

메타 블로그 사이트에서의 선형 SVM 기반 텍스트 분류 방법론 연구

이상진, 이승준, 박종현

서울대학교 산업공학과

내용

- 서론
- SVM 기반의 다계층 분류
- 실험 및 결과
- 결론

서론 (1/2)

- 배경

- 메타 블로그 사이트

- 실시간으로 생성되는 블로그 콘텐츠를 수집하여 사용자에게 제공

- 문제 정의

- 수집된 신규 포스트를 기존 카테고리 체계에 맞도록 자동 분류

- 미분류 포스트 대상

- 온라인 분류 가능

- 데이터

- 전체 포스트 개수:
69,300,000

- 기분류된 포스트 개수:
27,360,000(39%)

: 미분류 포스트

SPOTLIGHT

안드로이드 넥서스원 OS 2.2.1 업그레이드 안내

by Bluesky

전체

문화

스포츠

연예

IT/과학

정치/사회

edit >

실시간 인기글

최신 업데이트 글

오늘의 인기글

주간 인기글

월간 인기글

15

mixup

윤은혜 자선행사 / 윤은혜 자선행사 드레스, 글래머 몸매...
후지짱의 연예계 이야... | 4시간전 등록 | [더보기](#)

윤은혜 자선행사 / 윤은혜 자선행사 드레스, 숨겨왔던 글래머 몸매 공개! 인기 아이돌 그룹 '베이비복스' 출신 연기자 윤은혜씨가유방암 기금마련 자선행사에 참여해 평소에 볼 수 ...



15

mixup

K리그 신생 구단 광주의 이름 논란
강철지크와 지니어스의... | 1시간전 등록 | [축구](#) | [더보기](#)

내년부터 K리그에 참가하는 광주 시민축구단이 지난 9월에 구단 이름 공모를 했습니다. 설문 조사 결과 선정된 이름은 '광주 레이머스(Rayers)'였지요. 그런데 'ray'라는 단어의 뜻이...



14

mixup

쉽게 이해하는 문대표와 정부회장의 트위터 논쟁
강철지크와 지니어스의... | 1시간전 등록 | [더보기](#)

[등장 인물 프로필] http://twitter.com/green_mun나우콤 대표 문용식 트위터 <http://twitter.com/yjchung68>신세계 부회장 정용진 트위터사건의 발단은 정용진 부회장이 <직원 상...>



16

mixup

'대물' 이수경, 정치권 뒤흔들 뇌관?
꽃단비의 방송 연예 이... | 3시간전 등록 | [연예일반](#) | [더보기](#)

정치드라마 '대물'이 서해람 캐릭터 변화 문제로 여전히 시끄럽네요. 메인 작가와 PD교체로 여전히 고현정의 심기가 불편하가 봐요. 왜 안그렇겠어요? 4회까지만 해도 미실을 능가...



서론 (2/2)

- **포스트 개수**
 - 6,930만건 (일일 30만 신규 포스트 생성)
- **카테고리 체계: 2단계**
 - 1단계 대분류: 8개
 - IT, 정치,경제,환경,스포츠,생활,인물,문화
 - 2단계 소분류: 54개
 - 인터넷, 모바일,금융 등
- **학습에 사용된 포스트**
 - 사람이 직접 카테고리를 입력한(기분류된) 포스트
 - 80%-모델 학습, 20%-테스트에 사용

	원본 데이터	실험데이터
피드수	176,294	8,257
포스트 수	69,300,000	765,000
카테고리별 평균 포스트수	1,280,000	95,000
피드별 평균 포스트수	202.47	92.65
카테고리수(대분류)	8	8
카테고리수(소분류)	54	54

SVM 기반의 다계층 분류 (1/3)

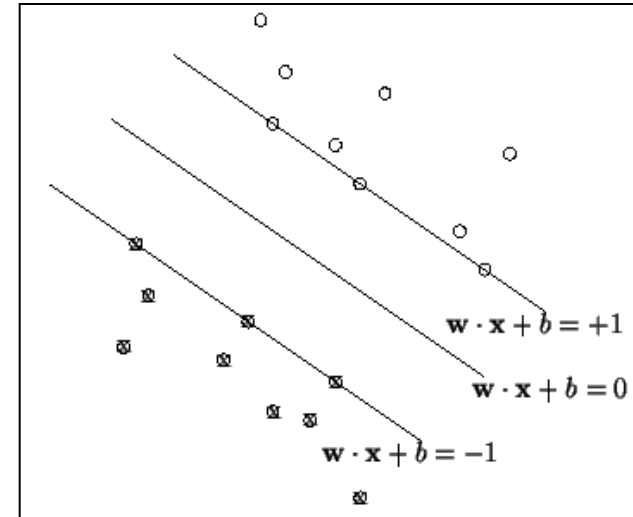
- SVM 개요
 - 커널기반의 2개 집단을 분리하는 기법
 - 최적의 초평면 계산
- 포스트(문서) 표현
 - 벡터 공간 모델(TF-IDF 벡터)로 문서를 표현

$$(tfidf)_{i,j} = tf_{i,j} \times idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \frac{|D|}{|\{j: t_i \in d_j\}|}$$
$$d_j = \{tfidf_{i,j} : t_i \in d_j\}$$

- 선형 커널 사용

$$K(x_i, x_j) = x_i^T x_j$$

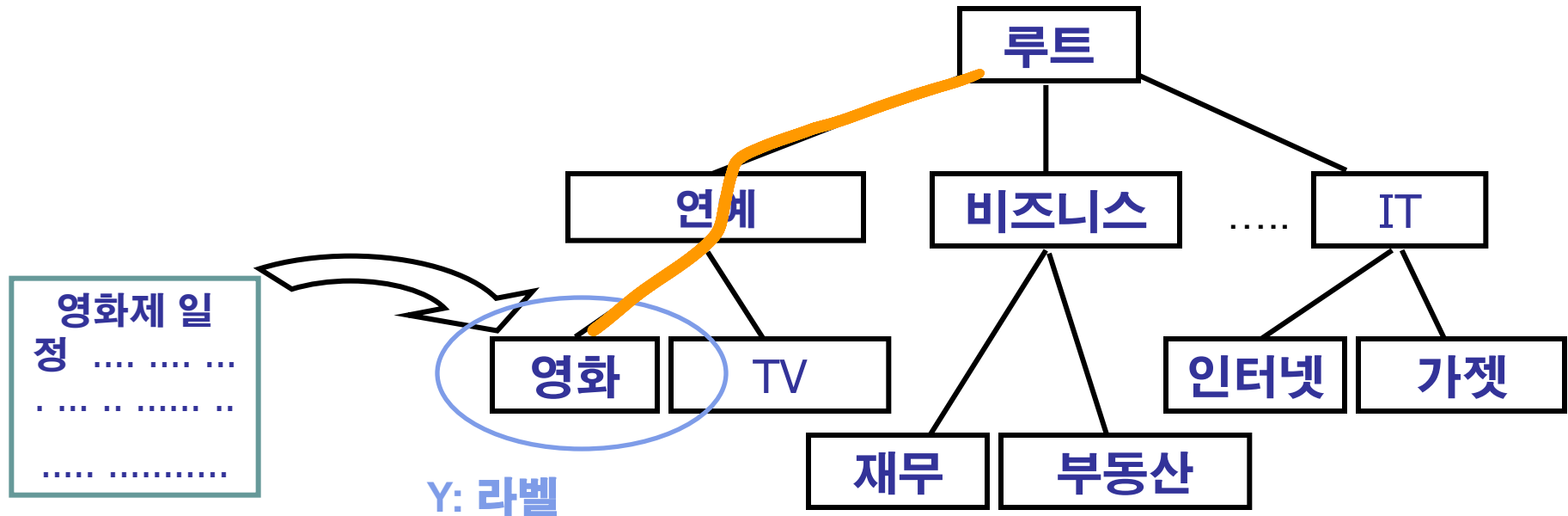
- 대용량 학습 데이터를 대상
 - 비선형 커널보다 선형 커널의 계산상 성능 우수
- 텍스트 분류 문제에서 선형 커널의 사용에 대한 기존 연구 결과



$$\begin{aligned} & \underset{w, b, \zeta_i}{\text{minimize}} && \frac{1}{2} w^T w + C(\sum_{i=1}^N \zeta_i) \\ & \text{subject to} && y_i (w^T x_i - b) + \zeta_i - 1 \geq 0, \quad 1 \leq i \leq N \\ & && \zeta_i \geq 0, \quad 1 \leq i \leq N \end{aligned}$$

SVM 기반의 다계층 분류 (2/3)

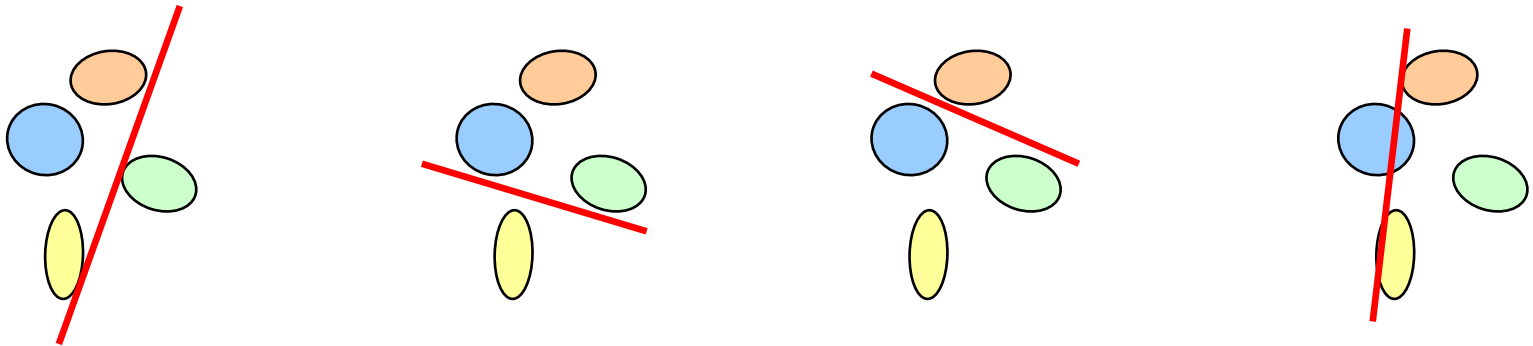
- 다계층 SVM(Hierarchical SVM)
 - 각 계층별로 해당 SVM 모델을 생성
 - Cf) 최하위 계층(Leaf node)에 대해서만 모델 생성하는 방법
 - 모델 생성
 - 1단계-1개 모델
 - 2단계-8개 모델



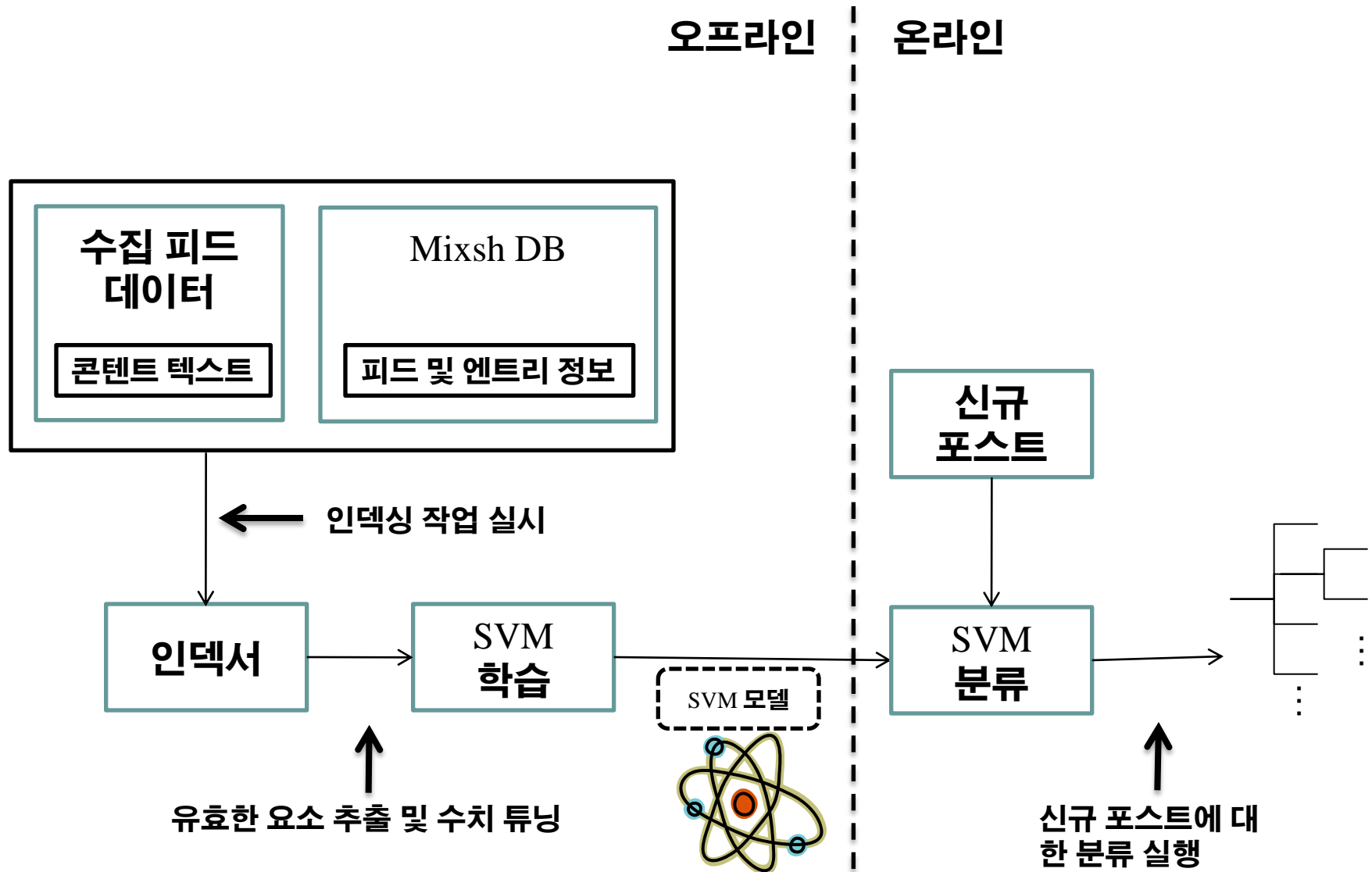
X: 포스트

SVM 기반의 다계층 분류 (3/3)

- **멀티 라벨 SVM(Multi-label SVM)**
 - **포스트는 여러 개의 카테고리에 포함될 수 있음**
 - E.g.음악 장르를 설명하는 포스트: 교육 및 음악 카테고리
 - **올바른 카테고리 선택을 위해 one vs. rest 방법 채택**
 - **각 카테고리 별로 SVM 분류기를 이용하여, 특정 기준치(threshold) 초과하는 카테고리를 선택**



실험 개요 (1/3)



실험 개요 (2/3)

신규 포스트 (정치 & 유머)



SVM 분류기

1단계
분류기

IT 사회 비즈니스 연예 스포츠
(-0.1, **0.4**, -0.5, **0.3**, -0.8, ...)

2단계 분류기
(사회)

2단계 분류기
(연예)

정치 경제 환경
(**0.6**, -0.4, -0.2, ...)

TV 영화 유머
(-0.3, -0.5, **0.3**, ...)

분류 결과

[정치, 유머]



실험 개요 (3/3)

- **구현 환경**

- Apache Lucene 3.0.2: **인덱싱 라이브러리**
- KoreanAnalyzer-20100525: **오픈 소스 한글 형태소 분석기**
- LIBLINEAR: **선형 SVM 라이브러리**
 - <http://www.csie.ntu.edu.tw/~cjlin/liblinear/>
- H/W
 - CPU: Quad-core Xeon 2.33GHz
 - RAM: 32GB
 - OS: Ubuntu 64bit

- **SVM 학습 시간** (610,000 posts)

- **전처리**: 36 min.
- **학습**: 6 min.

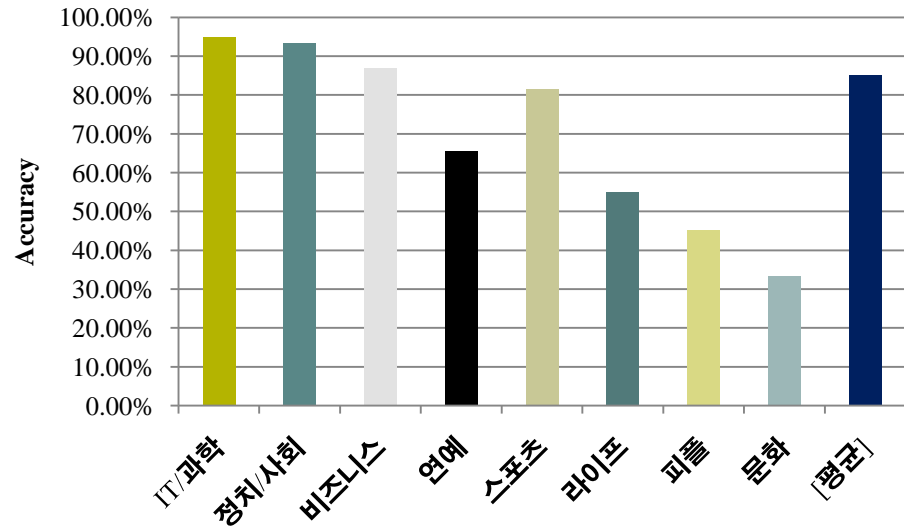
- **SVM 테스트 시간**

- **포스트 당** 30~50ms

실험 결과 및 분석(1/2)

• 대분류 자동분류 결과

	정확도
IT/과학	95.01%
정치/사회	93.35%
비즈니스	87.04%
연예	65.47%
스포츠	81.48%
라이프	55.00%
피플	45.12%
문화	33.33%
[평균]	85.11%



• 세부 분석

• 높은 정확율 항목

- IT/과학, 정치/사회

- 전문용어 등 분별력이 높은 피쳐들이 다수 존재

• 낮은 정확율 항목

- 연예, 라이프

- 다른 항목과 겹치는 경향을 지니는 분류집단임, 예) 연예인들의 럭셔리카

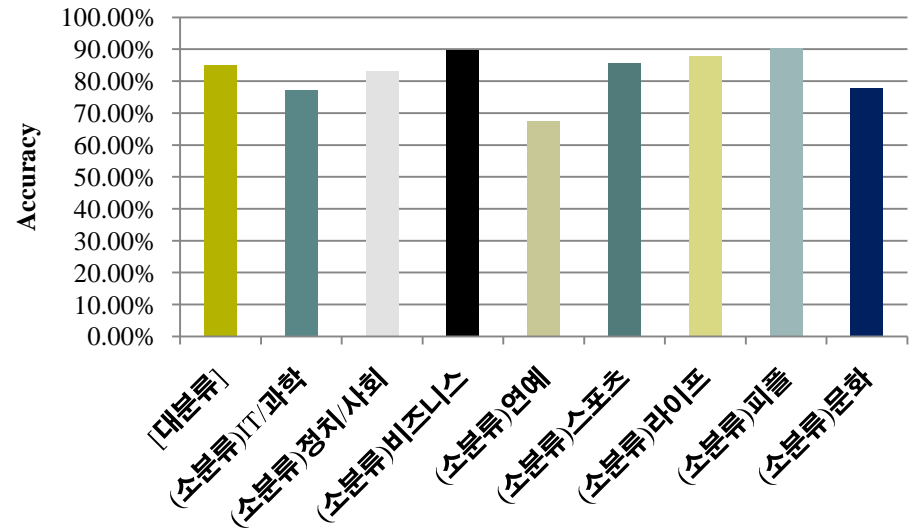
- 피플, 문화

- 주제의 다양성에 비해 학습 데이터수 미비

실험 결과 및 분석(2/2)

- 대분류의 분류결과를 사용하여 소분류의 분류를 시행

	정확도
[대분류]	85.11%
(소분류)IT/과학	77.08%
(소분류)정치/사회	83.36%
(소분류)비즈니스	89.61%
(소분류)연예	67.59%
(소분류)스포츠	85.71%
(소분류)라이프	87.78%
(소분류)피플	90.48%
(소분류)문화	77.78%
(소분류)평균	80.30%



- 세부 분석

- 평균 정확율 80.30%
- 낮은 정확율 항목
 - IT/과학
 - “104:블로그” 카테고리 – 다양한 주제가 혼합되어 있는 소분류 항목
 - 연예
 - 소분류들 사이에 유사성 지닌 분류 다수 존재, 예) TV vs. 드라마

- 실질 정확율: 68.25%

- 대분류 85% * 소분류 80.3% = 68.25%

결론

- 연구 내용
 - 국내 메타 블로그 사이트의 포스트를 대상으로 선형 SVM 기반 텍스트 분류 방법론 적용
 - 다계층 SVM 방법론
 - 멀티라벨 방법론
 - 높은 정확율 성능
 - 명확한 분류조건 및 풍부한 실험데이터
- 추가개선사항
 - 인물, 문화 등의 분류 등은 특화된 분류기법이나 전처리가 필요하다고 판단됨
 - 카테고리 최적화 방법론 탐색
 - 분류 프레임워크 내에서 분류별 평가 및 재구성 등의 방법 활용
 - 멀티라벨 방법론 적용 여부 판단
 - 멀티라벨 방법론의 성능 및 효용성 측정 필요

감사합니다