을 형성하고 만들어진 규칙을 바탕으로 개인에게 불필요한 메일이나 스팸메일을 삭제토록 하는 개인화된 분류를 위한 웹 메일 필터링 방법을 제안하였으며 정확도를 높이기 위해 베이지안 알고리즘을 적용하였다.

신경식 등[19]은 데이터 마이닝 기법 중 분류 문제에 많이 사용되는 인공신경망과 의사결정나무 기법을 이용해서 스팸메일의 분류와 예측을 가능케 하는 모형을 구축하였다. 의사결정나무 기법을 적용한 스팸메일 필터링이 조금 더 나은 분류 성과를 보이고 있으나 영문 이메일 데이터를 이용한 점은 연구의 한계점이 될 수 있다.

서정우 등[18]은 기존의 내용기반 이메일 필터링이 특정 단어의 패턴 매칭을 통해 필터링을 수행하나 스패머들이 제목이나 본문의 내용을 변형하여 스팸메일을 발송할 경우 이를 효과적으로 필터링하지 못하는 점에 착안하여 패턴 분류 문제에 특히 높은 성과를 보이는 SVM(Support Vector Machine)을 사용하여 생성된 색인어와 단어사전의 매칭을 통해 얻어진 데이터 셋을 SVM 분류기에 적용하여 정상 메일과 스팸메일을 분류하는 방법을 제시하였다. 특별히 많은 스팸메일의 내용을 차지하는 성인광고와 대출에 관련된 메일을 대상으로 실행한 결과 좋은 성능을 가지고 있음을 알 수 있었다.

조한철 등[6]은 스팸메일 필터링에 있어서 사용자가 직접 규칙을 작성할 필요 없이 학습을 통해 얻은 데이터로 자동으로 스팸메일을 필터링하는 시스템을 제안, 기존 문서분류에서 쓰이는 베이지안 학습방법 중 널리 쓰이는 통계적 알고리즘인 나이브 베이지안 분류자를 이용하였다. 기존의 규칙기반 시스템과 나이브 베이지안 분류자를 비교하여 분석한 결과, 오류율, 스팸재현율 등에 있어 나이브 베이지안 분류자를 이용하였을 때 정확도가 향상됨을 보이고 있다.

2.3 맥락 필터링(Contextual Filtering)

Contextual Filtering의 활용분야는 매우 다양하다. 지문인식분야에서의 Contextual Filtering은 인식된 지문 이미지의 상세분류를 위한 지문 이미지 강화(Fingerprint Image Enhancement)에서 활용되고 있다.

Contextual Filtering은 [그림 1]과 같이 지문인식 분야에서는 낮은 품질의 지문 이미지를 인식하여 분류하는데 있어서 강력한 도구로 자리잡고 있다.

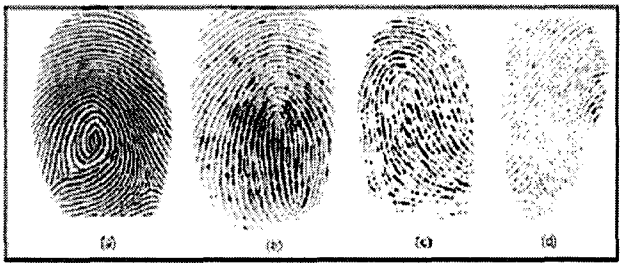


그림 1 Fingerprint Images

지문이미지에서 지문 융기선의 시작과 융기선의

빈도, 융기선의 계속성, 경향 등의 정보를 스펙트럼 분석을 실시하여 지문 이미지를 성공적으로 분류하고 있다[5].

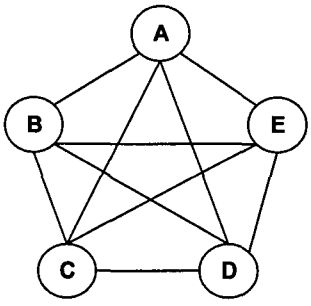
Uckun 등[20]은 일반 비행사고의 대부분은 비계기비행 등급의 조종사가 의도적으로 계기비행을 하여야 하는 기상조건에 진입하였을 경우와 계기비행이 가능한 조종사가 천둥 번개, 우박, 돌풍과 같은 불안정한 기상환경에 처했을 때 발생하는 점에 착안하여 이를 해결하기 위해 ARARE라는 기상 요약 및 보고 시스템을 개발하였다. AWARE 시스템은 텍스트 기반의 정보와 그래픽한 비행 기상정보를 통합하여 임무상황과 장비특성에 맞는 상황경고를 제공하는 시스템이다. 최상의 비행 기상 브리핑은 기상전문가가 조종사가 필요로 하는 기상정보를 원시 기상자료를 해석하여 제공하는 것이나 처리속도가 늦어 자동화된 서비스로 대체되었다. 그러나 자동화된 서비스는 방대한 양의 기상자료를 제공은 하나 계획된 비행임무 상황에 맞는 정보를 제공하지는 못하고 있다. 따라서 비행 기상 정보예보에 있어서 Contextual Filtering은 상황이나 지역적 수준에 맞도록 조종사가 실제로 필요로 하는 비행경로상의 기상현상을 제공하고 있으며 이러한 방대한 양의 기상정보를 축약하여 제공할 수 있는 것은 Contextual Filtering로 인해 가능하였다.

Ruch 등[16]은 의료보고서 편집의 향상을 목적으로 문장 내 존재하지 않는 단어로 인한 철자오류 정정 시스템을 개발하였다. 기존의 시스템들과 달리 의미론적이나 구문론적인 방법을 사용하였다.

2.4 연결빈도행렬

연결 빈도 행렬은 인접 행렬의 개념에서 출발한다. 인접 행렬(Adjacency Matrix)은 데이터의 인접성(Adjacency)을 이용하여 의사결정공간(Decision Space)에서 유용하게 쓰일 수 있는 개념[13]으로 품목 A가 품목 B가 존재할 때 품목 A와 품목 B가 동시에 구매되었는지, 또는 품목 A가 품목 B의 구매에 영향을 주었는지 그 여부를 확인할 수 있어 데이터 마이닝의 기법 중 연관규칙 분석[4]이나, 또는 추천 시스템[17] 및 데이터 시각화[9] 등에서 이용되고 있다.

이와 같은 데이터의 인접성은 유한개의 점과 선으로 구성된 도형인 연결그래프(Connected Graph)와 사상(Mapping)의 개념이나 도형의 위상적 성질을 이용하면 명확히 알 수 있다. 연결그래프에서 단위 정보를 표현하는 점을 "Vertex", 각 점을 잇는 선을 "Edge"라고 하고 또한 Edge에 방향성이 있는가에 따라서 유향 그래프(Directed Graph)와 무향 그래프(Undirected Graph)로 구분한다. 그 밖에 그래프 내에서 여러 Vertex 들의 연결과정을 경로(Path)라고 하며, 시작점과 끝점이 연결된 경로는 특별히 순환(Cycle)이라고 한다. 연결그래프에서 선(Edge)이 점(Vertex)을 공유하고 있으면 1, 공유하고 있지 않으면 0으로 나타낸 행렬을 인접행렬이라고 한다. 인접행렬은 N 개의 Vertex를 가지는 n * n 정방행렬이다 인접행렬의 어떤 원소 A*ij* = 1 이면 두 Vertex가 인접해 있다는 것이며, A*ij* = 0 이면 두 Vertex는 인접해 있지 않은 것이다. [그림 2]는 연결그래프와 이에 대한 인접행렬의 예이다. 연결빈도행렬은 이러한 인접행렬의 특성에 방향과 누적빈도를 추가한 개념이다.



	A	B	C	D	E
A	0	1	1	1	1
B	1	0	1	1	1
C	1	1	0	1	1
D	1	1	1	0	1
E	1	1	1	1	0

그림 2 연결그래프와 인접행렬의 예

2.5 브레인 매핑

인간의 뇌는 하나의 작은 우주라고 말할 수 있으며 인간의 뇌와 관련하여 연구할 분야는 무궁무진하며

특히 인간의 뇌에서의 정보처리과정은 비중이 있는 연구분야이다. 인간의 뇌는 정보처리과정에 있어 주어진 프로그램에 따라 한번에 하나의 명령을 정보로 변환하고 이 정보에 기초하여 다음의 과정을 결정하고 한번에 하나의 정보를 처리하는 컴퓨터의 직렬 정보처리 방식과는 다르다. 인간의 뇌는 다수의 뉴런이 복잡하게 연결된 네트워크를 이루고 입력정보가 들어오면 다수의 뉴런에 전달이 되며 이러한 상호작용이 뇌의 전체에 퍼져 동시에 병렬적으로 정보를 처리하고 있다. 이러한 인간의 뇌에서의 정보처리 능력을 인공지능에 많은 응용을 하고 있다. 신경과학(Neuroscience)의 발달로 인해 뇌의 정보처리 능력을 인공지능(Artificial Intelligence)에 응용하면서 뇌 연구의 지식영역이 점점 확대되고 있다. 인간의 뇌에 대한 연구는 기초과학, 공학, 의학, 심리학 등 여러 분야가 연관되어 있으며 미래형 핵심기술로 경제, 사회, 기술적 파급효과는 크다고 할 수 있다.

인간의 기능적 브레인 매핑(HFBM: Human Functional Brain Mapping) 분야는 크게 성장하고 있고 학제적(Interdisciplinary) 연구가 크게 집약되는 분야이다. 기능적 브레인 매핑의 개념은 최초 의학에서 뇌의 활동을 시각화하기 위해 두피의 다른 지점 사이의 뇌의 전기 활동성을 측정하는 데는 시작하였으며 측정의 결과는 뇌파도, 뇌전도라 하여 1928년 Hans Berger가 최초로 시도하였다[14]. 의학에서의 기능적 브레인 매핑에 대한 연구는 관련 학문들과의 학제적 연구를 통해 다양하게 연구되고 있다.

Fox 등[10]은 적절한 실험디자인을 통한 브레인 맵을 이용해 데이터 필터링에 응용할 수 있음을 보였으며 또한 브레인 맵 분류방법을 통해서 데이터베이스 구축을 위한 메타데이터 스키마를 만드는 데에도 유용함을 보였다. 또한 Law 등[14]은 전자 두뇌 그림 (EEG: electroencephalogram)의 사용을 통해서 두뇌가 무엇을 하고 있는지를 시각화하는 방법을 확장 발전시켰는데 EEG는 인간의 뇌는 끊임없는 전기적 활동을 하고 있으며 이러한 활동의 결과는 기록할 수 있다는 것으로 기록은 뇌에 존재하는 수 천 개의 뉴런의 활동의 결과이다. 뉴런의 활동패턴은 인간의 심리 상태에 따라 달라지는데 빠르고 느린 뇌파의 패턴을 단순한 시각화로 나타내면 다음 [그림 3]과 같은 패턴 특징을 나타낼 수 있다.

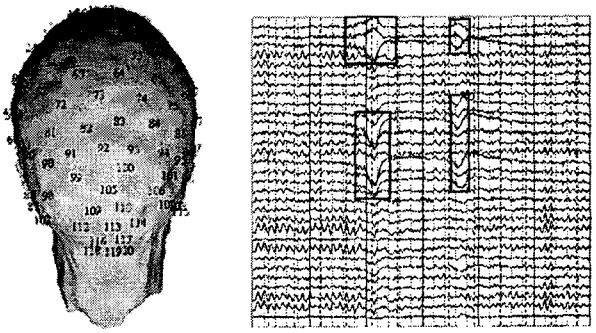


그림 3 뉴런의 활동 패턴과 뇌에서의 기억장소

최근에는 [그림 4]와 같이 두뇌의 전기적 활동의 파라미터들을 EEG 지형도에 3차원과 칼라로 묘사하여 그릴 수 있는 소프트웨어의 개발로 두뇌의 전기적 활동을 3차원으로 재구성하여 나타내고 있다.

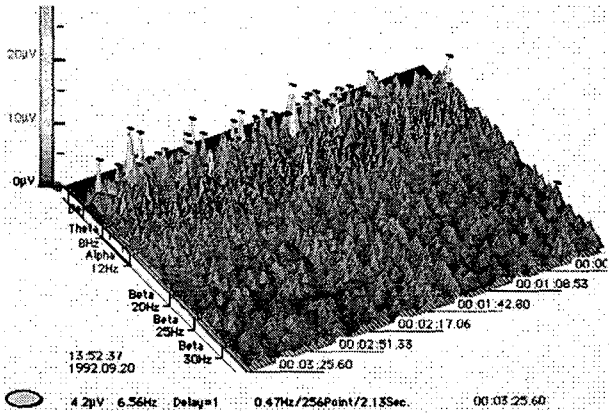


그림 4 3D Color EEG brain topography

3. 연구 방법

3.1 자료 수집

인간의 정보 처리과정을 시뮬레이션하는 인지 필터링 시스템을 구현하기 위해 자료 수집 대상으로 남녀 대학생 500명을 대상으로 하여 문장을 작성하도록 하였다. 문장은 50개의 단어를 이용하여 문장을

작성하였으며 사용된 단어는 국립국어연구원에서 2003년 5월에 발표한 한국어 학습용 어휘 목록 중 50개를 무작위로 선정하여 사용하였다. 피실험자는 1인당 5개의 문장을 주어진 단어를 이용하여 자유롭게 작성하도록 하였으며 한 문장당 주어진 50개의 단어 중 최소 3개 이상의 단어를 사용하여 문장을 작성하게 하였다.

수집된 총 2,500여 개의 문장 중 설문 목적에 맞지 않는 문장을 제외한 2,300개의 문장을 정리하였다. 정리된 문장은 다시 250명의 대학생에게 배부하여 각각의 문장을 "보관"과 "삭제"로 분류하게 하였다.

문장을 50개의 단어를 이용하여 작성한 이유는 신경망 분석의 경우 연구에서 처리하고자 하는 데이터에 포함되는 단어가 100개 이상이 되면 모형구축이 불가능한 점이 있어 사용단어의 수를 제한하였다. 이와 같은 실험결과는 베이지안 네트워크, 의사결정나무, 신경망, K-NN을 이용하여 뉴스기사 자동분류시스템 구축 연구[2]에서도 신경망의 경우 실험에 포함된 단어 수가 증가할 경우 입력이 되지 않아 단어 100개 이상은 처리할 수 없어 다량의 단어를 이용한 실험은 베이지안 망과 의사결정나무를 이용한 것에서 확인할 수 있다.

3.2 분석절차

본 연구는 1단계에서는 분류 예측 분석을 위해 수집된 각 문장을 사전 처리하는 단계로 각 문장을 보관과 삭제로 분류하였다. 2단계에서는 분류된 문장을 트레이닝(2,000개)과 테스트(300개)로 분류하여 테스트 데이터가 트레이닝 데이터에 포함되지 않도록 하였다. 3단계에서는 보관할 문장과 삭제할 문장으로 분류된 트레이닝 데이터를 이용, 학습 후 테스트 데이터로 기존의 분류 모델들과 본 연구에서 제안하고자 하는 인지 필터링 시스템과의 분류 예측 정확도를 비교하였다. 4단계에서는 3단계에서 실행된 모델들에서 산출된 결과의 정확성을 검증하기 위해 총 10회에 걸친 교차검증을 실시하였다.

3.3 연구모형

본 연구에서 제안하는 cognitive mapping의 주요 개념은 인간의 정보처리과정을 시뮬레이션하는 것이다. 인간의 뇌 (Brain)는 약 수 십억 개의 신경세포 (Neuron)와 이들을 상호 연결하는 약 수 십조 개의 시냅스 (Synapse)로 구성되어 이들의 복합 작용에 의해 사람들은 사물을 인식하고 어떻게 행동할 것인지를 판단한다. 인지과학 (Cognitive Science) 분야에서는 이와 같은 인간의 두뇌에 의한 정신활동이나 신체기능을 추상적으로 다루지 않고, 구체적인 기술로서 재현하려고 한다. 즉 인간이 느끼고, 사고하고, 말로 표현하는 것을 추상적으로 표현하는 것이 아니라, 구체적 공식이나 절차로 재현하려고 한다. 인간의 뇌는 시각 (Vision) 정보를 통하여 문자를 인식하고 의미를 이해한다. 인간의 뇌는 시각을 통해 획득한 정보들을 각각의 지정된 장소에 저장하며 저장된 정보는 지도와 같은 형태로 [그림 5]와 같이 나타낼 수 있다. 이러한 인간의 능력을 컴퓨터로 실현하려는 것이 패턴인식 (Pattern Recognition)의 분야이며 이 분야에는 광학 문자 인식 (Optical Character Recognition), 우편물 자동 분류, 문서인식, 도면인식 등의 분야가 부분적으로 실용화가 이루어졌으며, 최근에는 인공지능 (Artificial Intelligence)의 최신 기법인 신경망(Neural Network), 퍼지(Fuzzy), 유전 알고리즘 (Genetic Algorithm) 등의 응용과 자연어처리 (Natural Language Processing), 심리학, 생리학, 인지과학 (Cognitive Science) 등 관련 학문과의 접목에 의해 문자인식 기술은 새로운 단계에 접어들고 있다.

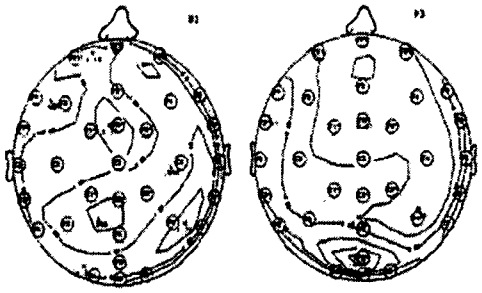


그림 5 뇌에서의 기억장소

본 연구에서는 문서에 사용한 단어의 출현 빈도와 단어와 단어간의 연결은 새로운 의미를 창출하고 창출된 정보는 선택적 주의집중을 받으며 정보를 처리하는데 중요한 역할을 수행하게 된다는 정보처리 메커니즘을 기반으로 하여 필터링에 응용하는 것이다. 예를 들어 다음의 문장에 "여배우", "죽음"의 단어가 출현할 경우 이용자는 읽기를 선택하기 보다는 삭제할 문장으로 분류할 확률이 높다. 그러나 유명 여배우의 죽음이라는 상황적 요인이 발생하였을 경우 이용자는 두 단어의 연결이 창출한 새로운 의미를 정보처리과정에 활용할 것이다. 이와 같이 이용자가 살펴 본 단어들은 인간의 정보 인지과정에 따라 뇌의 특정 주소에 기록, 저장이 되고 저장이 된 정보를 개인이 축적해 온 경험과 학습에 따라 해당 단어들이 주는 정보의 가치를 판단하게 될 것이다. 인간의 뇌에서 획득된 정보를 브레인 매핑을 응용하여 획득된 정보를 평면의 공간에 나타낼 수 있다.

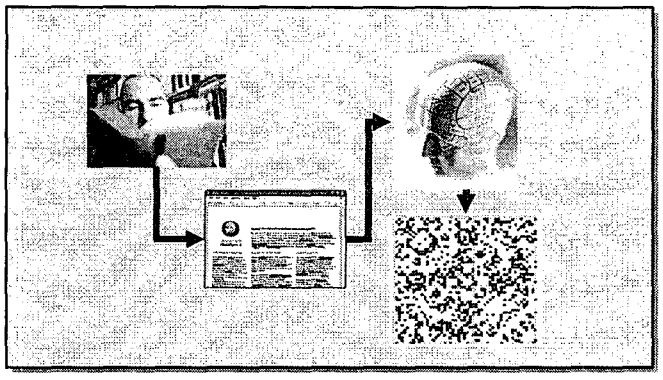


그림 6 인간의 정보처리개념

문서 분류에 있어 문서에 사용된 단어의 출현빈도와 인접한 단어의 연결과 연결된 단어가 창출하는 의미가 중요한 역할을 한다. 특히 단어의 빈도는 문서를 분류하는데 있어서 중요 역할을 수행해 문서 분류방법으로 많이 이용되고 있다.

본 연구에서는 획득된 정보가 뇌에서의 특정 장소에 위치하는 것을 나타내는 브레인 매핑과 인접성을 이용하여 분류에 사용하는 연결빈도행렬, 문서에 사용된 단어의 출현빈도를 이용하여 분류하는 웨이티드 키워드 매칭방법을 종합하여 필터링에 이용하였으며 특히 단어 출현빈도에 다양한 가중치를 적용하여 분류 예측에 있어 가중치의 역할을 분석하였다. Cognitive Mapping을 그림으로 구체화하면 다음과 같은 n * n 행렬로 나타낼 수 있다.

	a_1	a_2	a_3	⋮	⋮	⋮	⋮	⋮	⋮	a_n
a_1	43	36	36	78	65	98	55	99	5	78
a_2		56	26	92	60	57	95	14	79	38
a_3			32	49	59	68	79	64	88	93
.				77	49	71	55	04	22	51
:					45	49	43	74	75	82
:						92	82	67	91	86
:							77	96	67	49
:								56	87	97
:									53	38
a_n										81

그림 7 Cognitive Mapping

Cognitive Mapping은 Microsoft의 Visual C++를 이용하여 구체화였다. 기존의 분류방법들과의 Cognitive Mapping과의 차이를 그림으로 도식화한 것은 [그림 8]과 같다.

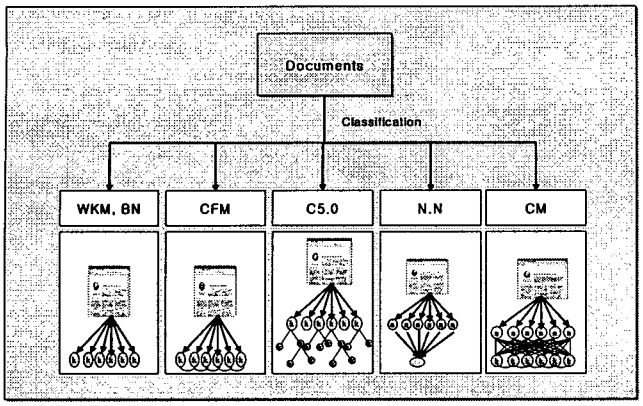


그림 8 기존 분석방법과 비교

3.4 분석결과

인지 필터링 시스템을 구체화한 Cognitive Mapping의 유용성을 검증하기 위한 비교대상으로는 문서분류시스템에서 활용도가 높은 키워드매칭, 웨이티드 키워드매칭, 인공지능에서 성과가 높은 신경망, 의사결정나무, 확률적 학습방법을 채택하여 분석하는 베이지안 망, 단어의 인접성을 이용하는 연결빈도행렬을 이용하여 비교·분석하였다.

본 연구에서는 분류 예측의 정확도 측정에 있어서 타당성을 높이기 위해 총 10회에 걸친 Cross Validation을 실시하였다. 또한 문서 필터링에 있어서 출현된 단어의 가중치가 중요한 역할을 하는 점을 이용하여 제안하고자 하는 Cognitive Mapping에 출현 단어별로 가중치를 적용하였으며 적용한 가중치의 범위는 0.1 ~ 1 사이로 변화시키면서 분석을 실시하였다.

각각의 분석기법이 테스트 데이터를 보관과 삭제로 분류하는 정확도 분석결과 높은 예측정확도를 보이는 분석 기법은 본 연구에서 제안하는 Cognitive mapping, 신경망, 베이지안 네트워크, 웨이티드 키워드 매칭, 의사결정나무 순으로 분류예측의 정확도를 보이고 있다. 키워드 매칭의 경우 단어의 출현여부에 따라 분류하는 것이므로 키워드의 존재여부로만 분류하는 것은 예측정확도가 낮은 것이라는 예상과 같이 예측정확도는 매우 낮아 결과에서는 제외하였다.

또한 문서의 분류에 있어 사용 단어 빈도의 역할을 파악하고자 가중치를 Cognitive Mapping에 적용한 결과 가중치 0.1~0.3에서 예측정확도가 증가하고 그 이후의 수치까지 변화시켰으나 정확도의 변화가 없음을 알 수 있었다. 또한 가중치를 1 이상 적용하였으나 분석 결과에 영향은 없었다[그림 9 참조].

본 연구에서 제안한 Cognitive mapping은 분석에 포함되어야 하는 변수인 단어의 개수에 제한을 받지 않는 것으로 입력변수의 제한을 받는 타 분석 기법과 비교를 할 때 예측 정확도뿐만 아니라 실용성 측면에서도 우수함을 보이고 있다.

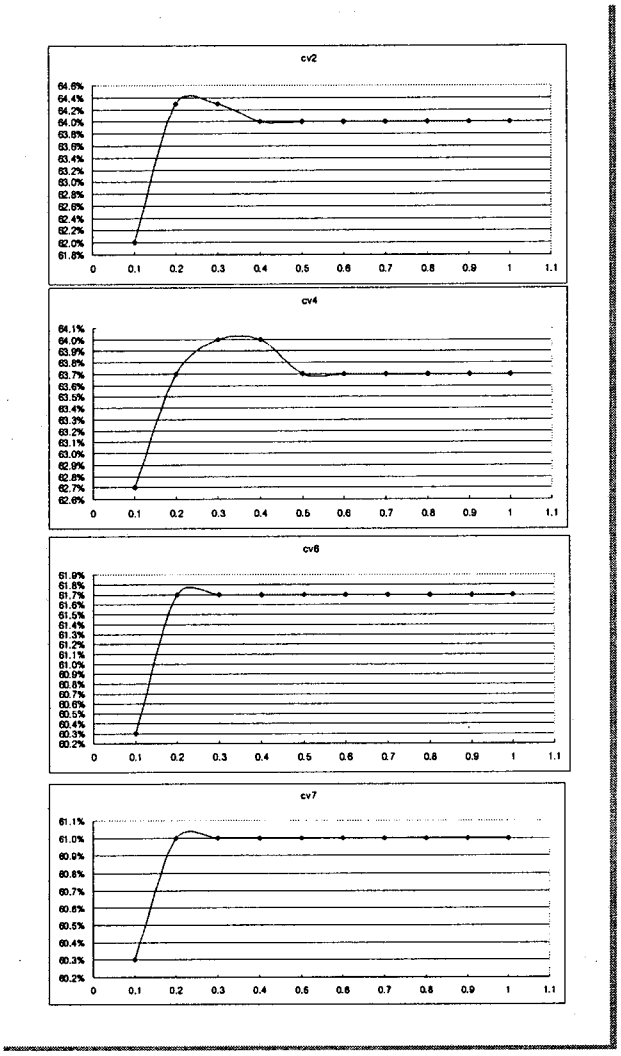


그림 9 가중치 적용결과

4. 결론

대량의 정보를 획득하여 필요한 정보만을 의사결정 또는 지식창출에 이용해야 하는 현대 사회에서 정보 필터링은 중요 연구영역이기도 하다. 특정 시스템만으로 정보 필터링을 할 수 있는 것이 아니라 일상생활에서 자신도 모르는 사이 정보 필터링을 수행하고 있는 것과 같이 인간의 뇌에서 정보를 획득하여 필요/불필요의 정보를 분류하는 것을 시뮬레이션하여 정보 필터링에 응용하고자 한 본 연구는 기존의 분류모델들인 신경망, 의사결정나무, 베이지안 망, 키워드 매칭, 웨이티드 키워드 매칭, 연결빈도행렬 등과의 예측 정확도의 비교시에도 우수성이 입증되었으며 분석 측면에서도 입력단어 수에 제한을 받지 않으므로 활용측면에서도 우수하다고 할 수 있다.

이용자의 요구에 맞는 정보를 얻기 위해 사용하는 정보 필터링 시스템이 이용자의 의도와 다르게 정보를 분류하거나 이용자의 다양한 요구를 반영하지 못할 때는 정보 필터링을 사용하지 않은 경우 보다 못할 수 있다. 본 연구의 의의로는 정보 필터링의 정확도를 향상시키기 위해 인간의 뇌에서의 정보처리과정을 시뮬레이션하는 인지적 매핑의 정보 필터링 시스템을 제안한 것과 특정 단어 또는 패턴만을 이용하여 필터링하는 기존 시스템과는 달리 단어의 존재, 단어와 단어의 연결이 창출하는 의미와 단어의 가중치를 종합하여 정보를 필터링하는 점에서 의의가 있다.

표 1 분석결과

단위 : %

구분	웨이티드 키워드매칭	연결빈도행렬	의사결정나무	신경망	베이지안 망	Cognitive Mapping
1차	61.67	58.00	47.33	59.67	59.33	62.00
2차	64.00	59.00	45.67	61.67	64.00	64.30
3차	59.30	60.70	54.00	61.00	59.33	59.70
4차	64.00	61.00	48.33	63.33	64.00	64.00
5차	61.00	60.30	52.00	61.33	61.00	63.00
6차	61.70	57.00	52.00	64.33	61.67	61.70
7차	61.00	56.30	51.33	61.00	61.00	61.00
8차	55.70	50.30	55.00	60.33	55.67	59.70
9차	59.00	54.00	50.00	57.00	59.00	59.00
10차	65.00	61.70	43.00	64.00	65.00	65.30
평균	61.19	57.83	49.87	61.37	61.19	61.57

References

[1] Alper, K. C., and Collin, G. H.(1997). *Agent Sourcebook,* Wiley, New York.

[2] Baeg, Y. G., and Seo, Y. M. (2003). "An Auto Classification Information System for Intenet News," *Proceedings of the Conference on The Korea Society of Management Information System, Fall,* pp. 574-581.

[3] Belkin, N. J. and Croft, W. B.(1992). "Information Filtering and Information Retrieval: Two Sides of the Same Coin?," *Communication of ACM,* Vol. 35, pp. 29-38.

[4] Berry, M., Linoff, G. (1997). *Data Mining Techniques,* Wiley, New York.

[5] Chikkerur, S., Wu, C., and Govindaraju, V. (2004). "Systematic approach for feature extraction in Fingerprint images," *1st International Conference on Biometric Authentication,* Hong Kong.

[6] Cho, H. C. and Cho, G. S. (2002). "Spam-mail Filtering System Using Naïve Bayesian Classifier and Message Rule," , *Proceedings of the Conference on The Korea Information Science Society,* Vol. 29, No. 1, pp. 223-225.

[7] Cho. Y. Y. (2003). *Artificial Intelligence System,* Hongneung.

[8] Choi. H. Y. (1998). " A Study on Information Service on Demand Using Information Filtering," *Journal of the Korean Society for Information Management,* Vol. 1, pp. 63-81.

[9] Condon, E., Golden, B., Lele, S., Raghavan, S., and Wasil, E. (2002). "A visualization model based on adjacency data," *Decision Support Systems,* Vol. 33, No. 4, pp. 349-362.

[10] Fox, P.T., Laird, A.R., Fox, S.P., Fox, P.M., Uecker, A.M., Crank, M., Koenig, S.F., and Lancaster, J.L. (2005). "BrainMap Taxonomy of Experimental Design Description and Evaluation," Human Brain Mapping, Vol. 25, pp.185-198..

[11] Jeong, O. R. and Cho, D. S. (2003). "Design and Implementation of Web Mail Filtering Agent for Personalized Classification," *Korea Information Processing Society Journal B,* Vol. 10, No. 7, pp.853-862.

[12] Jung, J. J. (2000). *E-mail Marketing,* Web-mania, Seoul.

[13] Kim, J. H., Nam, K. C. and Byun, H. S.(2004). "Application for Visualizing Web Navigation Patterns and Recommendation," *Proceedings of the Conference on The Korea Society of Management Information System, Fall,* pp. 534-540.

[14] Law, S. K., Nunez, P. L., Westdorp, A. F., Nelson, A. V., and Pilgreen, K. L. (1991). "Topographical mapping of brain electrical activity," *Proceedings of the IEEE Conference,* pp.194-201.

[15] Orad, D. W. (1996). " A Conceptual Framework for Text Mining," *User Modeling and User-Adapted Interaction,* Vol. 7, NO. 3, pp. 141 – 178.

[16] Ruch, P., Baud, R., Antonie, G., Lovis, C., Rassinoux,, A., Riviere, A. (2001). "Using Part-of-Speech and Word-Sense Disambiguation for Boosting String Edit Distance Spelling Correction," *Artificial Intelligence Medicine*: 8th Conference on AI in Medicine in Europe, AME.

[17] Schafer, J., Konstan, J., and Riedl, J. (2001). "E-commerce recommendation applications," *Data Mining and Knowledge Discovery,* Vol. 5, No. 1&2, pp. 115-153.

[18] Seo, J. U., Son, T. S., Seo, J. T., and Mun, J. S.(2004). " A Study on the Filtering of Spam e-mail Using n-Gram indexing and Support Vector Machine," *Journal of The Korea Institute of Information Security and Cryptology,* Vol.14, No.2. pp. 23-33

[19] Shin, K. S., and Ahn, S. S.(2002). "Spam Mail Classification Modeling Using Data Mining Techniques," *Ehwa Management Review,* Vol. 20, pp.89-105.

[20] Uckun, S., Ruokangas, C., Donohue, P., Tuvi, S. (1999). "AWARE: Technologies for interpreting and presenting aviation weather information," *Aerospace Conference, 1999. Proceedings 1999* IEEE, Vol. 2.

[21] Yang, J. Y., Hong, G. H., and Choi, J. M.(1999). "An Intelligent Collaborative Information Filtering Agent for Effiecient Information Filtering, *Proceedings of the Conference on The Korea Information Science Society, Fall*, Vol. 26, No. 2, pp. 69-71.