

01

# 비정형(텍스트) 마이닝의 개요

BIG

DATA

# 1. 비정형(텍스트) 마이닝의 개요



## 가. 비정형 데이터의 이해

### 비정형 데이터

- 정형화되지 않은 데이터
- 미리 정의된 데이터 모델(구조)을 가지고 있지 않은 데이터

예

아주 많은 양의 데이터를 가지고, 구조와 형태가  
다르고 정형화되지 않은 문서, 영상, 음성 등

» 책, 저널, 문서, 메타데이터, 건강 기록, 오디오, 비디오,  
아날로그 데이터, 이미지, 파일, 이메일 메시지,  
웹페이지, 워드프로세스 문서 등

# 1. 비정형(텍스트) 마이닝의 개요



## 가. 비정형 데이터의 이해

- 비정형 데이터의 유형은 크게 텍스트, 이미지, 음성과 영상, 로그 파일로 구분

텍스트  
(Text)

» 트위터, 페이스북 등 소셜 미디어에서의 실시간 대화, 온라인 모바일을 통한 SMS, 이메일 메시지, 블로그, 커뮤니티에서의 게시물, 전문정보, 뉴스 기사 등

이미지

» 인터넷 매체에서 업로딩되는 모든 사진, 그림 등

# 1. 비정형(텍스트) 마이닝의 개요



## 가. 비정형 데이터의 이해

- 비정형 데이터의 유형은 크게 텍스트, 이미지, 음성과 영상, 로그 파일로 구분

### 음성과 영상

» 음악파일, 유튜브 등과 같은 동영상 전문 웹사이트가 제공하는 영상, UCC, 뉴스 동영상, 애니메이션 등

### 로그 파일 (Log File)

» 웹로그, 인터넷 검색 인덱싱, 페이지 뷰 인덱싱, 웹상에서 모든 흔적들의 데이터 파일

# 1. 비정형(텍스트) 마이닝의 개요



## 가. 비정형 데이터의 이해

### ● 비정형 데이터는 불규칙 정도에 따라 반정형 데이터로 구분

- » 반정형 데이터는 관계형 데이터베이스나 다른 형태의 데이터 테이블로 조직된 데이터 모델의 정형적 구조를 따르지 않지만, 어의적 요소를 분리시키고 데이터 내의 레코드와 필드의 계층 구조가 있게 하는 태그(Tag)나 다른 마커(Marker)를 포함하는 정형 데이터
- » 반정형 데이터는 같은 클래스에 속하는 속성들을 순서에 상관없이 서로 묶을 수 있고 다른 속성을 포함 가능
- » 최근 객체지향 데이터베이스에서 반정형 데이터가 많이 등장
- » 마크업(Markup) 언어, 이메일, EDI(Electronic Data Interchange) 등

# 1. 비정형(텍스트) 마이닝의 개요



## 가. 비정형 데이터의 이해

### XML(eXtensible Markup Language)

- » 데이터 구조와 데이터를 스스로 기술하는 수단을 가진 비교적 최근에 나온 마크업 언어
- » 이전에는 기능적 수준에서 구조적 엄격함이 못 미치는 인상을 가지게 하여 비정형 형태로 보았으나, 실제로는 **아주 엄격한 요소 구조와 데이터 형식과 인간 중심 흐름 및 계층구조를 가능하게 하는 '유연성 있는 구조'로 언급**

# 1. 비정형(텍스트) 마이닝의 개요



## 나. 비정형 데이터 분석과 마이닝

빅데이터 환경에서 거의 80% 이상이 비정형 데이터이므로,  
빅데이터의 데이터 마이닝은 **비정형 데이터 마이닝에 초점**

통계 기반의  
데이터  
분석도구 사용

OLAP 분석을 통해  
다양한 관점으로  
조명하여  
의미 있게 해석

데이터 사이에  
숨겨진 관계, 패턴,  
경향 등을 추출

# 1. 비정형(텍스트) 마이닝의 개요



## 나. 비정형 데이터 분석과 마이닝

- 비정형 데이터의 내용 파악과 비정형 데이터 속 패턴(Pattern) 발견을 위해 데이터 마이닝, 텍스트 분석, 비표준 텍스트 분석 등과 같은 다양한 기법을 사용
- 비정형 데이터를 정련 과정을 통해 정형데이터로 만든 후, 분류, 군집화, 회귀분석, 요약, 이상감지 분석 등의 데이터 마이닝을 통해 의미 있는 정보를 발굴



# 1. 비정형(텍스트) 마이닝의 개요



## 나. 비정형 데이터 분석과 마이닝

### 텍스트를 정형화하는 방법

- » 주요 단어 등의 추출 등 정제 과정을 거쳐 정형화된 데이터 구조로 변환하는 것이 가장 일반적인 방법
- » 메타데이터(Meta Data)로 직접 태그(Tag)
- » 고도의 텍스트 마이닝 기반 정형화를 위해 텍스트 속 단어와 스피치(Speech)의 한 부분이 대응되게 태그

# 1. 비정형(텍스트) 마이닝의 개요



## 나. 비정형 데이터 분석과 마이닝

- 정제된 데이터베이스를 기반으로 일정한 기준이 적용된  
상식적 범위에서 부분적인 데이터를 다루는  
정형 데이터 마이닝의 한계를 뛰어넘는 기법 존재



02

# 텍스트 마이닝

BIG

DATA

## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념

## 텍스트 마이닝 (Text Mining)

인간의 언어로 이루어진 비정형 텍스트 데이터들을  
자연어 처리(Natural Language Processing)방식을  
이용하여 대규모 문서에서 정보 추출, 연계성 파악,  
분류 및 군집화, 요약 등을 통해  
데이터에 숨겨진 의미를 발견하는 기법

## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념

## 텍스트 마이닝 (Text Mining)

- 기존 통계분석이나 데이터마이닝을 적용하기에 부적합한 데이터를 다룸
- 텍스트 데이터 마이닝 (Text Data Mining), 텍스트 분석 (Text Analytics), 텍스트 데이터베이스로부터 지식발견 (Knowledge Discovery in Textual Database), 문서 마이닝 (Document Mining) 등으로 호칭

## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념

- 텍스트 마이닝은 대규모의 텍스트에서 고품질 정보를 도출
  - » 고품질 정보는 통계적인 패턴 학습 등의 수단을 통해 패턴과 추세를 파악함으로써 도출
- 텍스트 마이닝은 일반적으로 입력 텍스트를 정형화한 다음, 정형화 데이터 내에서 패턴을 추출하고 난 후, 출력을 평가하고 번역하는 과정을 포함
  - » 정형화는 입력 텍스트를 파싱(Parsing)할 때 추출되는 언어적 특징은 추가시키고 그 이외의 것들은 제거하면서 데이터베이스와 같은 정형화된 구조 속에 삽입하는 것
- 그리고 고품질 정보는 보통 새롭고 적절하며 관심을 끄는 데이터들의 집합으로서 어떤 목적과 관련하여 의미 있는 정보

## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념

- 텍스트 데이터마이닝 또는 텍스트 분석

» 정보 검색, 단어 빈도 분포를 연구하는 어휘 분석, 패턴 인식, 태그 및 주석, 정보 추출, 링크 및 연결 분석을 내포하는 데이터마이닝, 시각화, 예측 분석 등이 필요

## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념

- 정보 검색 기법(information retrieval)
- 자연어 처리(natural language processing) 기법
  - » 의사소통언어처리기법, 구어처리기법
- 특징 추출(feature selection) 기법
- 텍스트 범주화(text categorization) 기법
- 군집화(Clustering) 기법
- 연결분석(text link analysis) 등의 기법



## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념

“

여러 가지 종류의 텍스트 데이터로부터  
지식을 발견하는 과정

”

## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념

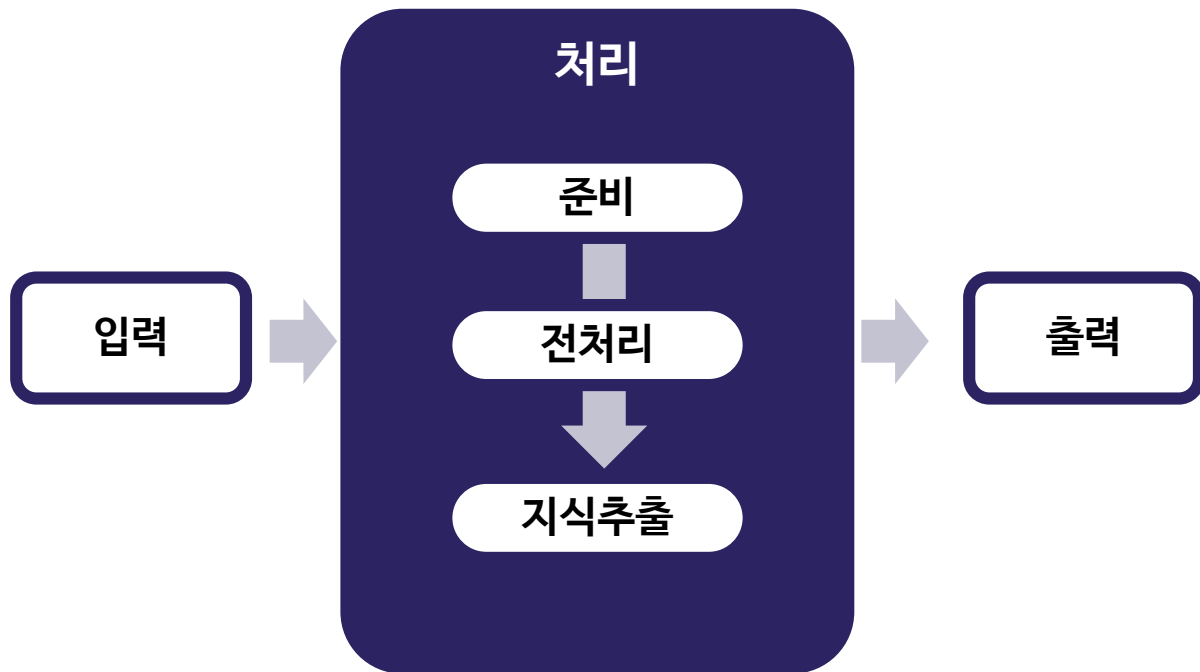
“여러 가지 종류의 텍스트 데이터로부터  
지식을 발견하는 과정”

텍스트 마이닝의 목적은 비정형  
데이터나 정형 데이터, 반정형  
데이터를 처리하여 의사결정을  
위해 필요한 **고차원적이고 의미  
있는 정보나 지식을 추출**하는 것

## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념



## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념



#### 준비

- 입력되는 여러 가지 텍스트 문서의 데이터들을 **문제 범위에 적절한 것으로 확립**

## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념



#### 준비

- ➔ 입력되는 여러 가지 텍스트 문서의 데이터들을 **문제 범위에 적절한** 것으로 확립

#### 전처리

- ➔ 조직화된 텍스트들을 **정형화**된 표현 양식으로 만듦

## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념



#### 준비

- ➔ 입력되는 여러 가지 텍스트 문서의 데이터들을 **문제 범위에 적절한** 것으로 확립

#### 전처리

- ➔ 조직화된 텍스트들을 **정형화**된 표현 양식으로 만듦

#### 지식추출

- ➔ **정형 데이터**에서 의미 있는 패턴이나 관계와 같은 **지식 발견**
- ➔ 분류, 클러스터링, 개념 및 개체 추출, 세분화된 분류 체계의 생간, 심리 분석, 문서 요약, 개체 관계 모델링

## 2. 텍스트 마이닝



### 가. 텍스트 마이닝의 개념

- 전자·소매유통업종과 금융업종을 중심으로 빅데이터 기반  
고객의 소리 분석 사례 증가
- 인터넷과 모바일 등이 활성화되면서 고객의 소리 창구가  
확대돼 접수 민원 급증



“

”

텍스트마이닝과 데이터마이닝으로  
고객의 소리 분석 기술(음성인식기술, 텍스트분석 기술)이  
급격히 발전해 고객의 소리에 대응

03

# 웹 마이닝

BIG

DATA



### 3. 웹 마이닝



#### 가. 웹 마이닝의 개념

“

데이터마이닝 기술의 응용분야로서  
인터넷을 통해 웹 서비스를 이용하면서  
웹에서 패턴을 발견 하는 것

”



- 데이터의 속성이 반정형이거나 비정형이고,  
링크(Link) 구조를 가지고 있기 때문에  
전통적인 데이터마이닝 기술에 추가적인 분석기법이 필요

### 3. 웹 마이닝



#### 가. 웹 마이닝의 개념

##### 웹콘텐츠 마이닝

웹 페이지에서 유용한 데이터, 정보,  
지식을 마이닝하고 추출하고  
통합하는 것

##### 웹 사용 마이닝

- » 웹 사이의 연결 분석
- » 웹 사이트의 노드와 연결 구조를  
분석하기 위해 그래프 이론을  
사용하는 과정

04

# 오피니언 마이닝

BIG

DATA

## 4. 오피니언 마이닝



### 가. 오피니언 마이닝의 개념

“

어떤 사안이나 인물, 이슈, 이벤트 등과  
관련 원천 데이터에서 **의견이나 평가, 태도, 감정 등과  
같은 주관적인 정보를 식별하고 추출하는 것**

”

## 4. 오피니언 마이닝



### 가. 오피니언 마이닝의 개념

- 오피니언 분석, 평판 분석, 정서 분석
- 오피니언 분석의 기본적인 작업  
문서, 문장, 특징, 관점 수준에서 표현된 견해가  
긍정적인지, 부정적인지, 중립적인지, 진보적인지 주어진  
텍스트의 특성을 분류하는 것

## 4. 오피니언 마이닝



### 가. 오피니언 마이닝의 개념

- 온라인 쇼핑몰에서의 잠재 구매자의 상품평 검색효율을 높이기 위해 **상품평 데이터에 순위를 결정**하는데 이용
- 영화관람 후기를 요약하고 **긍정/부정을 평가**하는데 이용
- 법률 분야의 블로그를 대상으로 오피니언 마이닝을 이용해 **고객의 반응이나 법률적 이슈에 대한 모니터링**에 이용
- 소셜 미디어에서 나타나는 **오피니언들을 조기 감지**하여 기업의 위기 상황을 인지하고 위기에 대응할 수 있는 **위기관리 모델의 핵심 정보**로 활용
- **오피니언**을 경제적 관점에서 정량화하여 **금액으로 환산**
- SNS에서 실시간으로 핫 토픽을 추출하고 오피니언의 흐름을 분석하여 **이벤트, 마케팅, 트렌드 분석 등 다양한 활용**
- 트위터에서 감지되는 시장 분위기를 이용하여 **주가의 흐름 예측**

# 4. 오피니언 마이닝

## 가. 오피니언 마이닝의 개념

### ✓ 오피니언 마이닝으로 본 원자력 여론지수 추이



빅데이터  
기반의  
여론분석의  
장점

- » 기존의 설문조사 방법보다 **저비용으로** 수행할 수 있음
- » **촘촘한 시간 간격**으로 여론의 모니터링이 가능함

05

# 소셜 데이터 마이닝





# 5. 소셜 데이터 마이닝



## 가. 소셜 데이터 마이닝의 개념

소셜 네트워크 서비스  
(SNS, Social Network Service)

개인의 생각이나 의견, 비전이나 가치 등을 디지털 콘텐츠(Digital Content) 형태로 공유하거나 교환할 수 있도록 사회적 관계를 맺는 쌍방관계를 갖는 커뮤니티 서비스(Community Services)

- 예 ) 페이스 북(Facebook), 구글 플러스(Google+) 등
- 트위터(Twitter), 마이크로 블로그(Micro Blog)도 일방적 관계를 서로 맺게 되면 쌍방 관계가 되므로 소셜 네트워크 서비스로 간주함

# 5. 소셜 데이터 마이닝



## 가. 소셜 데이터 마이닝의 개념

### 국외

- » 사용자의 로그, 관심사, 정보를 분석하여 **트렌드 감지**
- » 브랜드를 모니터링, 감성분석, 마케팅 등을 제공할 수 있는 **기반환경 서비스**
- » 구글은 구글 트렌드를 통해 실시간 핫이슈 검색, 실시간 순위 및 순위차트 제공, 카테고리별 이슈 분류, 기간별 설정 및 검색 기능을 제공

### 국내

- » 소셜 미디어 분석에서 언어 분석 기술을 적용해 **검색어에 대한 기간별 소셜 모니터링, 연관어 탐색, 감성 분석 서비스** 등을 제공
- » **다음소프트**의 경우 소셜 매트릭스 서비스 제공

# 5. 소셜 데이터 마이닝



## 가. 소셜 데이터 마이닝의 개념

### 국외

- » 사용자의 로그, 관심사, 정보를 분석하여 트렌드 감지
- » 브랜드를 모니터링, 감성분석, 마케팅 등을 제공할 수 있는 기반환경 서비스
- » 구글은 구글 트렌드를 통해 실시간 핫이슈 검색, 실시간 순위 및 순위차트 제공, 카테고리별 이슈 분류, 기간별 설정 및 검색 기능을 제공

### 국내

- » 소셜 미디어 분석에서 언어 분석 기술을 적용해 검색어에 대한 기간별 소셜 모니터링, 연관어 탐색, 감성 분석 서비스 등을 제공
- » 다음소프트의 경우 **소셜 매트릭스** 서비스 제공

소셜 미디어 정보를 **모니터링** 하여  
주별 급증한 **키워드** 순위를 제공,  
**연관어** 기반 탐색건수 제공,  
**감성** 기반 **연관어** 등으로 제공

## 5. 소셜 데이터 마이닝



### 가. 소셜 데이터 마이닝의 개념

- 개인의 일상 정보가 연결된 사회적 관계망을 분석하는 것이 필요한데 그것이 소셜 네트워크 애널리틱스(Social Network Analytics : SNA)임
- 노드와 링크로 구성되는 네트워크 이론에 의해 사회적 관계(우정, 연대감, 조직력, 성향)를 보여주는 것

# 5. 소셜 데이터 마이닝



## 가. 소셜 데이터 마이닝의 개념

- 개인의 일상 정보가 연결된 사회적 관계망을 분석하는 것이 필요한데 그것이 소셜 네트워크 애널리틱스(Social Network Analytics : SNA)임
- 노드와 링크로 구성되는 네트워크 이론에 의해 사회적 관계(연대감, 조직력, 성향)를 보여주는 것
  - » 노드(node) : 점, 행위자
  - » 링크(link) : 선, 각 노드의 관계

# 5. 소셜 데이터 마이닝



## 가. 소셜 데이터 마이닝의 개념

- 개인의 일상 정보가 연결된 사회적 관계망을 분석하는 것이 필요한데 그것이 소셜 네트워크 애널리틱스(Social Network Analytics : SNA)임
- 노드와 링크로 구성되는 네트워크 이론에 의해 사회적 관계(우정, 연대감, 조직력, 성향)를 보여주는 것
- 소셜 네트워크 연결구조 및 연결강도 등을 바탕으로 노드의 복잡도를 측정하여, 소셜 네트워크 상에서 연결의 중심 역할을 하는 영향력이 있는 행위자를 파악  
➔ 파악하고 관리하는 것이 마케팅 관점에서 매우 중요!

# 5. 소셜 데이터 마이닝



## 가. 소셜 데이터 마이닝의 개념

- 보험사기인지시스템(IFAS)

소셜 네트워크 분석(SNA) 기법을 활용해 보험설계사와 피보험자, 병원과의 관계를 분석하고 보험사기 혐의가 짙은 패턴을 가려내는 시스템

- » IFAS상의 보험계약 및 보험금 지급 데이터를 활용하여 계약자, 설계사, 병원 등 개별 혐의자들 간의 상호연관성을 분석하고, 보험사기 혐의 그룹을 시스템적으로 추출하는 기법
- » IFAS는 SNA 기법을 통해 보험사기 혐의 가능성을 계량화하여 보험설계사와 병원을 혐의그룹 형태로 분류, 그 연계도 (혐의그룹 모델)를 추출하고, 선정된 혐의 그룹 모델에 대해 시각화하고 그 특성 분석

# 5. 소셜 데이터 마이닝

## 가. 소셜 데이터 마이닝의 개념

### ☑ 브로커 개입형 보험사기 구조

