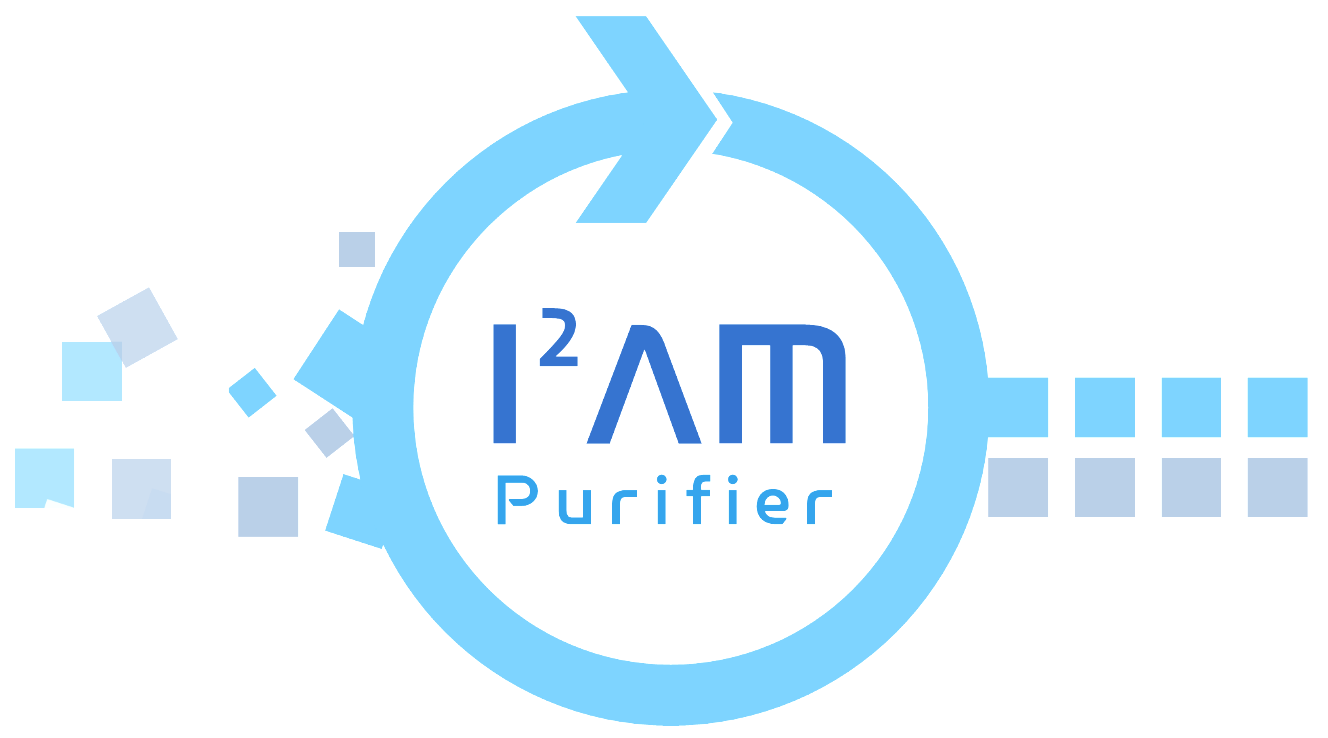
****

2016. 11.

강원대학교, 연세대학교

필터링 기술 분석서

문서 정보

|  |  |
| --- | --- |
| **버 전** | 0.1 |
| **작성일** | 2016-11-21 |
| **상 태** | 🞎 완료 🞎 진행 중 🗹 초안 |
| **작성자** | 김하진(강원대학교) |
| **검토자** | 이름(소속) |
| **승인자** | 이름(소속) |

**목 차**

[필터링 기술 분석서 1](file:///D:\대학원\1.연구실\2.프로젝트\IITP_I2AM\2팀\평가자료\필터링%20기술분석서%20161110.docx#_Toc467486547)

[문서 정보 2](#_Toc467486548)

[1. 개요 1](#_Toc467486549)

[1.1. 정의 1](#_Toc467486550)

[1.2. 응용 분야 1](#_Toc467486551)

[2. 필터링 기법 2](#_Toc467486552)

[2.1. 학습 기반 필터링 2](#_Toc467486553)

[2.1.1. 내용기반 필터링 3](#_Toc467486554)

[2.1.2. 협업 필터링 4](#_Toc467486555)

[2.1.3. 베이지안 필터링 5](#_Toc467486556)

[2.2. 비학습 기반 필터링 6](#_Toc467486557)

[2.2.1. 질의 필터링 7](#_Toc467486558)

[2.2.2. 블룸 필터링 7](#_Toc467486559)

[2.2.3. 칼만 필터링 8](#_Toc467486560)

[3. 기술 동향 9](#_Toc467486561)

[3.1. 관련 제품군 9](#_Toc467486562)

[3.2. 관련 기술연구 10](#_Toc467486563)

# 개요

## 정의

필터링이란 스팸메일과 같이 정보 이용자가 원하지 않는 정보를 걸러 내는 일 또는 그런 기능이나 과정을 말한다. 데이터베이스에서 사용자가 원하는 정보를 뽑아내거나 원하지 않는 정보를 걸러내기 위해 사용하며, 이상치나 결측치를 탐지하여 데이터를 정제하기 위해 사용하기도 한다. 또한 스트림 환경과 같이 대량의 데이터가 끊임없이 유입되는 상황에서 실시간으로 데이터를 처리하기 위해 필터링하여 데이터를 줄이는데 사용하기도 한다. 필터링은 사용자가 원하는 필터 조건을 입력하여 ‘필터(filter)’를 생성하고, 필터에 데이터를 통과시켜 원하는 데이터만 추출하거나 원하지 않는 데이터를 걸러내는 방식으로 이루어진다. 예를 들어, 온도 데이터베이스에서 온도가 30도가 넘는 날만 추출하고자 할 때, 필터 조건을 ‘30도 이상’으로 설정하여 필터를 생성하고 데이터에 적용하는 것이다.

## 응용 분야

필터링은 웹, 음악, 데이터 마이닝, 신호처리, 이미지처리 등 데이터를 생성, 처리하는 다양한 분야에서 사용된다. 웹 분야에서 필터링은 불건전하거나 저작권을 위반한 음원 또는 유해한 내용이 포함된 웹 페이지나 이메일에 대한 접근 및 사용을 차단하기 위해 사용된다. 웹 분야에서 필터링은 원하지 않는 내용의 문자열 등을 확인하는 방식으로 이루어진다.

데이터 마이닝 분야에서는 고객이 선호할만한 아이템을 추천해주기 위해 필터링을 사용한다. 고객들의 선호도와 관심 표현을 바탕으로 필터를 생성하여 비슷한 패턴을 가진 고객들을 식별해내고, 이를 바탕으로 비슷한 취향을 가진 고객들에게 서로 아직 구매하지 않은 상품들을 교차 추천하거나 관련 상품을 추천하는 서비스를 제공한다. 인터넷 서점 아마존(Amazon)이 이런 필터링 기술을 사용하고 있는데, 루시디의 “한밤중의 아이들”을 구매한 사람은 아룬다티 로이의 “작은 것들의 산”도 좋아할 것이라고 예측하여 추천하는 식이다.

신호처리에서 필터링은 원하는 신호의 주파수 범위로부터 노이즈를 제거하는 것으로, 주파수 스펙트럼에서 원하지 않는 부분을 제거하는 것을 의미한다. 일반적으로 저역통과, 고역통과, 대역통과, 대역 저지 등 4가지 필터로 나뉘며 이는 아날로그와 디지털 필터로 구현될 수 있다. 신호처리 분야에서 쓰이는 필터는 대표적으로 디지털 필터 중 유한 임펄스 응답(finite impulse response) 필터와 무한 임펄스 응답(infinite impulse response) 필터가 있다. 유한 임펄스 응답 필터는 필터의 길이가 한정된 필터로, 설계가 쉽다는 장점이 있다. 무한 임펄스 응답 필터는 필터의 길이가 무한한 필터로, 설계가 어렵지만 필터의 성능이 유한 임펄스 응답 필터 보다 더 우수하다는 특징이 있다.

영상처리에서 필터링은 영상에 있는 공간 주파수 대역을 제거하거나 강조하는 것으로, 영상의 노이즈를 제거하는데 많이 사용된다. 일반적으로 영상은 밝기 값이 급격하게 변하지 않기 때문에 저주파 신호의 형태를 보이는 반면, 노이즈는 무작위 값을 가지는 밝기 값을 가지기 때문에 상대적으로 고주파 신호의 형태를 보이게 된다. 이런 특징을 이용하여 영상 내에 존재하는 고주파 신호 성분을 제거함으로써 영상의 노이즈를 제거할 수 있다. 영상처리에서 쓰이는 필터는 저주파 성분을 남기고 고주파 성분을 제거하는 저주파 통과 필터, 고주파 성분을 남기고 저주파 성분을 제거하는 고주파 통과 필터, 경계선을 검출하는 예지 강화 필터 등이 있다.

이상으로 여러 분야에서 쓰이는 필터링 기법을 설명하였다. 신호처리와 영상처리에서 쓰이는 필터링은 특정 분야에 특화된 기법이므로 보편적인 샘플링 기법을 소개하는 본 보고서에서는 다루지 않는다.

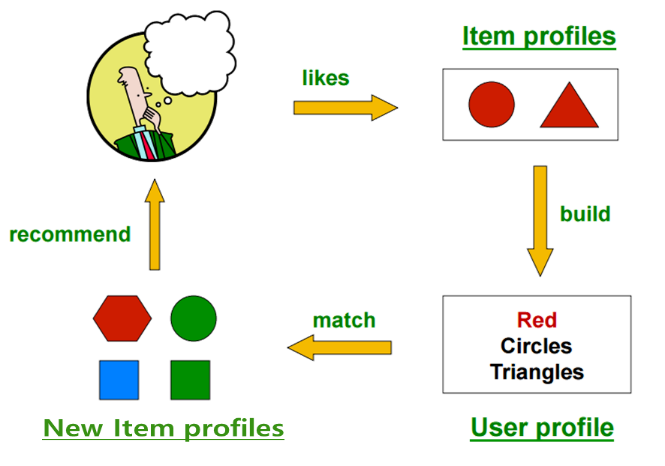
# 필터링 기법

본 절에서는 다양한 필터링 기법에 대해 설명한다. 필터링은 크게 학습 기반 필터링과 비학습 기반 필터링으로 나뉜다. 학습 기반 필터링은 필터링 수행 전 필터 모델을 학습시키는 과정이 필요한 필터링 기법이고, 비학습 기반 필터링은 필터링 수행 전 필터 모델을 위한 별도의 학습과정이 필요하지 않은 필터링 기법이다.

## 학습 기반 필터링

학습 기반 필터링은 필터링 수행 전 필터 모델을 학습시키는 과정이 필요한 필터링 기법으로 데이터 마이닝 분야에서 많이 사용된다. 필터링을 통해 걸러내는 정보를 결정하는 과정이 복잡하기 때문에 필터를 사전에 학습시켜 필터 모델을 생성해야 한다. 예를 들어, 수많은 이메일 중 스팸메일을 걸러내야 하는 경우, 스팸메일과 스팸이 아닌 메일을 분류하고 스팸메일이 가지는 특징과 스팸이 아닌 메일이 가진 특징을 필터 모델에게 학습시켜야 한다. 따라서 필터링 수행 전, 필터 모델 생성을 위한 데이터가 필요하고, 필터 모델 생성 과정이 복잡하고 시간이 많이 소요된다. 필터링 수행 중에도 필터를 통해 추출된 데이터를 사용자가 확인하고 필터링 결과에 대한 피드백을 제공하여 필터를 계속 학습시켜 더 나은 필터링 결과를 추출하도록 할 수도 있다. 다만 이 경우는 사용자가 필터 결과에 대한 피드백을 제공해야 하기 때문에 실시간으로 수행할 수 없다. 학습 기반 필터링 알고리즘으로 내용기반, 학습, 베이지안 필터링이 있다.

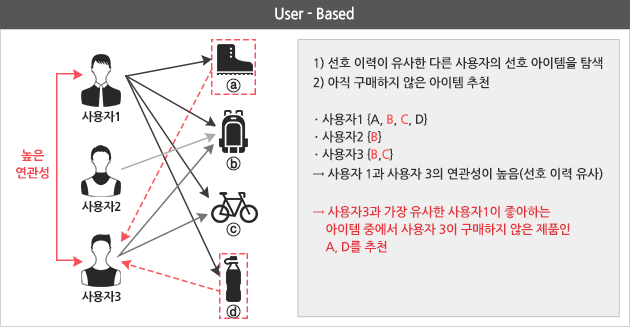
### 내용기반 필터링

내용기반 필터링은 사용자의 프로파일을 이용하여 필터 모델을 만들고, 필터 모델을 바탕으로 사용자에게 적절한 정보를 추천해주는 필터링 기법이다. 내용기반 필터링은 사용자가 과거에 사용했거나 평가했던 상품들로 이루어진 사용자 프로파일과 아이템들 간의 유사도를 측정하고 가장 유사한 아이템의 정보를 제공한다. [그림 1]은 내용기반 필터링을 설명한 그림이다. 사용자가 선호하는 아이템인 빨간색 동그라미, 빨간색 세모의 특성을 추출하여 필터 모델을 만들고, 해당 조건에 부합하는 유사한 아이템을 추출해 사용자에게 추천하는 것이다. 정보검색 주요 모델인 불리안 모델, 벡터공간 모델, 확률모델, 인공신경망 모델, 퍼지집합 모델과 같은 기법을 사용하여 주제에 적절한 텍스트 정보를 찾는데 효과적이다.

[그림 ]. 내용기반 필터링 예시[[1]](#footnote-1).

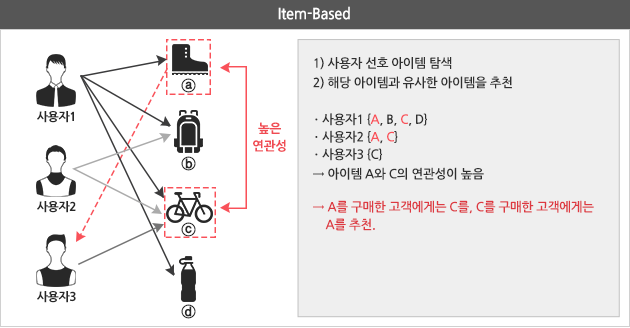
### 협업 필터링

협업 필터링은 많은 사용자로부터 얻은 아이템, 즉 상품에 대한 선호도와 관심 표현을 바탕으로 사용자들의 관심사를 예측하는 필터링 기법이다. 협업 필터링은 크게 사용자 기반(User-based)과 아이템 기반(Item-based) 필터링으로 나뉜다. 사용자 기반 협업 필터링은 선호 이력이 유사한 다른 사용자의 선호 아이템을 탐색하여 아직 구매하지 않은 아이템을 추천하는 방법이다. [그림 2]는 사용자 기반 협업 필터링을 나타낸 그림이다. 사용자1과 사용자3의 선호 이력이 유사하여 두 사용자의 연관성이 높은 것으로 나타났고, 사용자3과 가장 유사한 사용자1이 선호하는 아이템 중에서 사용자3이 구매하지 않은 제품인 A와 D를 사용자3에게 추천한다.



[그림 ]. 사용자 기반 협업 필터링 예시[[2]](#footnote-2).

반면 아이템 기반 협업 필터링은 사용자가 선호하는 과거 구매 혹은 조회 아이템을 기반으로 그 아이템과 선호 연관성이 높은 아이템을 추천하는 방식이다. [그림 3]은 아이템 기반 협업 필터링을 나타낸 그림이다. 사용자들이 선호하는 아이템을 탐색하여 해당 아이템과 유사한 아이템을 추천한다. 그림에서는 아이템 A와 C의 연관성이 높으므로 A를 구매한 고객에게는 C를, C를 구매한 고객에게는 A를 추천한다. 따라서 C를 구매한 사용자3에게 A를 추천한다.



[그림 ]. 아이템 기반 협업 필터링 예시.

사용자 한 명의 프로파일을 분석하는 것이 아니라 여러 사용자의 평가 내용을 기반으로 다른 사용자에게 아이템을 추천하는 점이 내용기반 필터링과의 차이점이다. 협업 필터링은 복잡한 아이템을 어렵게 분석하지 않아도 된다는 장점이 있다. 하지만 아이템 평가내용이 없는 사용자에게 협업 필터링을 사용하기 어렵다는 문제점이 있다. 주로 영화, 도서, VOD, 음악 추천 등에 사용한다.

### 베이지안 필터링

베이지안 필터링은 조건부 확률을 이용하여 각 아이템이 가진 속성에 따라 특정 범주로 분류하는 필터링 방법이다. 예를 들어 설명하면 다음과 같다. 스팸메일을 분류하는 문제에서, 두 단어 과 가 포함된 메일이 스팸메일일 확률은 식 1과 같이 구할 수 있다.

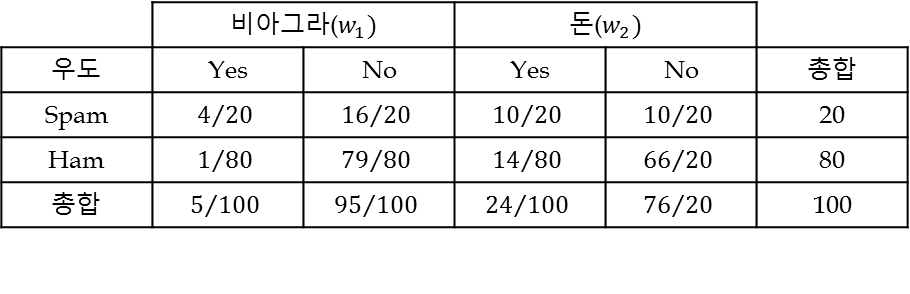
식 (1)

하지만 식 1과 같이 확률을 구할 경우, 과 가 메일에 포함될 사건이 독립적이지 않기 때문에 가능한 모든 중첩된 사건의 확률을 구해야 한다. 따라서 모든 확률을 저장하기 위해 메모리가 많이 필요하고, 계산도 복잡해지는 문제가 나타난다. 따라서 베이지안 필터링은 모든 단어가 나타날 사건을 독립이라고 가정하고, 해당 메일이 스팸메일일 확률은 식 2와 같이 간단하게 구한다. 마찬가지로 해당 메일이 스팸메일이 아닐 확률은 식 3과 같이 구할 수 있다.

식 (2)

식 (3)

[표 ]. 2개 단어의 출현에 대해 100개의 이메일에 대한 우도표.



두 단어 과 가 포함된 메일이 스팸메일일 확률과 스팸메일이 아닐 확률이 [표 1]과 같다면 이 메일이 스팸일 확률은 식 2에 따라 이고 스팸이 아닐 확률은 식 3에 따라 이다(두 식의 분모는 같으므로 계산하지 않았다). 따라서 이 메일이 스팸메일일 확률은 이고, 스팸메일이 아닐 확률은 같은 식에 의해 이다. 따라서 해당 메일을 스팸으로 분류한다.

베이지안 필터링은 모든 속성이 일어날 확률을 독립적이라고 가정하고 조건부 확률을 구한 후, 나온 확률에 따라 특정 범주로 분류한다. 따라서 스팸메일 필터링과 같은 문서 분류, 저자 식별이나 주제 분류, 네트워크에서 이상 행동 검출, 관찰된 증상을 고려한 질병 진단 등 결과의 확률을 추정하기 위해 많은 속성을 고려해야 하는 문제에 적합하다. 베이지안 필터링은 속성 값을 바탕으로 각 범주의 관찰된 확률을 계산하기 위해 훈련 데이터가 필요하다는 특징이 있다. 따라서 필터를 사전에 학습시켜 생성해야하고, 시간이 오래 걸린다는 단점이 있다. 반면 매우 단순한 개념이지만, 좀 더 세련된 알고리즘과 동등한 결과를 내고, 노이즈와 결측치에도 잘 수행한다는 장점이 있다. 하지만 모든 속성은 동등한 중요도를 갖고, 독립적이라는 결함 가정에 의존한다는 단점이 있다.

## 비학습 기반 필터링

비학습 기반 필터링은 필터링 수행 전 필터 모델을 위한 별도의 학습과정이 필요하지 않은 필터링 기법이다. 필터링을 통해 걸러내는 정보를 결정하는 과정이 비교적 간단해 필터 생성 과정이 간단하고 소요 시간이 적다. 가장 기본적인 필터링 기법인 SQL 구문을 이용한 질의 필터링과, 블룸 필터링, 칼만 필터링이 이에 속한다.

### 질의 필터링

질의 필터링은 가장 많이 쓰이는 필터링으로, 정형화된 데이터를 대상으로 SQL[[3]](#footnote-3) 질의를 사용하여 필터 조건에 부합하는 데이터를 추출하는 방법이다. 따라서 질의 필터링은 다른 필터링과 다르게 별도의 필터 모델 없이 필터링 알고리즘을 생성할 수 있다. 처리 속도가 빨라 데이터 검색 응용에 가장 많이 사용되고, 실시간으로 처리할 수 있고, 텍스트, 숫자 등 다양한 데이터에 적용할 수 있다. 질의를 어떻게 구성하느냐에 따라 수행 속도나 성능에 차이가 발생할 수 있기 때문에 질의를 최적화하는 것이 중요하다.

### 블룸 필터링

블룸 필터링은 해시함수를 사용하여 데이터가 사용자가 미리 정의한 집합에 속하는지 여부를 빠르게 검사하는데 사용되는 필터링 기법이다. 블룸 필터는 m개의 비트로 이루어진 자료구조이며, 사용자는 집합에 속한 데이터에 k개의 해시함수를 적용한 결과값에 해당하는 블룸 필터 인덱스 값을 1로 표시하여 블룸 필터를 만든다. 그리고 각 데이터에 해시함수를 적용하여 결과값을 블룸 필터 인덱스의 값과 비교하여 해당 데이터가 집합 내에 속하는지 여부를 판단한다. [그림 5]는 블룸 필터링의 동작과정을 보여준다. 블룸 필터의 크기 m은 18, 해시 함수의 개수 k는 3이다. 데이터 집합에 x, y, z라는 세 개의 데이터가 있을 때, x를 해시 함수에 적용한 결과는 2, 6, 14이고 y는 5, 12, 17, z는 4, 6, 12이다. 해시 함수에 적용한 결과값에 해당하는 블룸 필터의 인덱스 값을 1로 바꾼다. 블룸 필터를 생성한 후, 새로운 데이터 x가 들어왔을 때, 데이터 집합에 속하는지 여부를 판단하기 위해 블룸 필터에 적용해보면 해시 함수의 결과로 2, 6, 14를 얻게 되고, 모든 값이 1이므로 x는 데이터 집합에 포함된 데이터라는 결과를 얻게 된다.

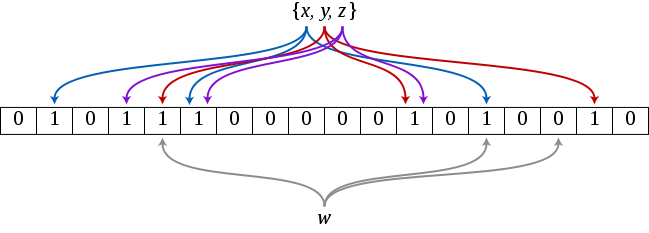


그림 . 블룸 필터링 예시.

블룸 필터링은 어떤 데이터가 집합에 속한다고 판단된 경우 실제로 데이터가 집합에 속하지 않는 긍정 오류(false positive)가 발생하는 것은 가능하지만, 반대로 데이터가 집합에 속하지 않는 것으로 판단되었으나 실제로는 데이터가 집합에 속하는 부정 오류(false negative)는 불가능하다는 특징이 있다. 또한 집합에 원소를 추가하는 것은 가능하나 집합에서 원소를 삭제하는 것은 불가능하다는 특징이 있다. 집합 내 데이터의 숫자가 증가할수록 긍정 오류 발생 확률도 증가하기 때문에 블룸 필터의 크기(m)와 해시 함수의 개수(k)를 최적화하는 것이 중요하다. 데이터가 집합에 속하는지 여부를 판단하기 위해 질의 필터링을 사용할 경우, 매번 데이터 집합을 확인해야 하기 때문에 수행시간이 오래 걸리는 반면, 블룸 필터링을 사용할 경우 해시 함수의 값만 확인하기 때문에 수행시간이 매우 빠르다는 장점이 있다. 따라서 데이터 포함여부 판단에 대한 디스크 I/O를 줄여야 할 때 블룸 필터를 사용하면 시스템의 성능과 처리량을 향상시킬 수 있다.

### 칼만 필터링

칼만 필터링은 유입되는 데이터에 포함된 노이즈를 제거하는 필터링 기법이다. 칼만 필터는 물체의 측정값에 확률적인 오차가 포함되어 있고, 특정 시점에서 물체의 상태가 이전 시점의 상태와 선형적인 관계를 가지고 있는 경우 적용이 가능하다. 예를 들어, 레이더 추적의 경우 특정 물체의 위치, 속도, 가속도 등을 측정할 수 있지만 이 측정값에 오차가 포함되어 있을 수 있다. 이 경우, 연속적으로 측정하는 값들을 칼만 필터를 이용해 해당 물체의 위치를 추정할 수 있다.

칼만 필터는 재귀적으로 동작한다. 즉, 칼만 필터는 바로 이전 시간에 추정한 값을 토대로 현재의 값을 추정하며, 바로 이전 시간 외의 측정값이나 추정값은 사용하지 않는다. 각 추정 계산은 예측(Prediction)단계와 보정(Update) 두 단계로 이루어진다. 예측 단계에서는 이전 시간에 추정된 상태에 대해, 그 상태에서 사용자 입력을 가했을 때 예상되는 상태를 계산한다. 보정 단계에서는 앞서 계산된 예측 상태와 실제로 측정된 상태를 토대로 정확한 상태를 계산한다. 칼만 필터는 시간에 따라 진행한 측정을 기반으로 하기 때문에 데이터가 끊임없이 유입되는 스트림 환경에도 적용할 수 있다는 장점이 있다.

# 기술 동향

## 관련 제품군

에스퍼(Esper)는 복잡한 이벤트 처리(CEP: complex event processing)를 위한 대표적인 DSMS 오픈소스이다. 이 때 이벤트는 데이터 스트림이며 CEP는 여러 데이터 스트림 소스로부터 발생한 대용량의 데이터를 실시간 처리하기 위해 스트림 일부를 범위로 하여 질의를 수행하며 이 범위를 옮겨가며 반복 수행한다. 이 때 질의 처리를 위해 스트림 도메인에 특화된 언어인 ESP(event processing language)를 사용한다. EPL은 기존 SQL과 유사한 질의 언어로, 질의 필터링을 구현하여 사용할 수 있다. 또한 베이지안 필터링, 블룸 필터링 등 다른 필터링 알고리즘도 EPL로 구현하여 사용이 가능하다.

오라클 Complex Event Processing은 오라클사에서 개발한 CEP이다. CEP엔진이 경량 서버이고, SQL의 확장판인 오라클 CQL(continuous query language)을 사용하여 사용이 용이하다는 특징이 있다. 오라클 CEP에서는 질의 필터링을 제공한다.

MATLAB(매트랩)은 MathWorks사에서 개발한 수치 해석 및 프로그래밍 환경을 제공하는 공학용 소프트웨어이다. 행렬을 기반으로 한 계산 기능을 지원하며, 함수나 데이터를 그림으로 그리는 기능 및 프로그래밍을 통한 알고리즘 구현 등을 제공한다. MATLAB은 수치 계산이 필요한 과학 및 공학 분야에서 다양하게 사용된다. MATLAB은 칼만 필터의 적용과 시각화를 제공한다.

아파치 카산드라(Apache Cassandra)는 구글의 BigTable 컬럼 기반 데이터 모델과 페이스북에서 만든 Dynamo의 분산 모델을 기반으로 제작되어 아파치 오픈소스로 공개된 분산 데이터베이스이다. 기존 관계형 데이터베이스와 다르게 SQL을 사용하지 않은 NoSQL의 제품중 하나이며 대용량 데이터 트랜잭션에 대해서 고성능 처리가 가능한 시스템이다. 카산드라는 주로 데이터 분석에 사용하는 데이터베이스이기 때문에 블룸 필터를 제공한다. 또한 하둡, BigTable 등 다양한 데이터베이스 저장소는 물론 Squid와 같은 웹 캐시 시스템에서도 블룸 필터를 제공한다.

디지털 콘텐츠 저작권 보호 기술업체 뮤레카가 내용기반 필터링으로 유해 동영상 차단 서비스를 제공한다. 유해 동영상 파일에서 오디오, 비디오 데이터를 분석하여 유통되는 동영상 파일과 유사한 파일을 차단하는 것이다.

아파치 머하웃은 2011년에 출시된 분산처리가 가능하고 확장성을 가진 기계학습 라이브러리이다. 맵리듀스를 이용하는 아파치 하둡 위에서 작동하며 추천, 클러스터링, 분류 기능을 제공한다. 추천 기능에서는 협업 필터링을 핵심으로 제공한다.

## 관련 기술연구

김병만 외 4명은 내용기반 필터링과 협력 필터링의 결합 기법을 제안하였다.[[4]](#footnote-4) 내용 기반 필터링에서는 내용에 기반을 두어 정보를 추출하는 반면 협력 필터링은 다른 사람의 의견을 이용하게 된다. 기존 협력 필터링 방법의 문제점을 해결하기 위한 방법의 일환으로 내용 기반 필터링과 협력 필터링 기법을 보다 유기적으로 결합시키는 연구를 수행하였다.

카운팅 필터(counting filter)는 블룸 필터에서 필터를 재생성하지 않고도 원소의 삭제가 가능하도록 변형된 필터이다. 블룸 필터에서 배열(필터) 내 각 버켓의 크기가 1비트였던 것을 카운팅 필터에서는 n비트로 확장하여 카운터로 사용한다. 원소를 추가하면 버켓의 값을 1씩 증가시키고, 원소의 삭제 시에는 버켓의 값을 1씩 감소시키는 방법으로 필터를 유지한다.

1. Cambridge University Press, “Mining of Massive Datasets,” Cambridge University Press, 2015. [↑](#footnote-ref-1)
2. LG CNS 블로그, http://blog.lgcns.com/940 [↑](#footnote-ref-2)
3. 관계형 데이터베이스 관리 시스템(RDBMS)의 데이터를 관리하기 위해 설계된 특수 목적의 프로그래밍 언어 [↑](#footnote-ref-3)
4. 김병만, et al. "추천시스템을 위한 내용기반 필터링과 협력필터링의 새로운 결합 기법." *정보과학회논문지: 소프트웨어 및 응용* 31.3 (2004): 332-342. [↑](#footnote-ref-4)