# 좋은 검색서비스를 만들기 위한 기계학습의 활용사례

김상범

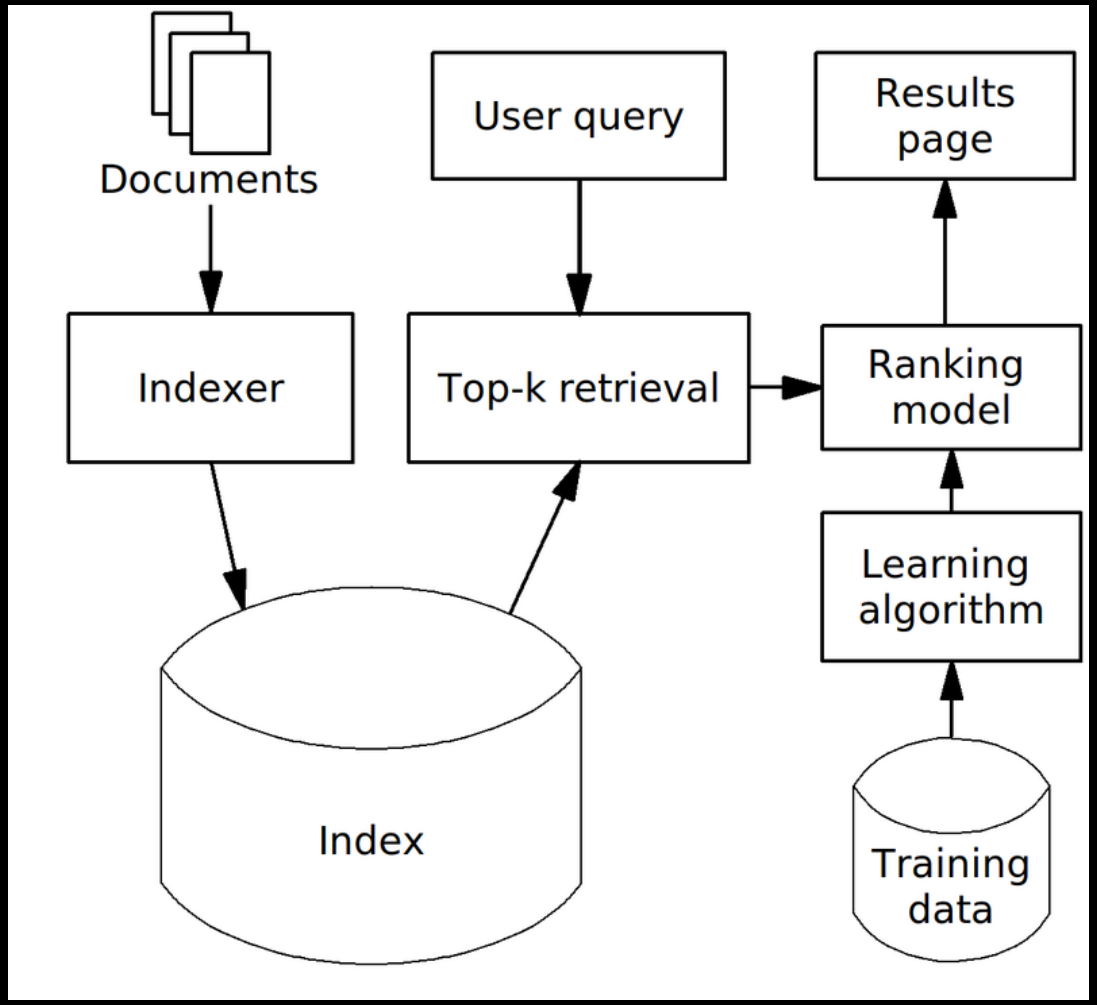
2016. 6.

NAVER

# Contents

1. 네이버 ?
2. 검색서비스 ?
3. 기계학습 활용사례
   - Ranking
   - Sequence Labeling
   - Text/Query Mining
   - Vision-Text Collaboration
   - Recommendation

# 검색서비스?




Coverage

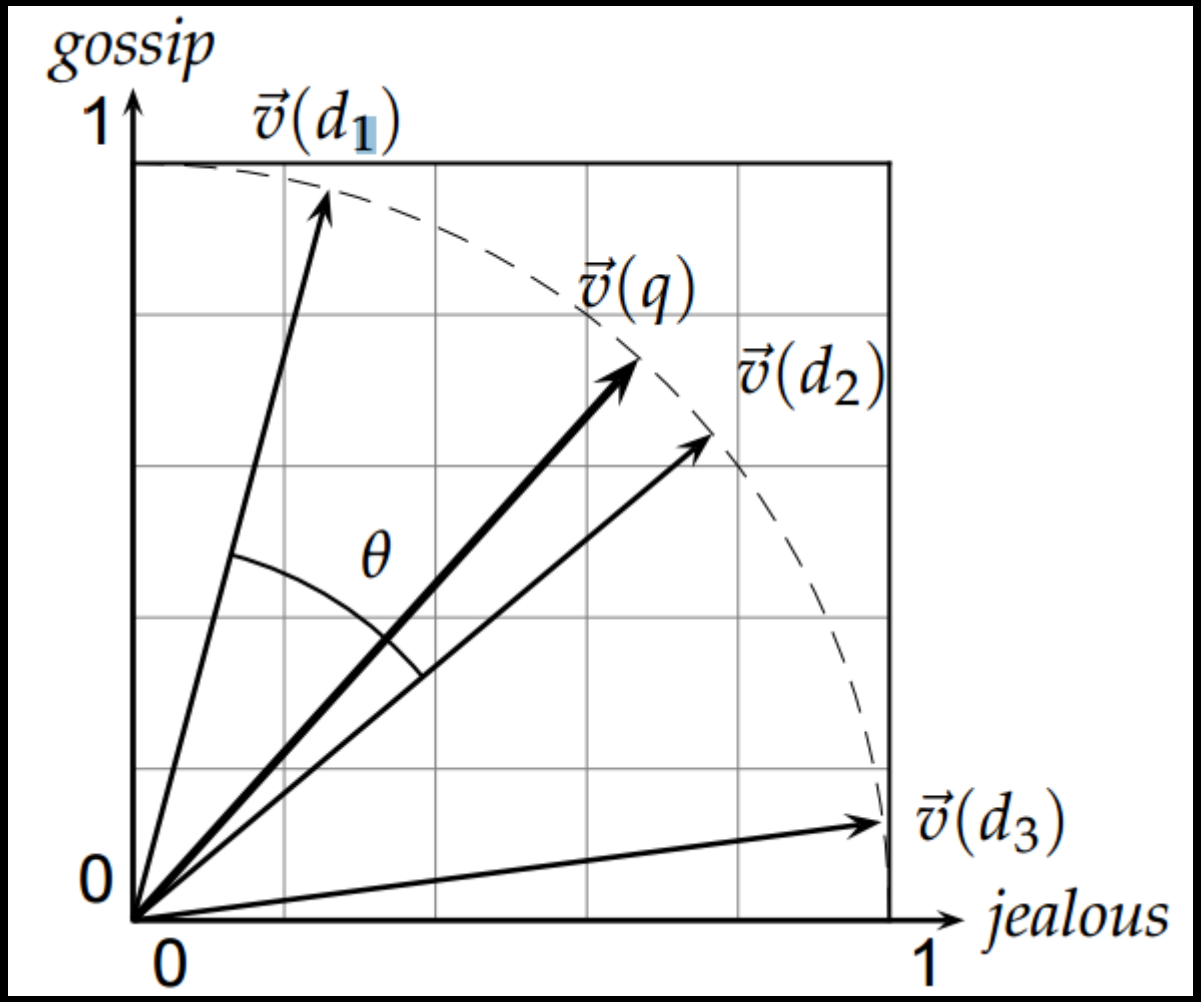
Efficiency

Effectiveness

Policy

# Ranking : Old Approaches

- Vector Space Model




http://nlp.stanford.edu/IR-book/pdf/06vect.pdf

# Ranking : Old Approaches

- Probabilistic Ranking Model

$$P(\mathrm{rel}|d,q) \propto_q \frac{P(\mathrm{rel}|d,q)}{P(\overline{\mathrm{rel}}|d,q)} \tag{2.1}$$

$$= \frac{P(d|\mathrm{rel},q)}{P(d|\overline{\mathrm{rel}},q)} \frac{P(\mathrm{rel}|q)}{P(\overline{\mathrm{rel}}|q)} \tag{2.2}$$

$$\propto_q \frac{P(d|\mathrm{rel},q)}{P(d|\overline{\mathrm{rel}},q)} \tag{2.3}$$

$$\approx \prod_{i \in \mathbf{V}} \frac{P(TF_i = tf_i|\mathrm{rel},q)}{P(TF_i = tf_i|\overline{\mathrm{rel}},q)} \tag{2.4}$$

$$w_i^{\mathrm{BM25}}(tf) = \frac{tf'}{k_1 + tf'} w_i^{\mathrm{RSJ}}$$

$$= \frac{tf}{k_1 \left((1-b) + b\frac{dl}{avdl}\right) + tf} w_i^{\mathrm{RSJ}}$$

# Ranking : Old Approaches

- Language Model based IR

$$P(d|q) \propto P(d) \prod_{t \in q} ((1-\lambda)P(t|M_c) + \lambda P(t|M_d))$$

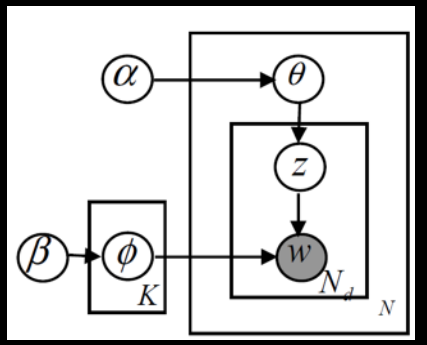
✓Simple smoothing

$$\hat{P}(t|d) = \frac{\mathrm{tf}_{t,d} + \alpha \hat{P}(t|M_c)}{L_d + \alpha}$$

✓LDA

$$P(w \mid D) = \lambda\left(\frac{N_d}{N_d + \mu} P'(w \mid D) + \left(1 - \frac{N_d}{N_d + \mu}\right) P'(w \mid coll)\right)$$

$$+ (1-\lambda)\left(\sum_{t=1}^{K} \frac{n_{-i,j}^{(w_i)} + \beta_{w_i}}{\sum_{v=1}^{V}(n_{-i,j}^{(v)} + \beta_v)} \times \frac{n_{-i,j}^{(d_i)} + \alpha_{z_i}}{\sum_{t=1}^{T}(n_{-i,t}^{(d_i)} + \alpha_t)}\right) \quad (9)$$




http://maroo.cs.umass.edu/pdf/IR-464.pdf

# Ranking : 실제상황

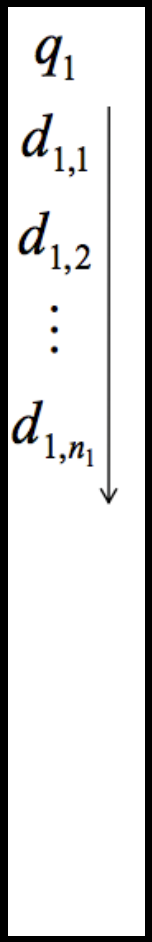
- 단어의 빈도만 갖고 랭킹을 하기에는 signal이 너무 많음
  - 질의와 매치된 단어의 폰트가 어떠한가?
  - 많은 사람들이 보았거나 링크를 걸었는가?
  - 다른 사람이 어떤 단어로 링크를 걸었는가?
  - 문서를 작성한 사람이 과거에 스팸작성자로 경고조치 됐었는가?
  - 언제 만들어진 문서인가?
  - 몇 명이 조회한 문서인가?
  - 문서의 제목과 본문의 관련성은 높은가?

- 상당수의 질의들에 대해서는 자동평가집합구축이 가능
  - 클릭수?
  - 위치를 고려한 클릭수?
  - 체류시간?

Feature가 많아지고
Evaluation Set확보가 비교적 쉬워짐
➔ 기계학습 기반 랭킹을 안 할 수 없음

# Ranking : Learning-to-Rank Overview

- 검색의 핵심은 **체계적으로** 줄세우기




= "한국 경제 전망"

= [ 0.66 , 2 , 0.08, 1 , 3, 1, 0, 4.2, 25, 0.43 ]

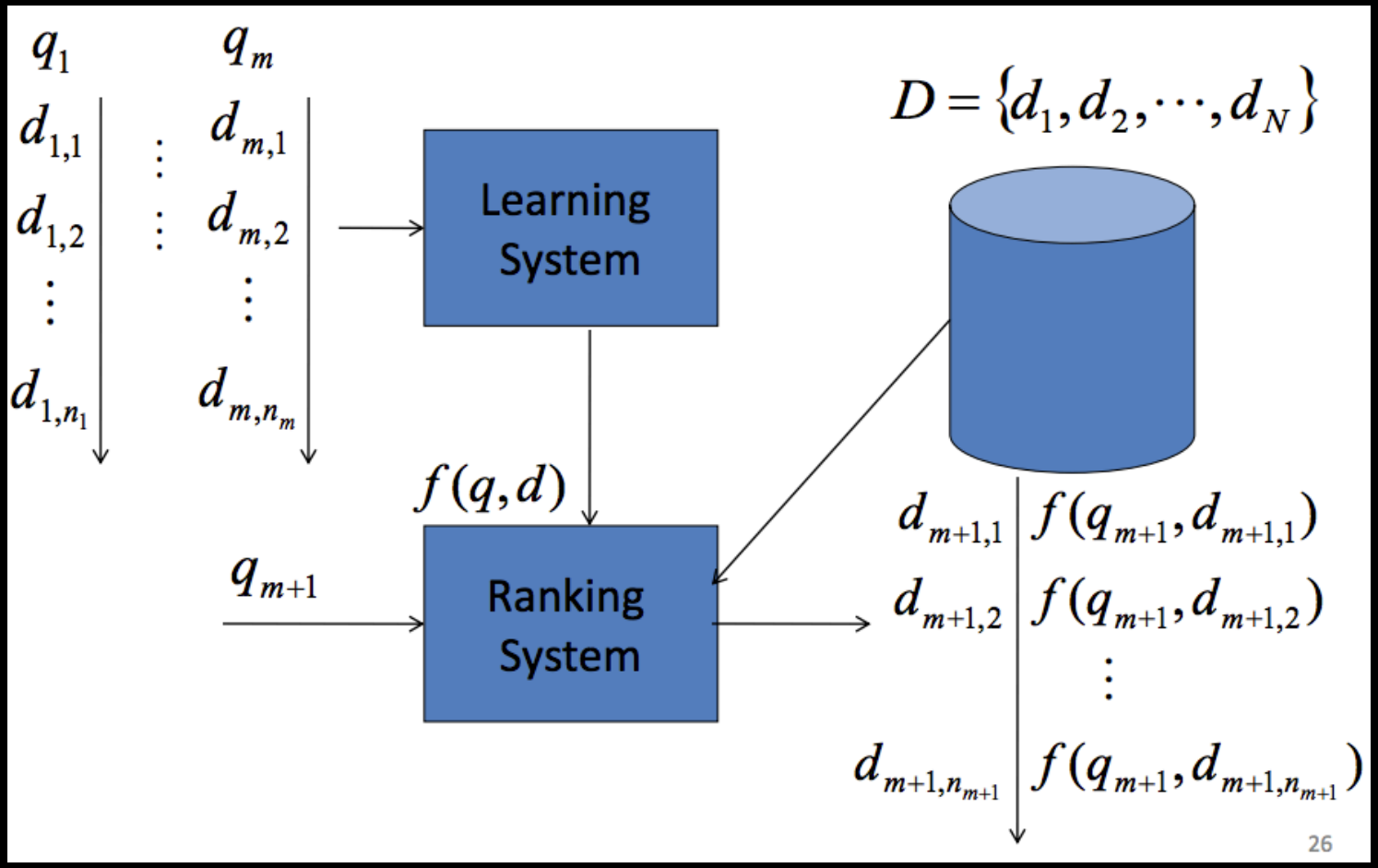
질의 단어 중 제목에 나타난 단어의 비율
제목에 출현한 질의 단어의 총 합
질의 단어별 "본문출현빈도/본문길이"의 합
질의 단어 중 본문에 나타난 단어의 비율

질의의 단어수
전문정보를 찾는 질의인가?
질의에 연예인 이름이 포함되어 있나?

문서가 포함된 사이트의 Site Authority
문서의 나이(Age)
문서의 품질(Quality)

# Ranking : Learning-to-Rank Overview

- 검색의 핵심은 **체계적으로** 줄세우기




http://research.microsoft.com/en-us/people/hangli/acml-tutorial.pdf

# Ranking : Learning-to-Rank Overview

- Training Data
  - Query : Document(URL) : Feature-Value-List : Grade
- Feature List
  - Matching Feature
    - ✓ Sum of tf*idf , Match term Ratio, etc
  - Document-specific Feature
    - ✓ Visit Count, Quality, Create Time, etc
  - Query-specific Feature
    - ✓ Length of query, HasPersonName, etc
- Grade
  - Perfect / Excellent / Good / Fair / Bad

# Ranking：Learning-to-Rank Overview

- Evaluation Measure：nDCG

## 4 documents: $d_1$, $d_2$, $d_3$, $d_4$

| i | Ground Truth | | Ranking Function$_1$ | | Ranking Function$_2$ | |
|---|---|---|---|---|---|---|
| | Document Order | $r_i$ | Document Order | $r_i$ | Document Order | $r_i$ |
| 1 | d4 | 2 | d3 | 2 | d3 | 2 |
| 2 | d3 | 2 | d4 | 2 | d2 | 1 |
| 3 | d2 | 1 | d2 | 1 | d4 | 2 |
| 4 | d1 | 0 | d1 | 0 | d1 | 0 |
| | $NDCG_{GT}$=1.00 | | $NDCG_{RF1}$=1.00 | | $NDCG_{RF2}$=0.9203 | |

$$DCG_{GT} = 2 + \left( \frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF1} = 2 + \left( \frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF2} = 2 + \left( \frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

$$MaxDCG = DCG_{GT} = 4.6309$$

http://web.stanford.edu/class/cs276/handouts/lecture8-evaluation_2014-one-per-page.pdf

# Ranking SVM

- Problem Definition
  - Input space: $X$
  - Ranking function $f : X \to R$
  - Ranking: $x_i \succ x_j \;\Leftrightarrow\; f(x_i; w) > f(x_j; w)$
  - Linear ranking function: $f(x; w) = \langle w, x \rangle$

    $$\langle w, x_i - x_j \rangle > 0 \;\Leftrightarrow\; f(x_i; w) > f(x_j; w)$$
  - Transforming to pairwise classification:

    $$(x_i - x_j, z),\; z = \begin{cases} +1 & x_i \succ x_j \\ -1 & x_j \succ x_i \end{cases}$$

# Ranking SVM

- Solution

$$\min_{w,\xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \xi_i$$

$$z_i \left\langle w, x_i^{(1)} - x_i^{(2)} \right\rangle \geq 1 - \xi_i \quad i = 1, \cdots, l$$

$$\xi_i \geq 0$$

$$\min_{w} \sum_{i=1}^{l} \left[ 1 - z_i \left\langle w, x_i^{(1)} - x_i^{(2)} \right\rangle \right]_+ + \lambda \| w \|^2$$

$$[s]_+ = \max(0, s) \quad \lambda = \frac{1}{2C}$$

# Ranking SVM

- Problems
  - Error 에도 그 중요도가 있는데 반영을 못한다
    (검색랭킹 specific한 평가척도를 직접 최적화하지는 못함)

Ranks: 3, 2, 1
ranking 1: 2 3 2 1 1 1 1
ranking 2: 3 2 1 2 1 1 1
ranking 2 should be better than ranking 1
Ranking SVM views them as the same

  - Query별 labeled 문서수에 따라 bias가 생길 수 있다

q1: 3 2 2 1 1 1 1
q2: 3 3 2 2 2 1 1 1 1 1
number of pairs for q1 : 2*(2-2) + 4*(3-1) + 8*(2-1) = 14
number of pairs for q2: 6*(3-2) + 10*(3-1) + 15*(2-1) = 31

개선된 RankSVM이나 Listwise 접근법 등 다양한 연구가 진행됨

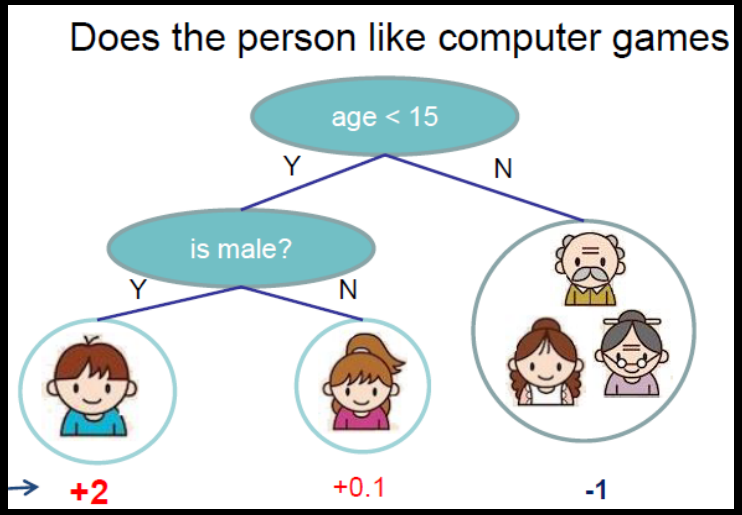
# Ranking만 잘하면 되나?



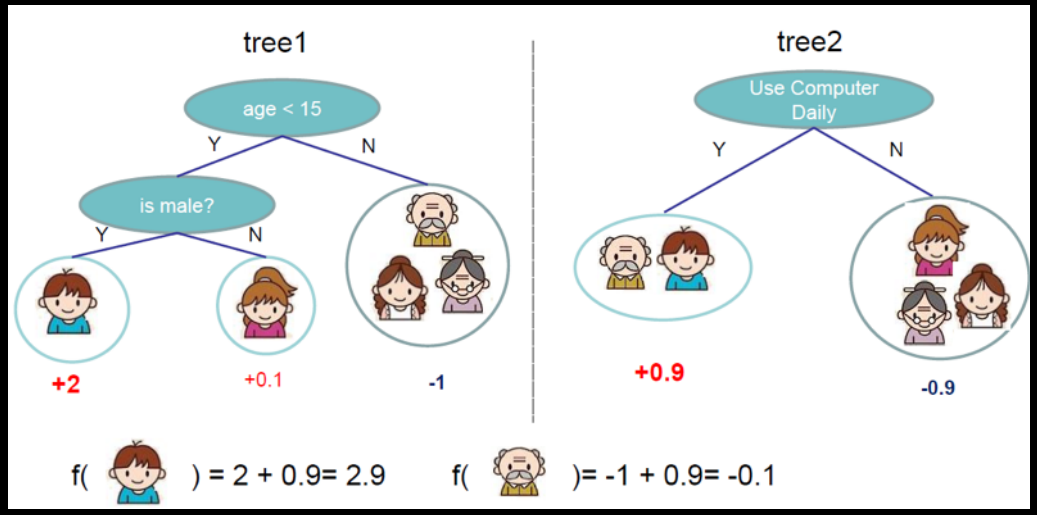

Ranking만 잘 하면 될까?
무조건 5개까지 보여주면 될까?

# Ranking and Regression : GBRT

- Regression Tree




- Regression Tree Ensemble




- Model: assuming we have K trees

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

- Objective

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Training loss　　Complexity of the Trees

https://homes.cs.washington.edu/~tqchen/pdf/BoostedTree.pdf

SVM, NN, LR같이
학습방법이 잘 연구되어온
Numerical vector/matrix기반
classifier가 아니라서…
학습방법 자체가 큰 연구토픽

# Ranking and Regression：GBRT

- GBRT (Gradient Boosted Regression Tree)

$$
\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\cdots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}
$$

New function

Model at training round t

Keep functions added in previous round

$$
\begin{aligned}
Obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + constant
\end{aligned}
$$

Goal: find $f_t$ to minimize this

$$
\begin{aligned}
&= \sum_{i=1}^{n} \left(y_i - (\hat{y}_i^{(t-1)} + f_t(x_i))\right)^2 + \Omega(f_t) + const \\
&= \sum_{i=1}^{n} \left[2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2\right] + \Omega(f_t) + const
\end{aligned}
$$

# Ranking and Regression : GBRT

- GBRT (Gradient Boosted Regression Tree)

$\hat{y}_i^{(0)} = 0$

$\hat{y}_i^{(1)}$

$\hat{y}_i^{(2)}$

$\hat{y}_i^{(t)}$

Model at training ro

v function

$Obj^{(t)} = \sum_i^n$

$= \sum_i^n$

ant

Goal: find $f_t$ to minimize this

$$= \sum_{i=1}^{n} \left( y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)) \right)^2 + \Omega(f_t) + const$$

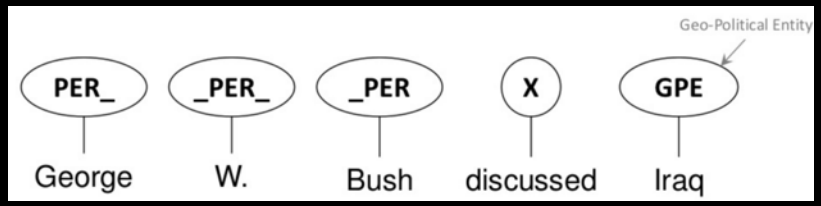
$$= \sum_{i=1}^{n} \left[ 2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2 \right] + \Omega(f_t) + const$$

# Sequence Labeling : 일반적인 적용분야

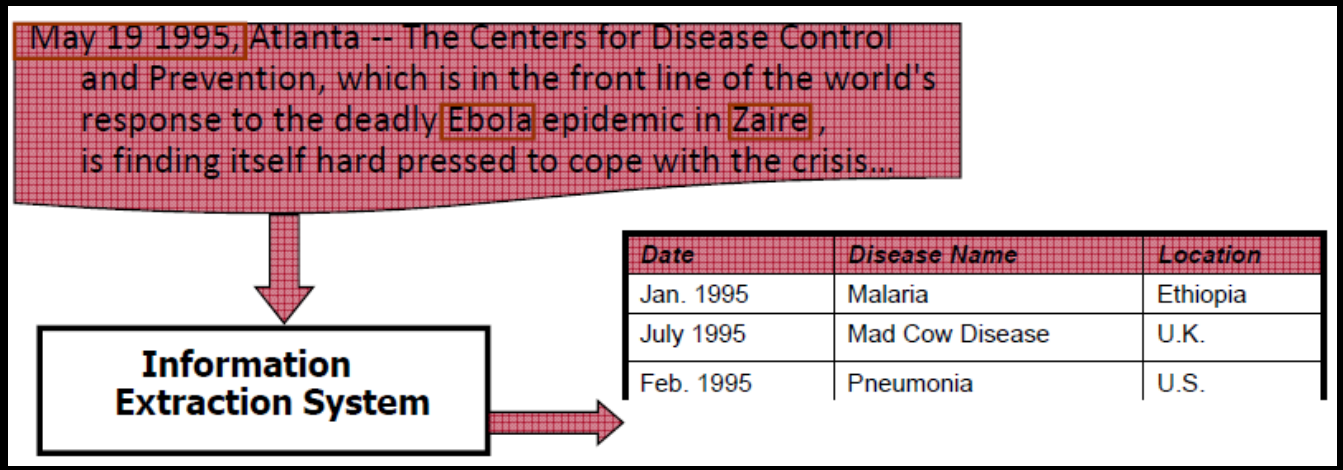
- Word Segmentation

| 검 | 색 | 시 | 스 | 템 | 용 | | 단 | 어 | 분 | 할 | 기 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| B | I | B | I | I | B | O | B | I | B | I | B |

- Named Entity Tagging



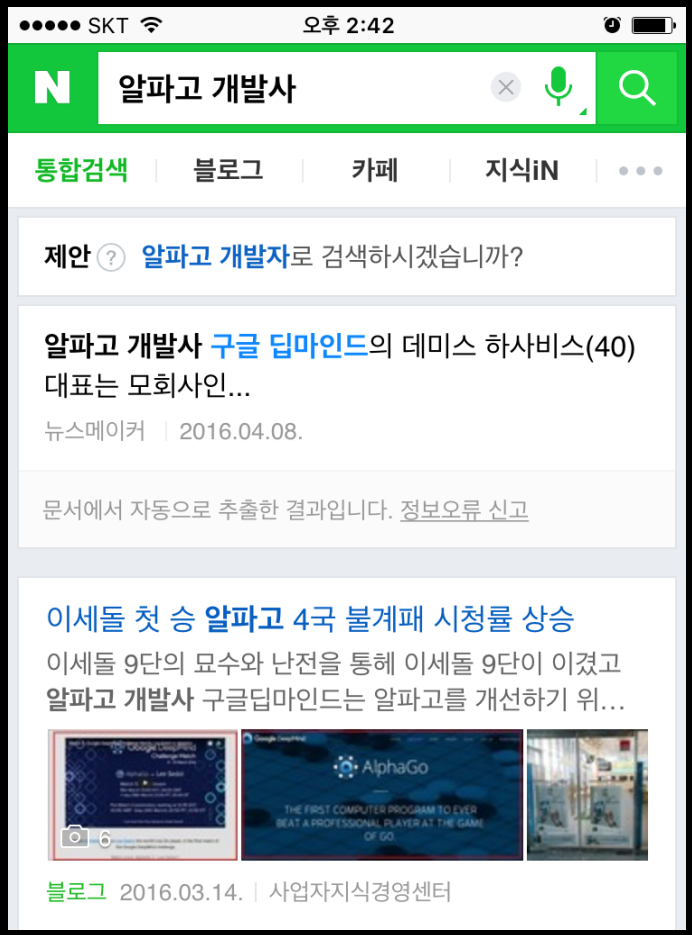

- Information Extraction



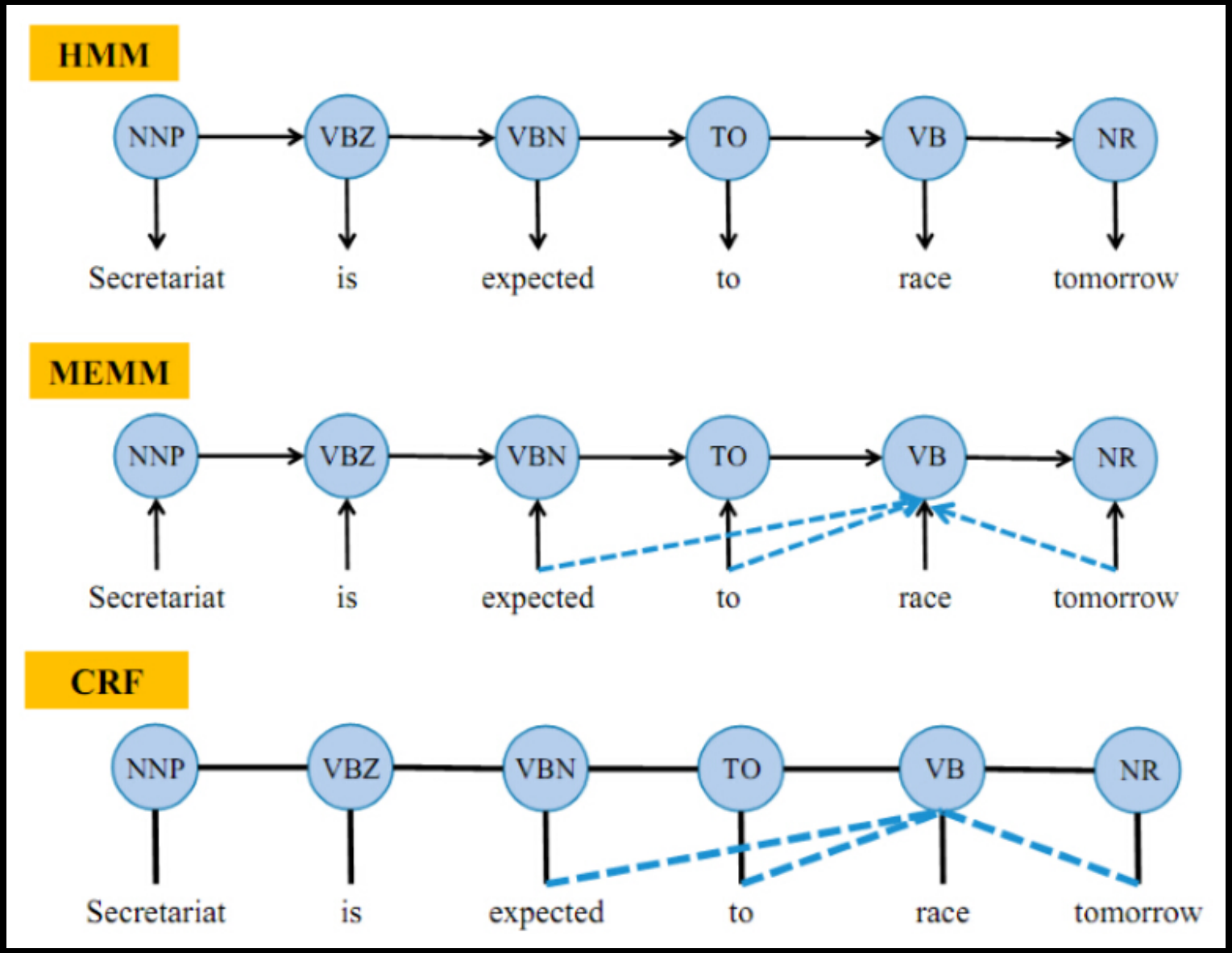
# Sequence Labeling : 실제 적용사례

- 자동번역기를 위한 분석기


- 즉답제공을 위한 관계추출



# Sequence Labeling : HMM / MEMM / CRF

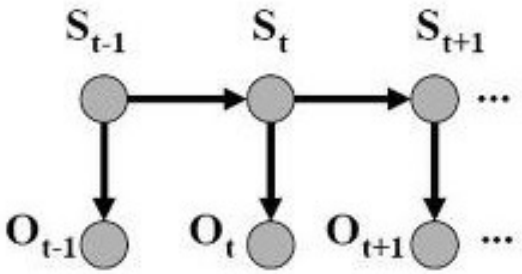


# Sequence Labeling : HMM / MEMM / CRF

$\vec{s} = s_1, s_2, \ldots s_n \qquad \vec{o} = o_1, o_2, \ldots o_n$

**HMM**

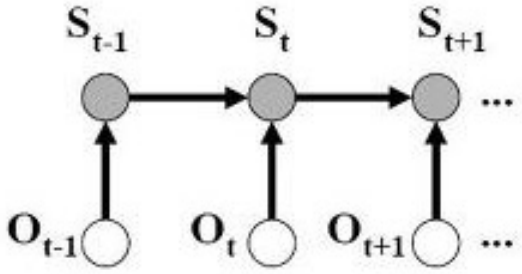
$$P(\vec{s}, \vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}) P(o_t \mid s_t)$$




**MEMM**

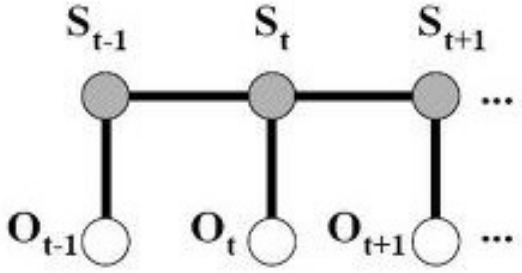
$$P(\vec{s} \mid \vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}, o_t)$$

$$\propto \prod_{t=1}^{|\vec{o}|} \frac{1}{Z_{s_{t-1}, o_t}} \exp\left( \sum_j \lambda_j f_j(s_t, s_{t-1}) + \sum_k \mu_k g_k(s_t, x_t) \right)$$




**CRF**

$$P(\vec{s} \mid \vec{o}) \propto \frac{1}{Z_{\vec{o}}} \prod_{t=1}^{|\vec{o}|} \exp\left( \sum_j \lambda_j f_j(s_t, s_{t-1}) + \sum_k \mu_k g_k(s_t, x_t) \right)$$



# Sequence Labeling : LSTM

- 새로운 강자 LSTM

Table 2: Comparison of tagging performance on POS, chunking and NER tasks for various models.

| | | POS | CoNLL2000 | CoNLL2003 |
|---|---|---|---|---|
| Random | Conv-CRF (Collobert et al., 2011) | 96.37 | 90.33 | 81.47 |
| | LSTM | 97.10 | 92.88 | 79.82 |
| | BI-LSTM | 97.30 | 93.64 | 81.11 |
| | CRF | 97.30 | 93.69 | 83.02 |
| | LSTM-CRF | **97.45** | 93.80 | 84.10 |
| | BI-LSTM-CRF | 97.43 | **94.13** | **84.26** |
| Senna | Conv-CRF (Collobert et al., 2011) | 97.29 | 94.32 | 88.67 (89.59) |
| | LSTM | 97.29 | 92.99 | 83.74 |
| | BI-LSTM | 97.40 | 93.92 | 85.17 |
| | CRF | 97.45 | 93.83 | 86.13 |
| | LSTM-CRF | 97.54 | 94.27 | 88.36 |
| | BI-LSTM-CRF | **97.55** | **94.46** | **88.83** (**90.10**) |

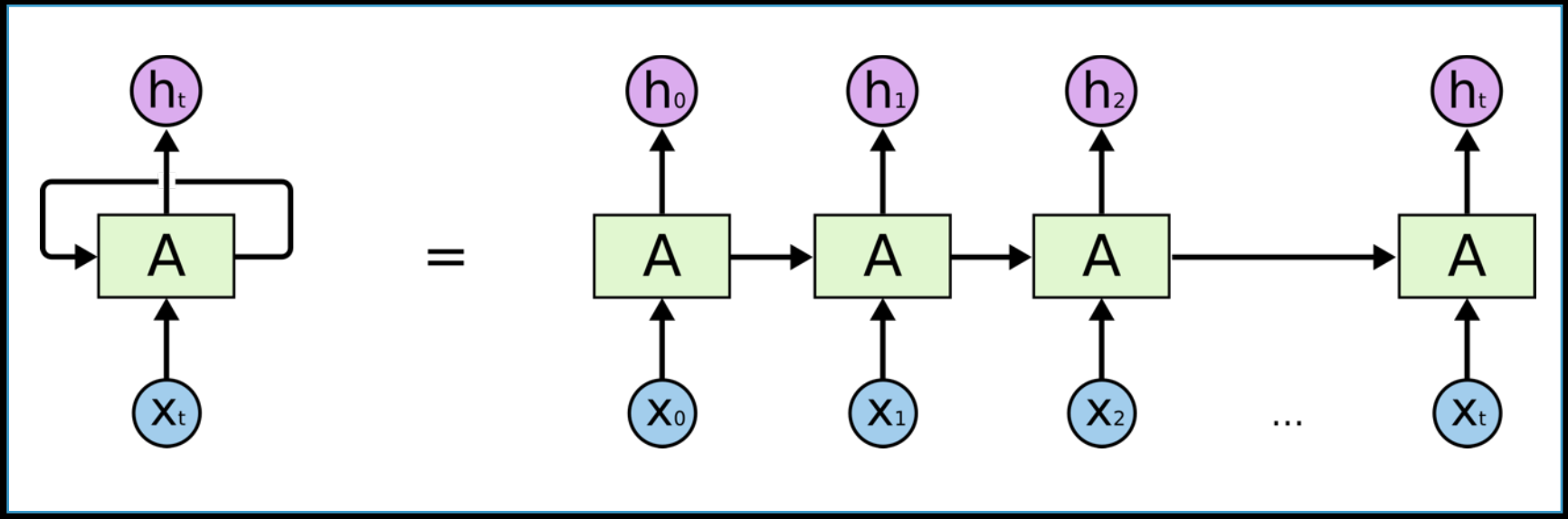
Table 3: Tagging performance on POS, chunking and NER tasks with only word features.

| | | POS | CoNLL2000 | CoNLL2003 |
|---|---|---|---|---|
| Senna | LSTM | 94.63 (-2.66) | 90.11 (-2.88) | 75.31 (-8.43) |
| | BI-LSTM | 96.04 (-1.36) | 93.80 (-0.12) | 83.52 (-1.65) |
| | CRF | 94.23 (-3.22) | 85.34 (-8.49) | 77.41 (-8.72) |
| | LSTM-CRF | 95.62 (-1.92) | 93.13 (-1.14) | 81.45 (-6.91) |
| | BI-LSTM-CRF | **96.11** (-1.44) | **94.40** (-0.06) | **84.74** (-4.09) |

http://arxiv.org/pdf/1508.01991v1.pdf

# Sequence Labeling : LSTM

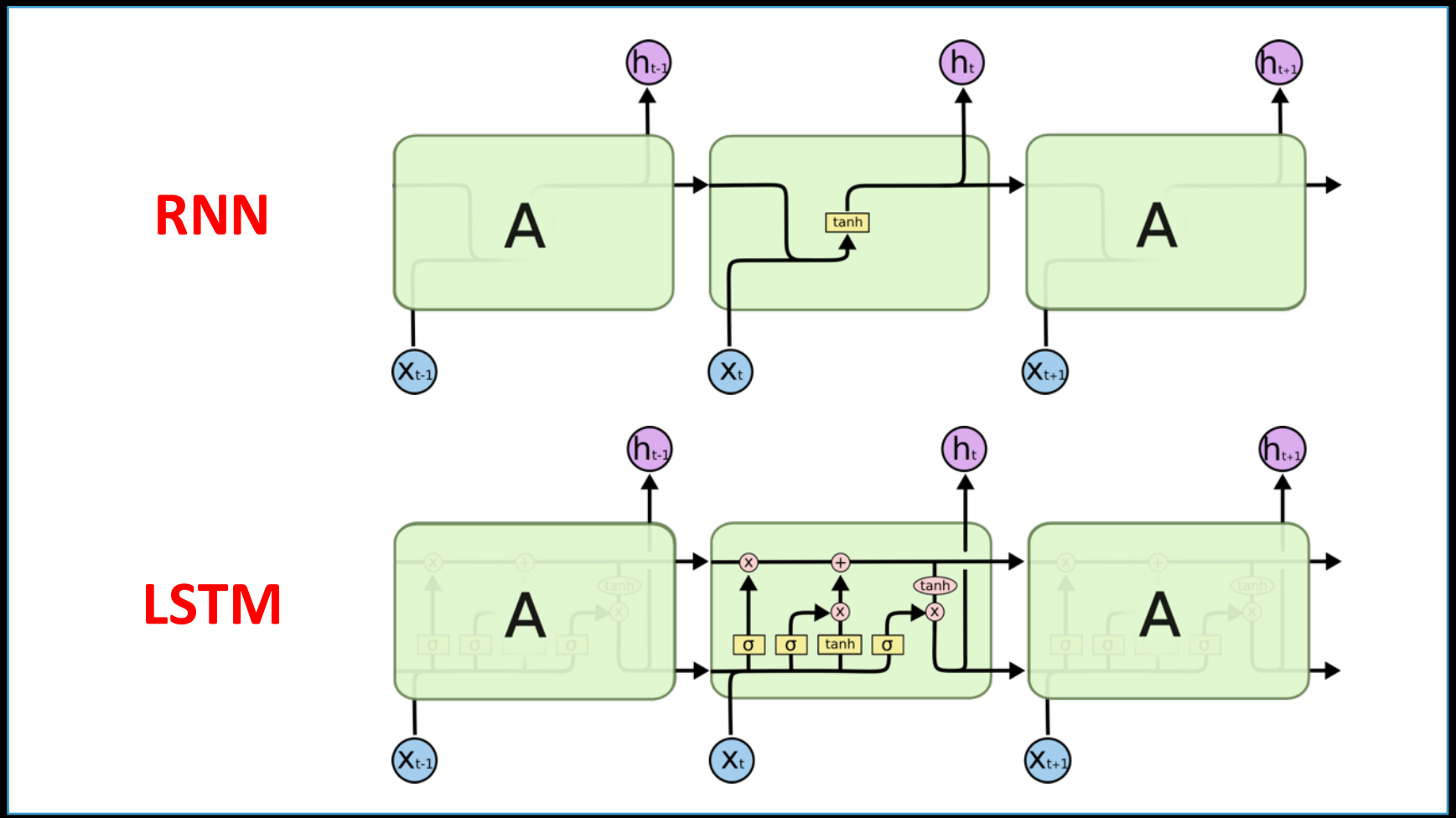
- Motivation : RNN의 원거리 의존관계 문제




- 이전상태를 기억하면서 현재 입력을 바탕으로 결과를 내놓는 다는 점에서 많은 진보를 가져다 줌
- “the clouds are in the _____ ”
- “ I grew up in France where my mother still lives. So I speak fluent _____” → 원거리 의존관계 문제 ( long-term dependency)

# Sequence Labeling : LSTM

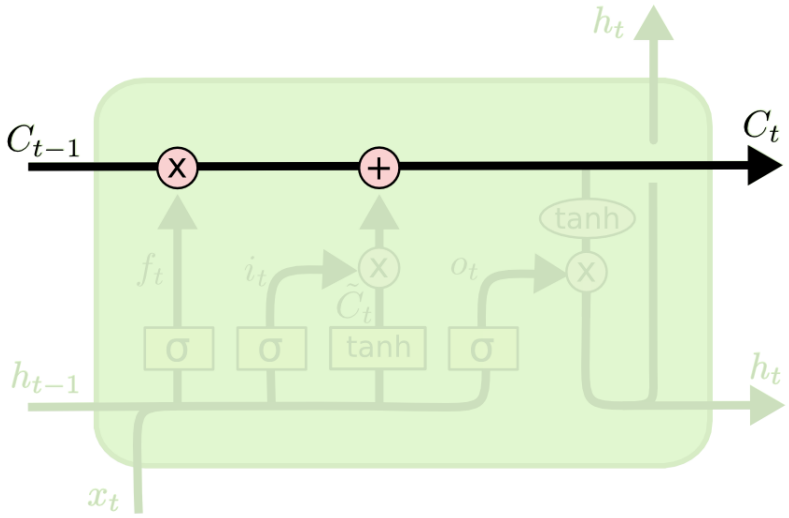
- RNN vs LSTM




http://colah.github.io/posts/2015-08-Understanding-LSTMs/

# Sequence Labeling : LSTM

- Cell state 의 도입




기존의 RNN에 없던 C는
이전 히스토리 중 의미있는 정보들을 갖고 있는 벡터.

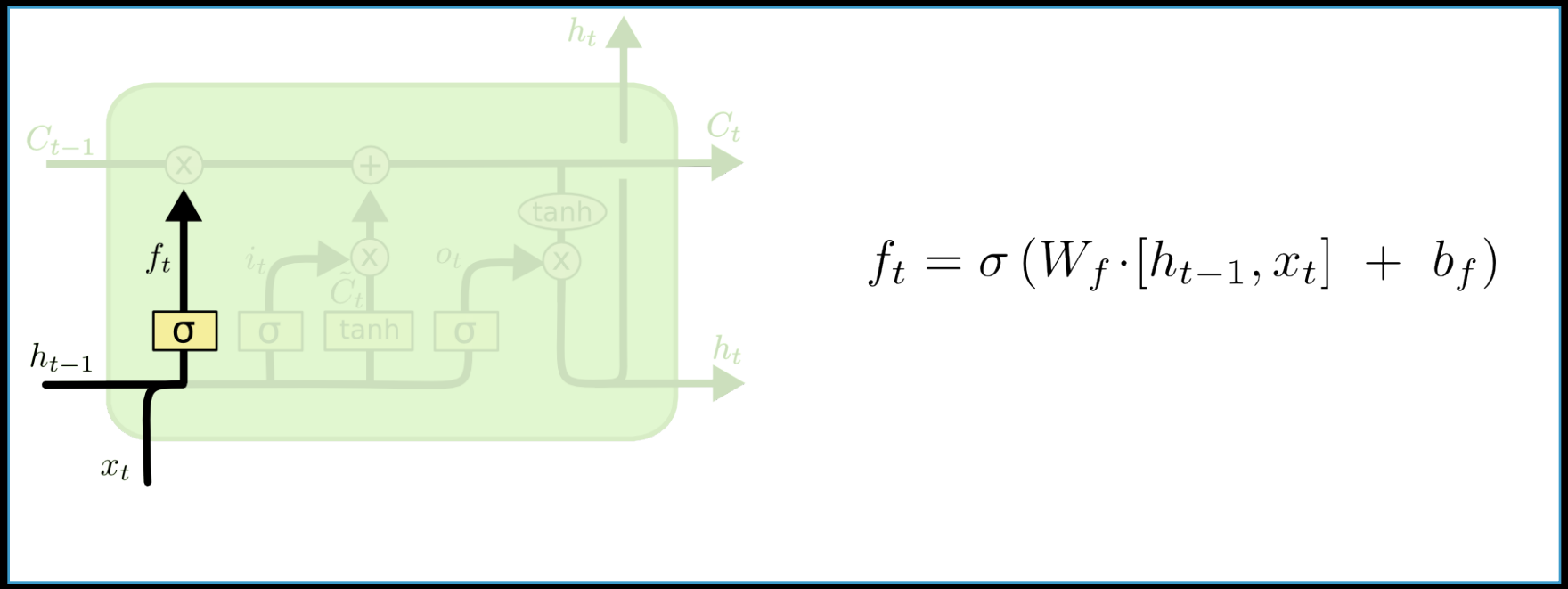
예를 들어 문장을 계속 읽어나가면서 다음 단어를 예측하는 LSTM이라면,
가장 최근에 나타난 주어의 성별정보를 C의 k번째 원소에 저장
(He, She 등을 적절히 생성해내려고)

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

- 이전셀상태에 뭔가 곱해지고 더해져서 최종 셀상태가 됨
- 곱해질때는 이전셀상태 각 원소의 유지여부가 결정되고, 더해질 때는 셀 걸러진 이전셀상태에 뭔가 새로운 것(정보)이 추가되는 것
- 출력(h)은 계산된 이번셀상태에 따라 결정됨

# Sequence Labeling : LSTM

- Forgot gate 를 통한 이전셀상태 억제장치 준비




$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

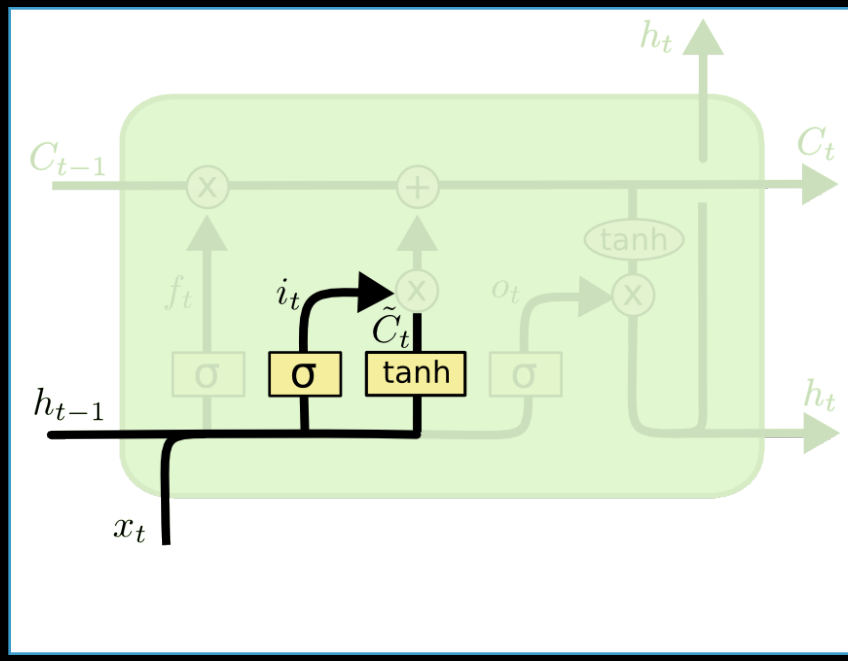
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

- 이전출력(h)과 현재입력(x)을 받아 f를 계산해서 이전셀상태에 곱함
- f에 따라 이전 셀상태벡터의 어떤 값은 리셋되고 어떤 값은 살아남음

이전출력+현재입력을 보아 “새로운 주어”라는 판단이 들면 f의 k번째 원소는 0

# Sequence Labeling : LSTM

- Input gate 를 통한 이번입력의 반영정도 준비




$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

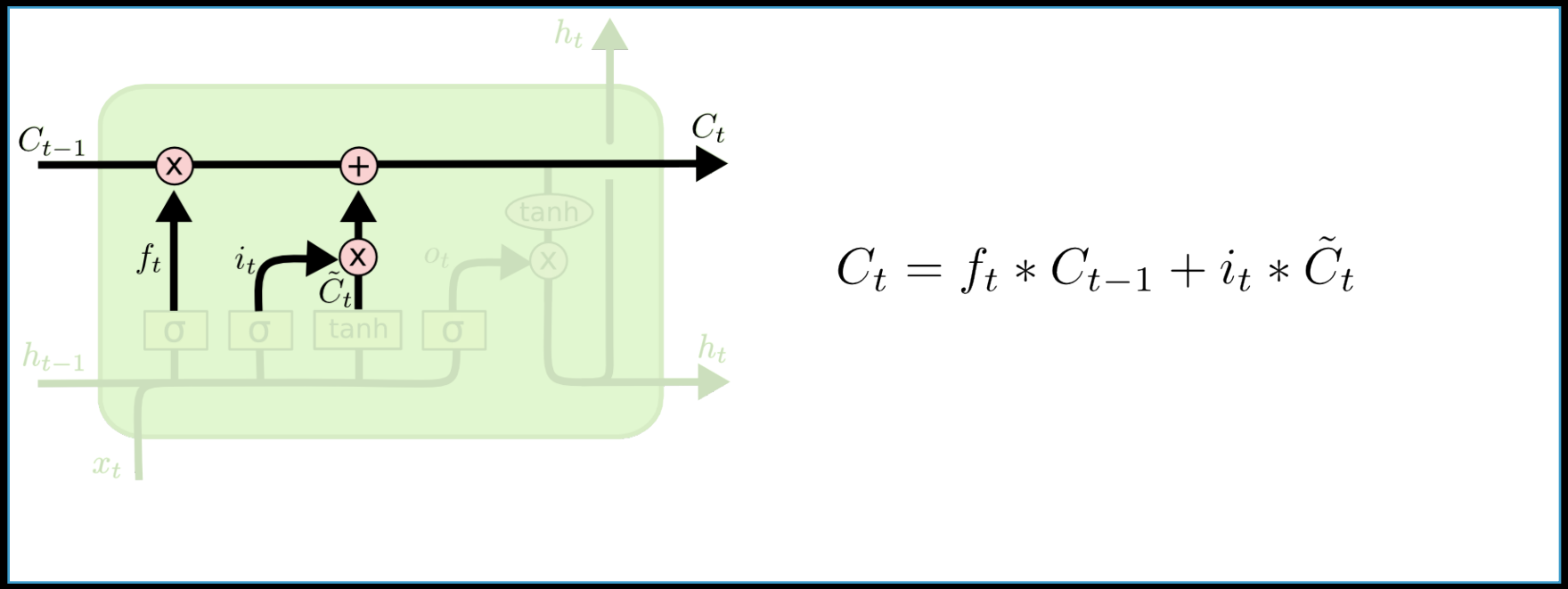
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

- 이전출력(h)과 현재입력(x)으로 i를 계산함
- 이전출력(h)과 현재입력(x)으로 이번입력에 기반한 셀상태를 계산함

이전출력+현재입력을 보아 “새로운 주어”라는 판단이 들면
i는 1이고 셀상태의 k번째 원소는 새로운 성별값

# Sequence Labeling : LSTM

- 최종셀상태 G의 계산




$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

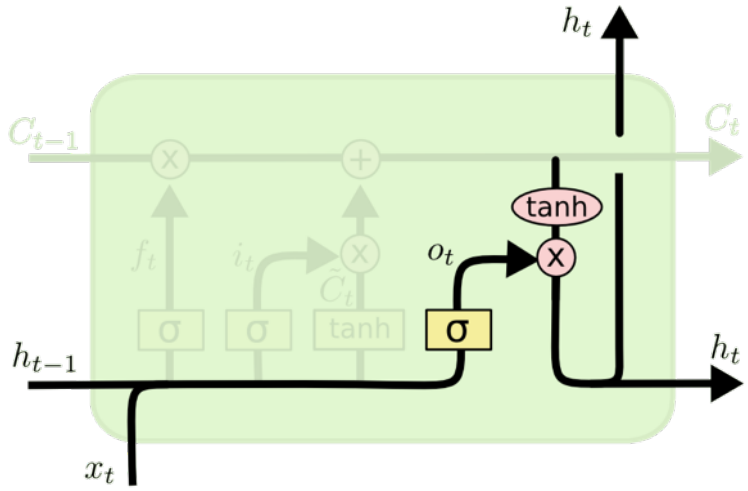
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

- f로 이전셀상태의 유지정도를 결정
- i로 단순 셀상태값이 최종 셀상태값에 끼치는 영향력을 결정

앞에서 구한 f와 i 및 셀상태를 사용하여 최종셀상태를 업데이트 ( 최근에 본 주어의 성별 )

# Sequence Labeling : LSTM

- 최종출력(h)계산을 위한 셀상태와 output gate 의 활용




$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

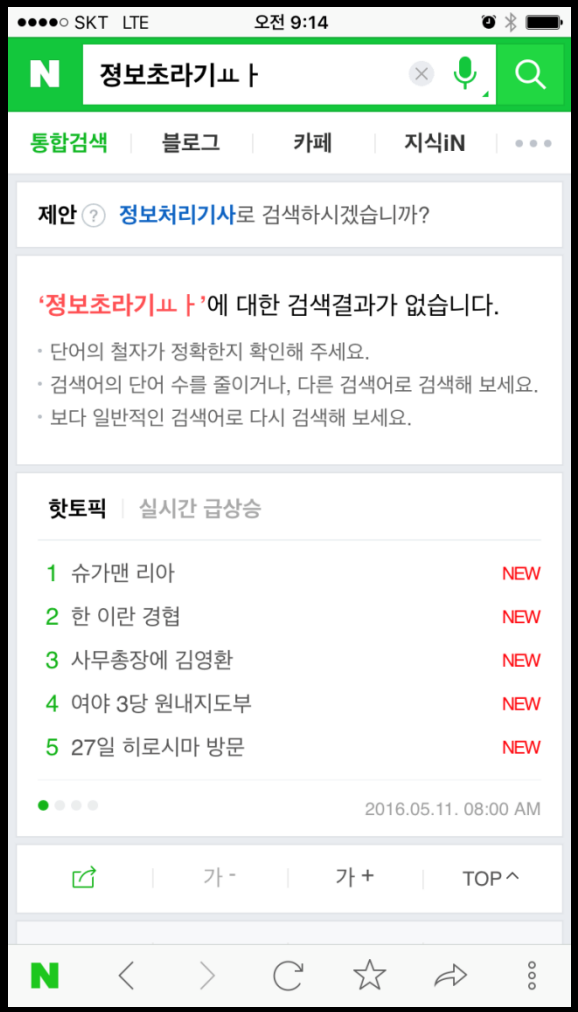
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

- 최종출력(h)은 최종셀상태(C)의 일부

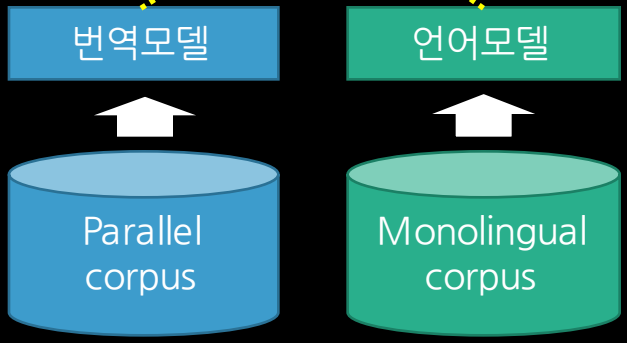
다음 단어가 무엇인지를 예측하기 위해서는,
최종셀상태 정보 중 방금 바뀐 최근 주어의 성별정보는 필요 없고
대신 동사의 단/복수 형 결정을 위해 주어의 “수”정보가 필요하므로
C의 k번째 정보는 내보내지 않는 대신 k+1번째 정보만 내보낼 수 있다.

# Text/Query Mining

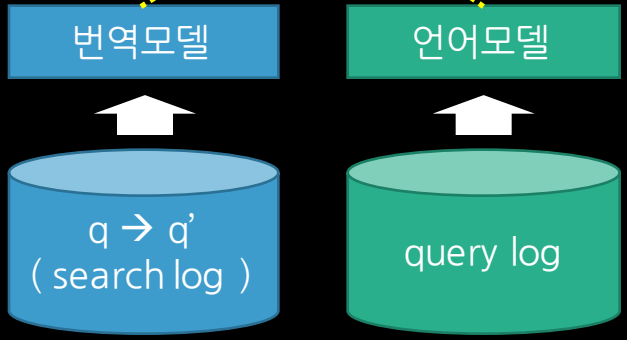
- Query Suggestion & Reformulation




$$e_{best} = argmax_e P(f|e)P(e)$$




$$q'_{best} = argmax_{q'} P(q|q')P(q')$$



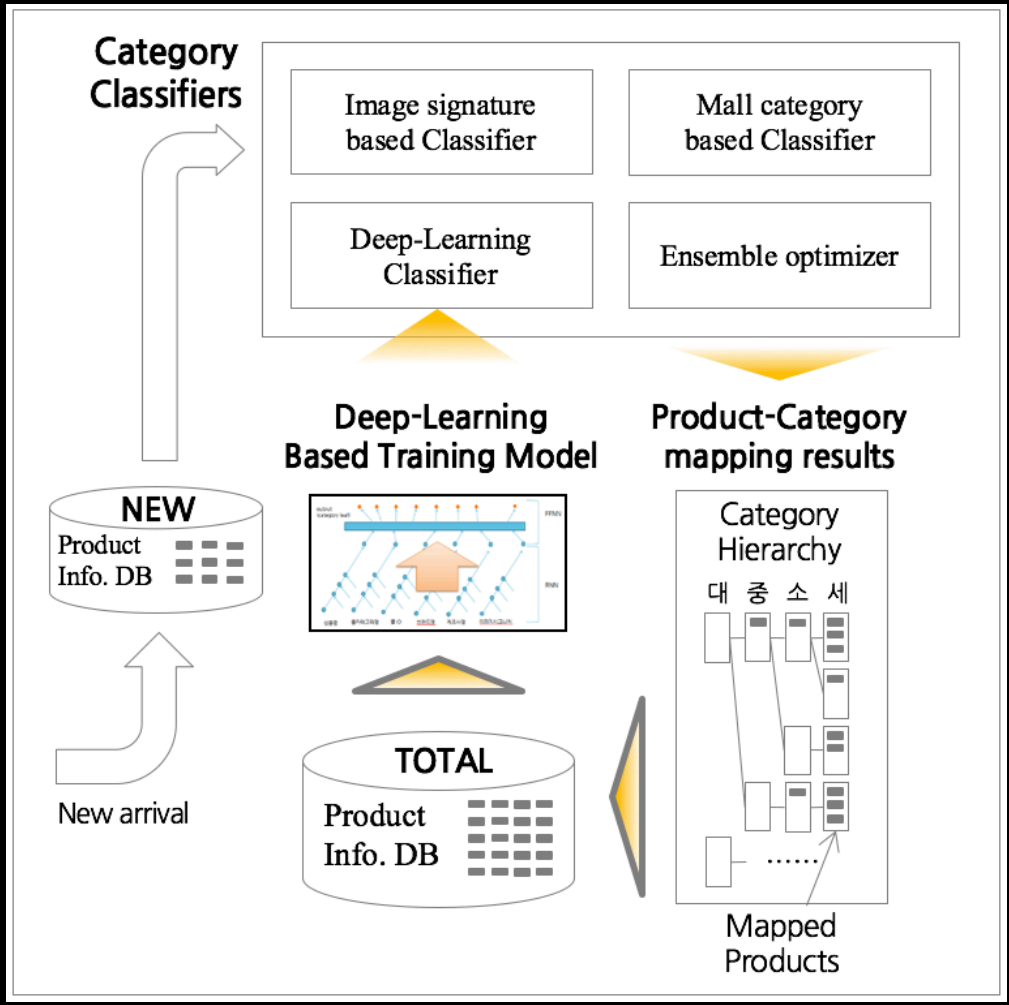
# Text/Query Mining

- Product Categorization

4억여개의 상품
4천여개의 카테고리

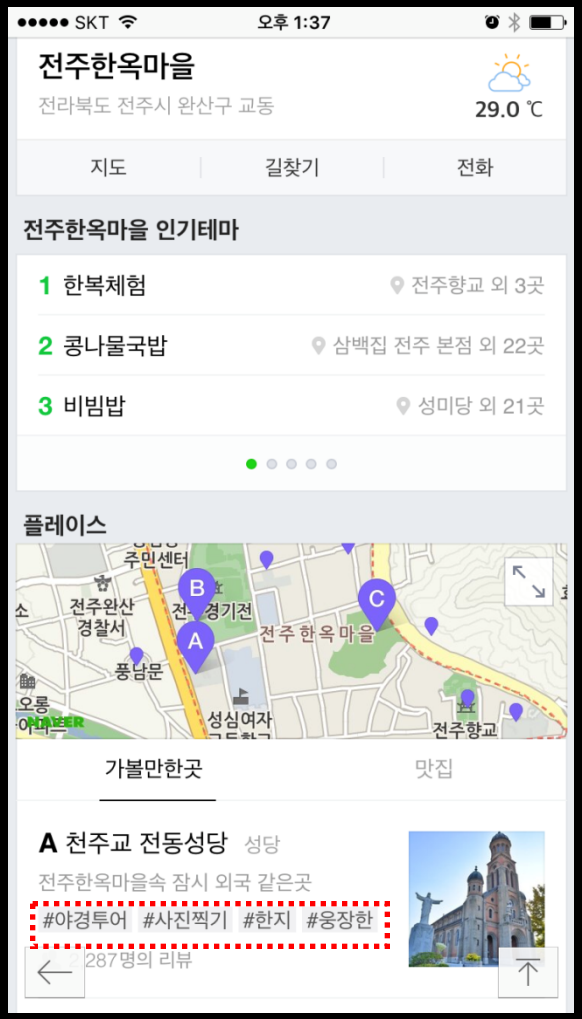
예)
스포츠/레저 > 수영 > 비치웨어 > 커플비치웨어
디지털/가전 > 음향가전 > 홈시어터 > 조합형홈시어터

카테고리 등록정보 신뢰도도 낮고
더 나은 쇼핑서비스를 위해
카테고리를 개편하기도 하고…



# Text/Query Mining

- Place Analysis




S1: 아이들과 당일치기로 전주를 다녀왔어요
S2: 아이들 교육에도 최적의 장소
S3: 비오는날 운치있는 전주여행
S4: 운치있고 여유로운 곳이었어요
S5: 야간에 산책하면서 다니는 재미
S6: 날씨 즐기며 천천히 걸어다니는 것도

S1: 아이들과 당일치기로 전주를 다녀왔어요
S2: 아이들 교육에도 최적의 장소
S3: 비오는날 운치있는 전주여행
S4: 운치있고 여유로운 곳이었어요
S5: 야간에 산책하면서 다니는 재미
S6: 날씨 즐기며 천천히 걸어다니는 것도

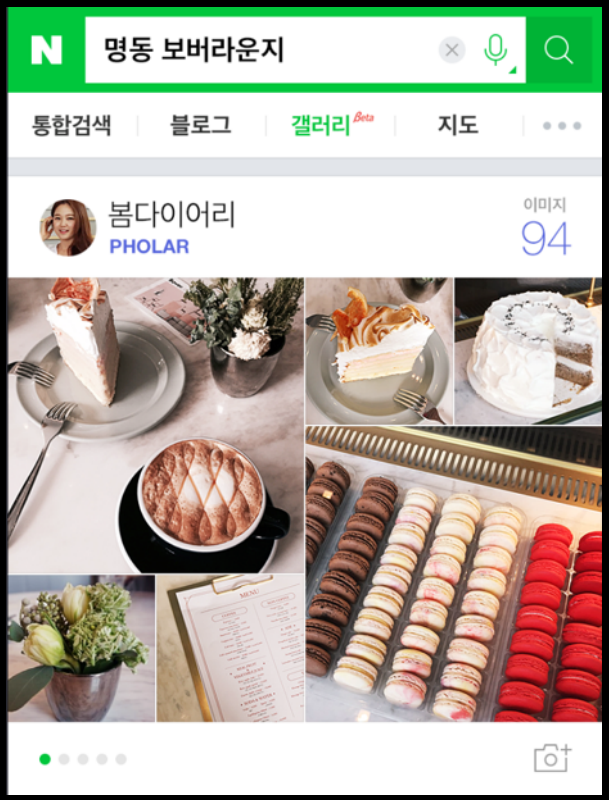
# Vision-Text Integration

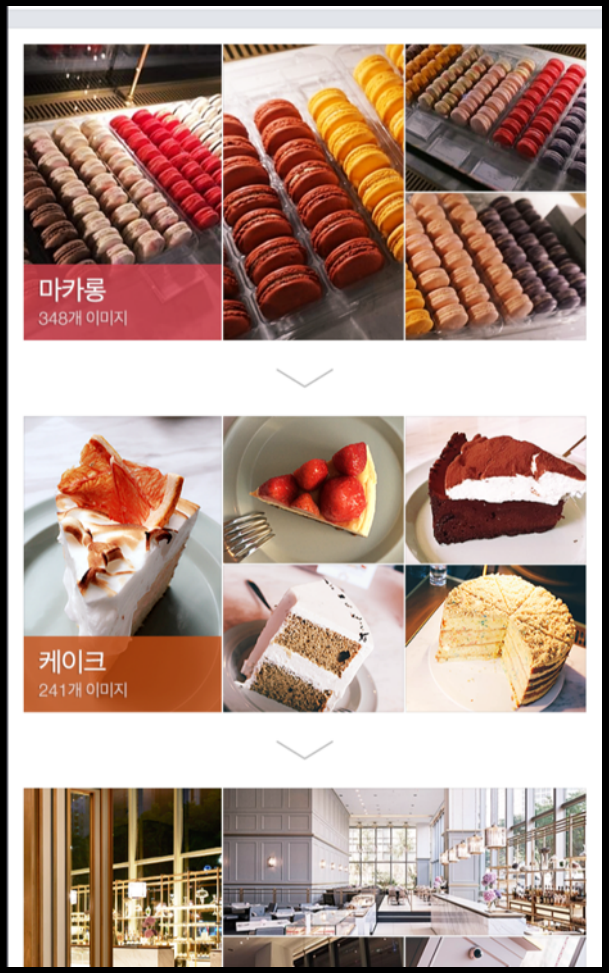
- 스타 타임라인



# Vision-Text Integration

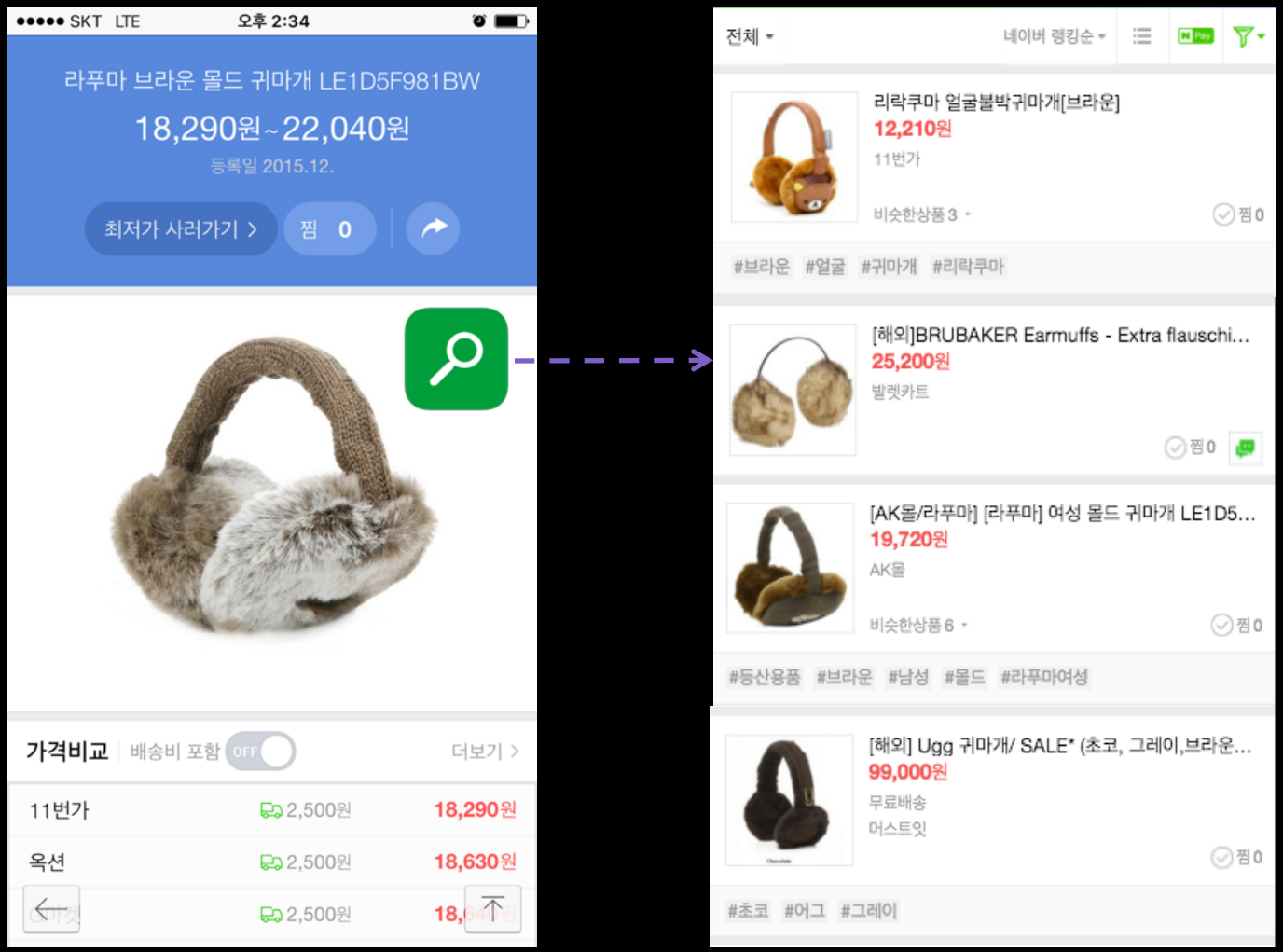
- 음식점 포토요약



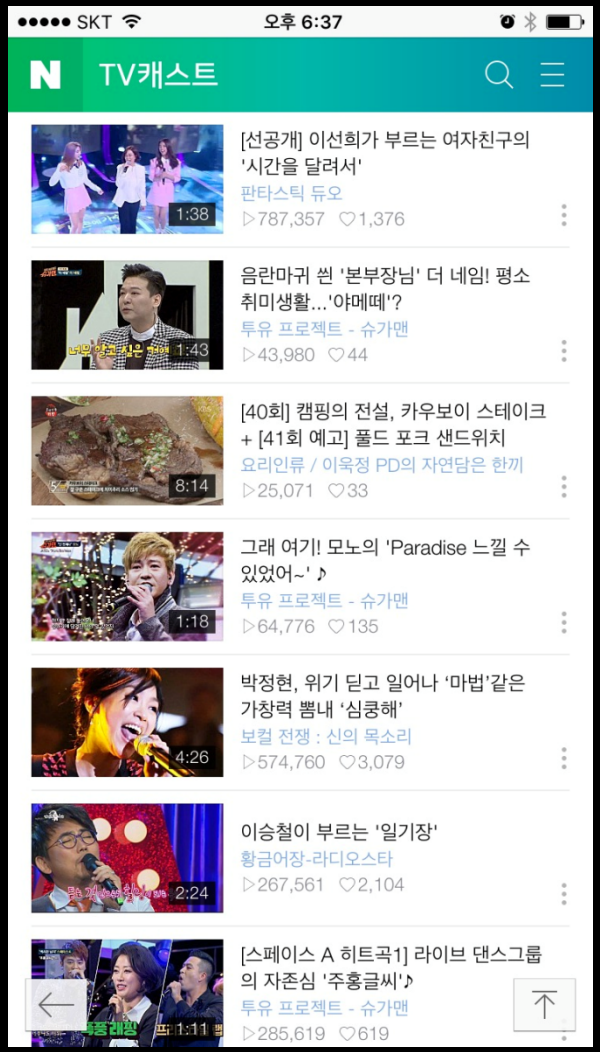

# Vision-Text Integration

- 스타일 서치



# Recommendation

- 어떻게 하면 네이버에서 더 오래 즐거운 시간을 보내실까?



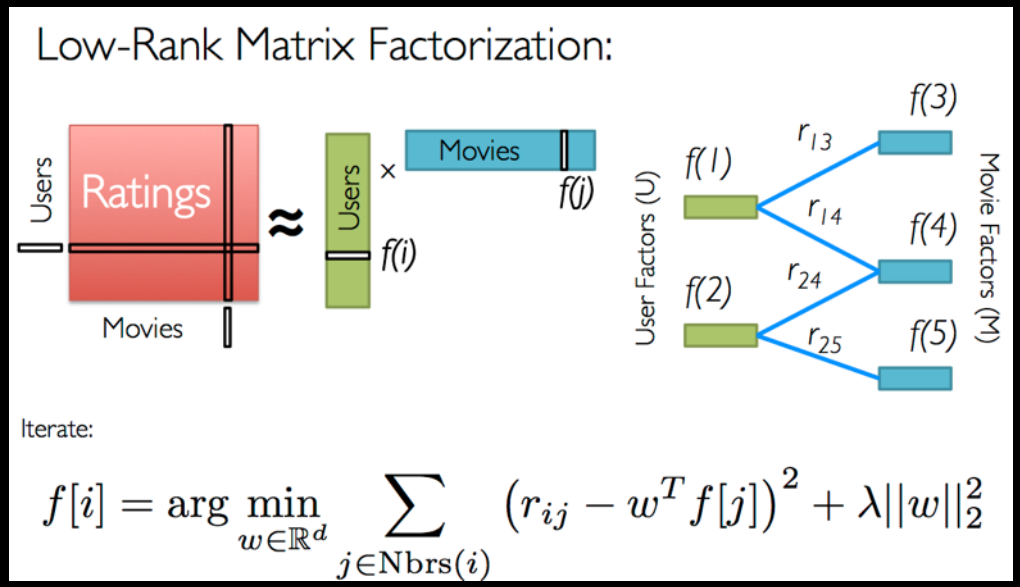

"투유프로젝트-슈가맨" 이라는 방송의 클립을 보고 난 사용자에게 어떤 동영상을 추천해주면 좋을까?

"슈가맨같은 동영상"

"의외의 다른 동영상"

# Recommendation

- Collaborative Filtering + Diversity




https://databricks-training.s3.amazonaws.com/movie-recommendation-with-mllib.html

DIVERSIFY($k$): Given query $q$, a set of documents $R_q$, a probability distribution of categories for the query $P(c|q)$, the quality values of the documents $V(d|q,c)$, $\forall d \in \mathcal{D}$ and an integer $k$. Find a set of documents $S \subseteq R_q$ with $|S| = k$ that maximizes

$$P(S|q) = \sum_c P(c|q) \left(1 - \prod_{d \in S} (1 - V(d|q,c))\right) \quad (3)$$

→ Novelty factor

감사합니다