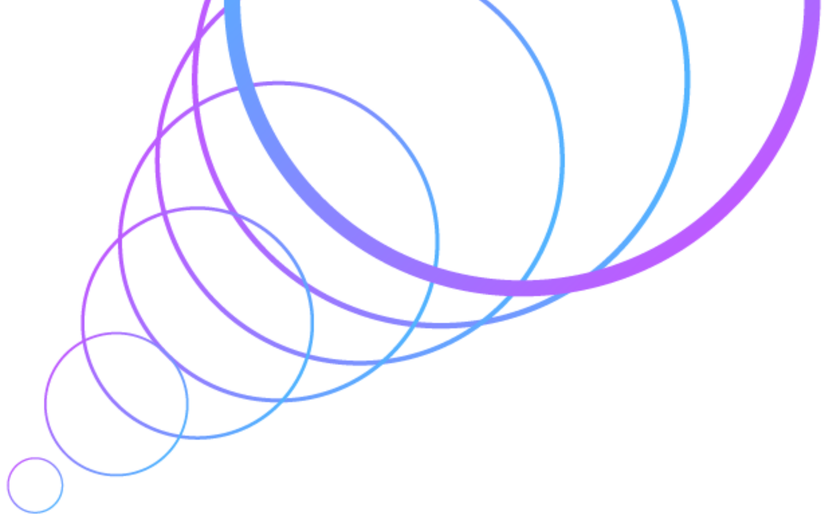
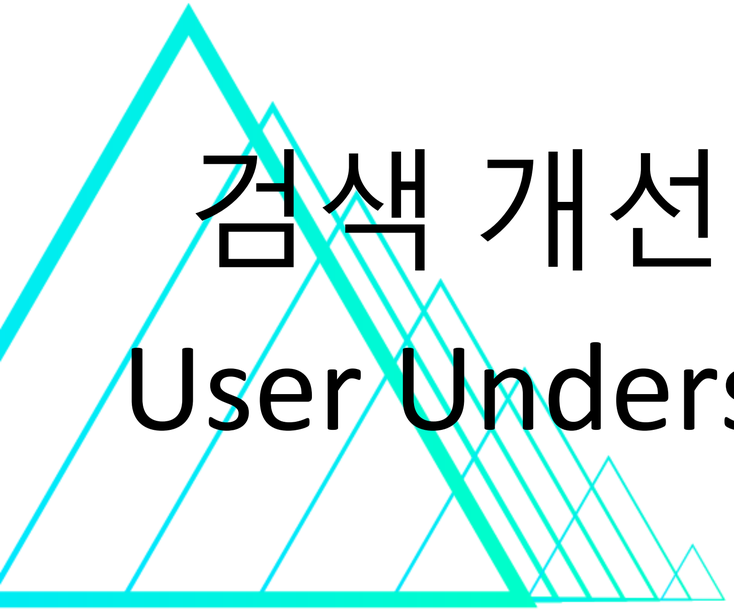




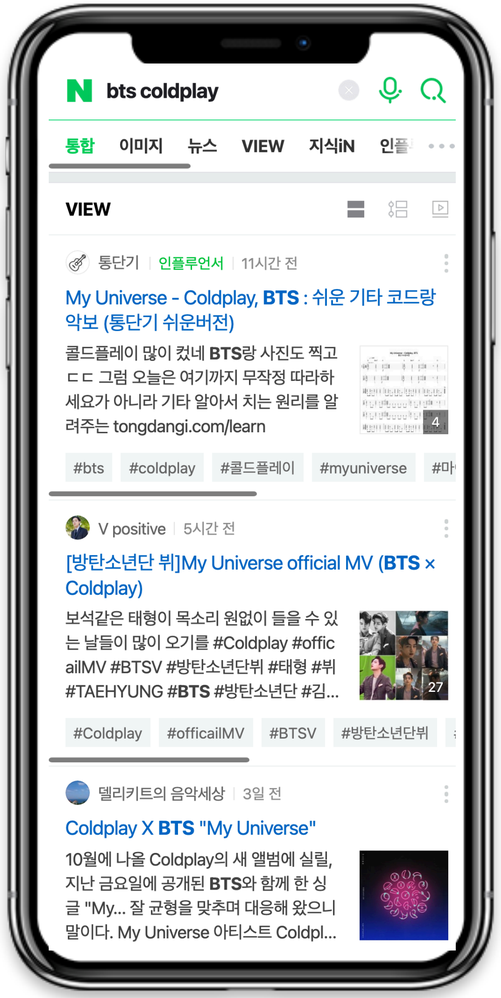
검색 개선을 위한 검색 사용자 이해하기
User Understanding for Search Enhancement

백준선, 백수현 NAVER

CONTENTS

1. 검색 서비스와 사용자

1.1. 검색 서비스의 구성요소



[네이버 검색서비스 예시]

[라인 검색서비스 예시]

1.1. 검색 서비스의 구성요소

DEVIEW 2021

검색 서비스를 통한 사용자 행위

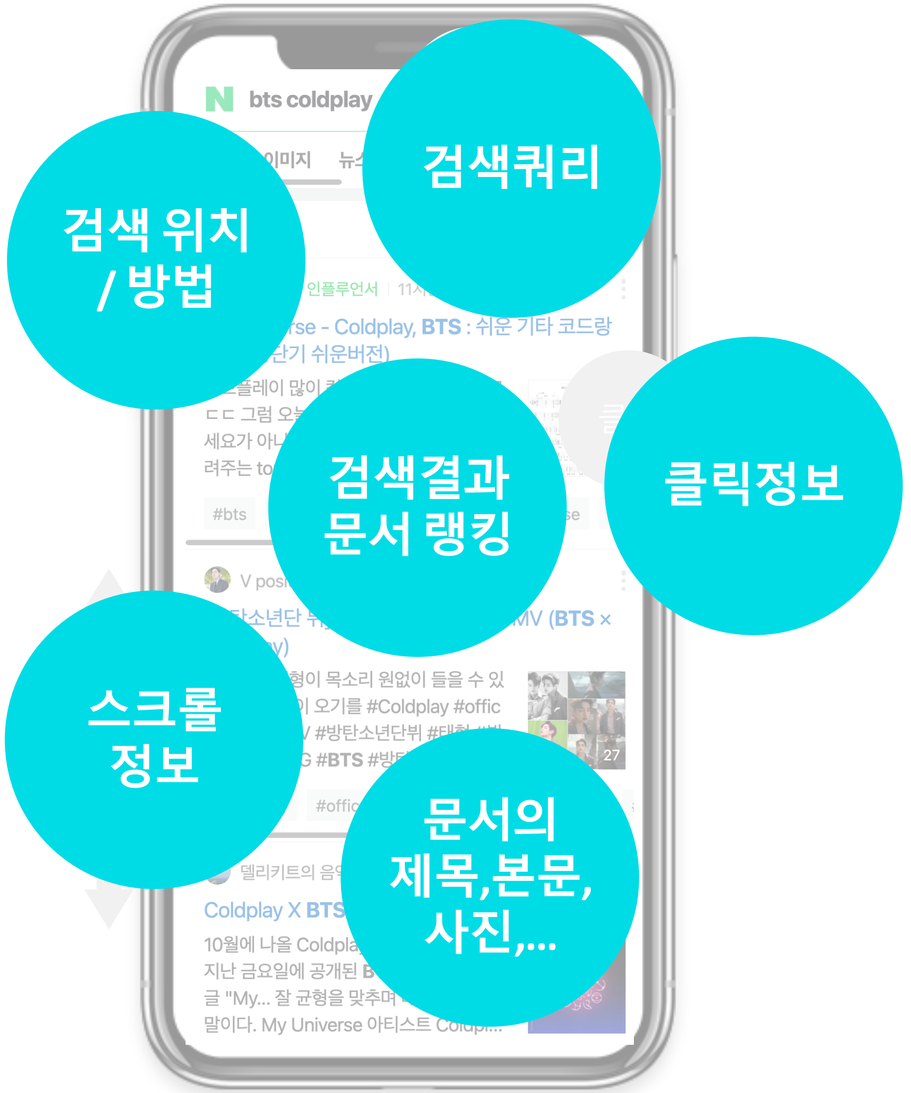
- 쿼리를 전달하여 검색 수행
- 검색결과를 검토, 원하는 문서를 파악
- 스크롤하여 추가적인 결과 확인
- 클릭하여 내가 원하는 문서인지 검토

1.2. 검색 사용자는 어떤 흔적을 남길까?

DEVIEW 2021

앞서 살펴본 사용자의 행위는
다양한 형태로 로깅됨

- 검색 쿼리, 검색 결과 및
  검색한 여러 정황 등에 대한 자세한 로그
- 클릭 행위에 대한 자세한 로그
- 스크롤 행위에 대한 자세한 로그



1.2. 검색 사용자는 어떤 흔적을 남길까?

DEVIEW 2021

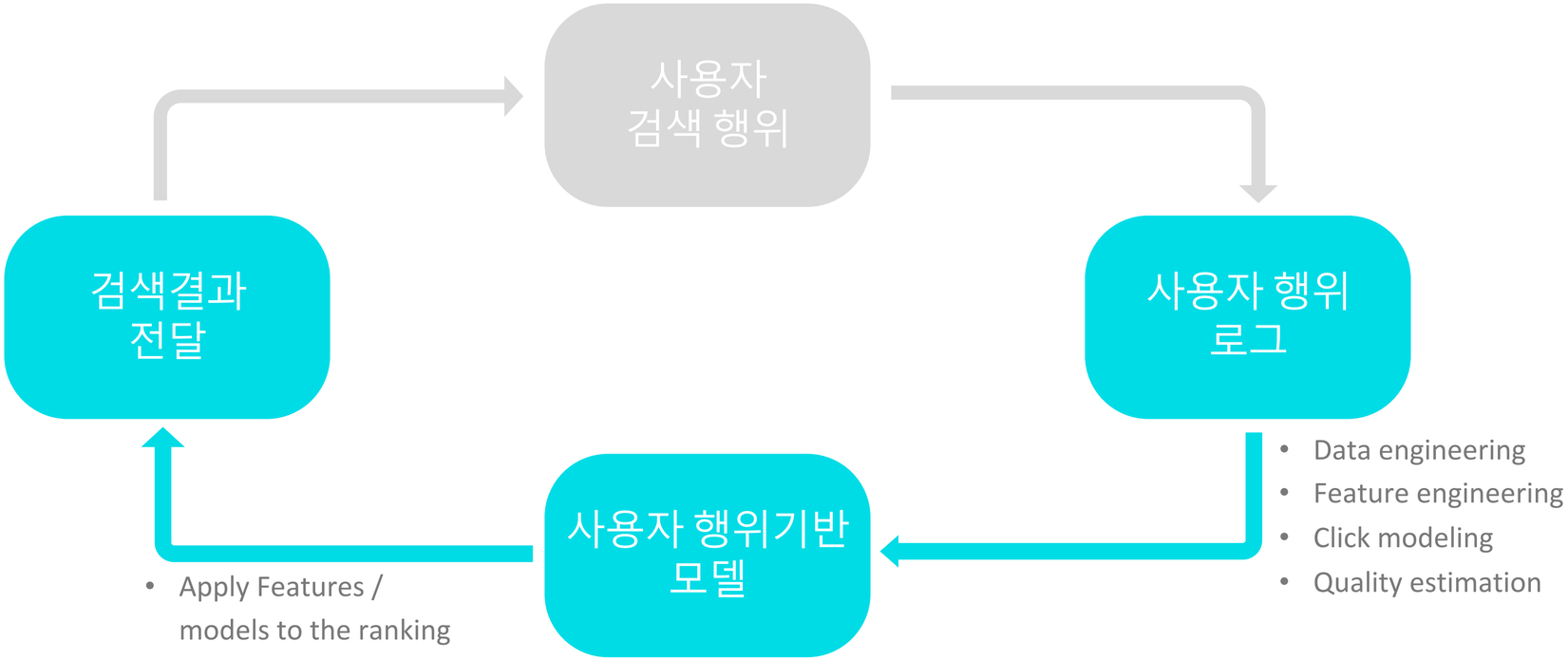
하루 약 30억건의 사용자 로그가 저장

- 네이버 자체구축 시스템을 통해 안전하고 신속하게 저장/유통됨

[네이버 로그 유통과정 단순화]

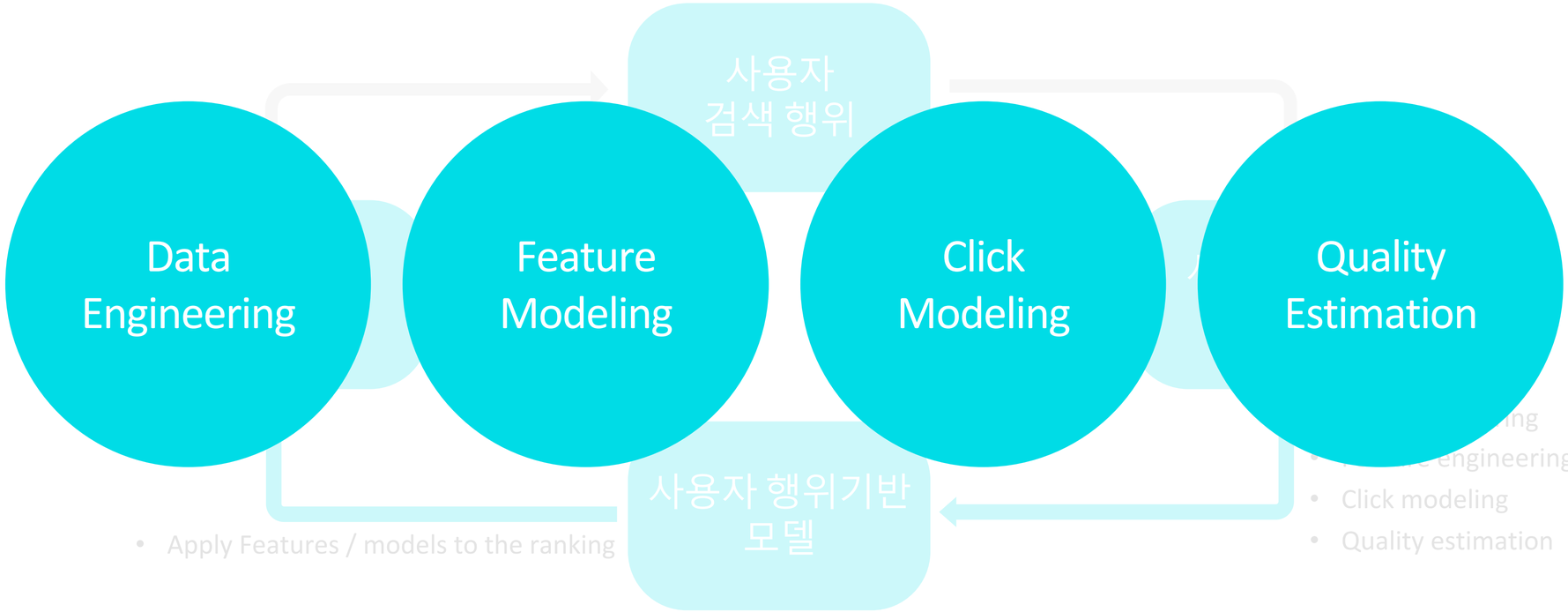
1.3. 검색 로그 분석팀은

사용자 로그에서 의미를 발견하고,
이를 검색랭킹에 기여하는 조직



1.3. 검색 로그 분석팀은

사용자 로그에서 의미를 발견하고,
이를 검색랭킹에 기여하는 조직



2. 사용자 행동 모델링

2.1. 사용자 행동 모델링

DEVIEW 2021

사용자 행동 모델링이란?

- 다양한 사용자 로그를 기반으로, 목적에 맞는 모델을 생성/검증
- 사용자 행동 모델은 Primitive Feature, Click Model로 구분하여 관리

2.1. 사용자 행동 모델링

Primitive Feature, Click Model 설명

	Primitive Feature	Click Model
데이터셋	사용자 로그	사용자 로그
모델링 대상	문서의 노출빈도, 클릭빈도 등	사용자 행동 자체를 모델링
모델링 기법	간단한 수식을 통한 연산	PGM 기반 확률 모델 Neural Net 모델

2.2. Primitive Feature

Primitive Feature, Click Model 설명

	Primitive Feature	Click Model
데이터셋	**사용자 로그**	사용자 로그
모델링 대상	**문서의 노출빈도, 클릭빈도 등**	사용자 행동 자체를 모델링
모델링 기법	**간단한 수식을 통한 연산**	PGM 기반 확률 모델 Neural Net 모델

2.2. Primitive Feature

Primitive Feature 예시

- 문서의 클릭률 : CTR (Click Through Rate)
- 문서 클릭 후, 결과 페이지에 머무른 시간 : DWELL TIME
- 해당 클릭이 몇번째의 클릭인지 : NUMBER OF HOPS
- 문서를 불만족스럽게 본 비율 : BOUNCE RATE

2.2. Primitive Feature

DEVIEW 2021

Dwell time 샘플

- www.daum.net 을 클릭한 대상질의의 Dwell time을 추출한 샘플
- 사용자가 원하는 문서와 연관이 높은 경우, Dwell time의 값이 비교적 높음
- 검색어와 문서의 연관도를 상당히 잘 나타냄

Query	Dwell time
라이코스한메일	**1.0000**
한메일닷컴	**1.0000**
다음포털	**1.0000**
다응넷	**0.9429**
www,daum	**0.8006**
다으0ㅇ	**0.6953**
피아노애드립책	0.2874
따음	0.2086
다음티비	0.2071
mnet보는방법	0.2029

2.2. Primitive Feature

어느 feature가 좋을까? feature의 평가지표

- DCG
  - 문서와 쿼리의 연관도(CG)를 랭킹 위치에 따라 패널티를 부여한 값
  - 한마디로, 랭킹의 퀄리티를 나타내는 점수
- NDCG
  - 모델이 예측한 DCG를 정답셋의 DCG로 정규화 한 값
  - 정답셋에 대한 랭킹의 퀄리티를 나타낸 점수
- NDCG를 중요한 평가지표로 활용하고 있음

2.2. Primitive Feature

Primitive Feature 성능평가

- 자체 평가데이터 기준, Baseline대비 NDCG 0.29~0.57%의 상승률을 보임
- Feature Importance 기준 상위 30% 이내의 중요도

Primitive Feature	NDCG Improvement Rate
CTR	+ 0.29%
DWELL TIME	+ 0.49%
NUMBER OF HOPS	+ 0.45%
BOUNCE RATE	+ 0.57%

2.3. Click Model 이란?

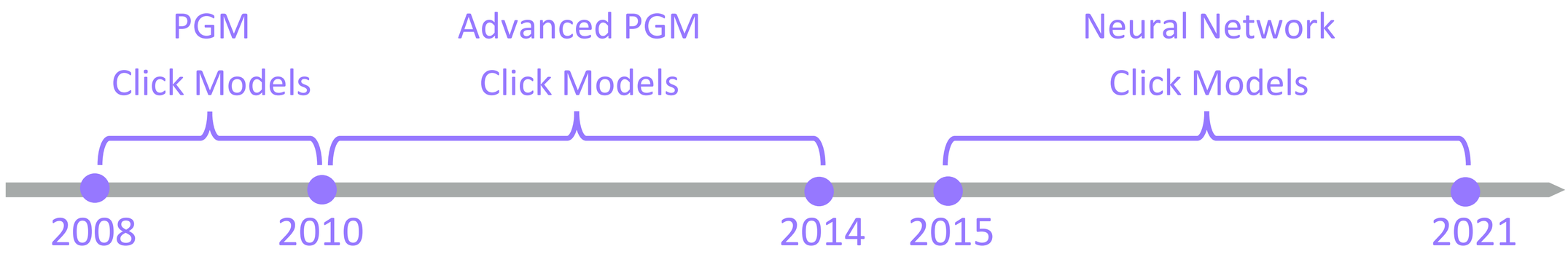
Primitive Feature, Click Model 설명

	Primitive Feature	Click Model
데이터셋	사용자 로그	**사용자 로그**
모델링 대상	문서의 노출빈도, 클릭빈도 등	**사용자 행동 자체를 모델링**
모델링 기법	간단한 수식을 통한 연산	**PGM 기반 확률 모델** **Neural Net 모델**

2.3. Click Model 이란?

2008 ~ 현재까지 사용자 행동을 모델링하는 다양한 연구 진행중

- Microsoft, Yandex 등이 연구 주체
- KDD, SIGIR, WSDM, CIKM 등을 통해 논문 발표

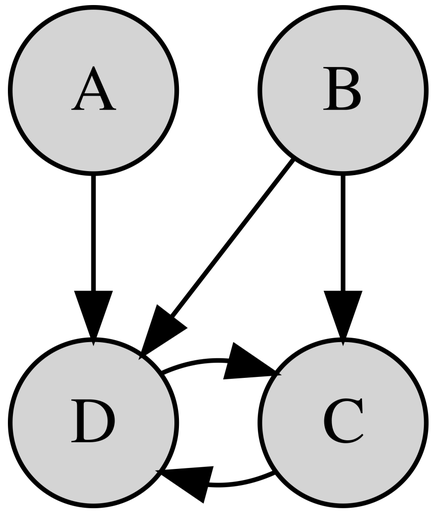


[Click Model 관련 연구논문 Timeline]

2.4. PGM Click Model

PGM(Probabilistic Graphical Model) 모델이란?

- 여러 변수간의 의존성을 그래프로 표현한 확률모델
- 확률이론, 베이지안 통계분석 및 ML에 폭넓게 활용됨



[PGM 모델 예시]

출처: https://en.wikipedia.org/wiki/Graphical_model

2.4. PGM Click Model

DEVIEW 2021

PGM기반 Click Model?

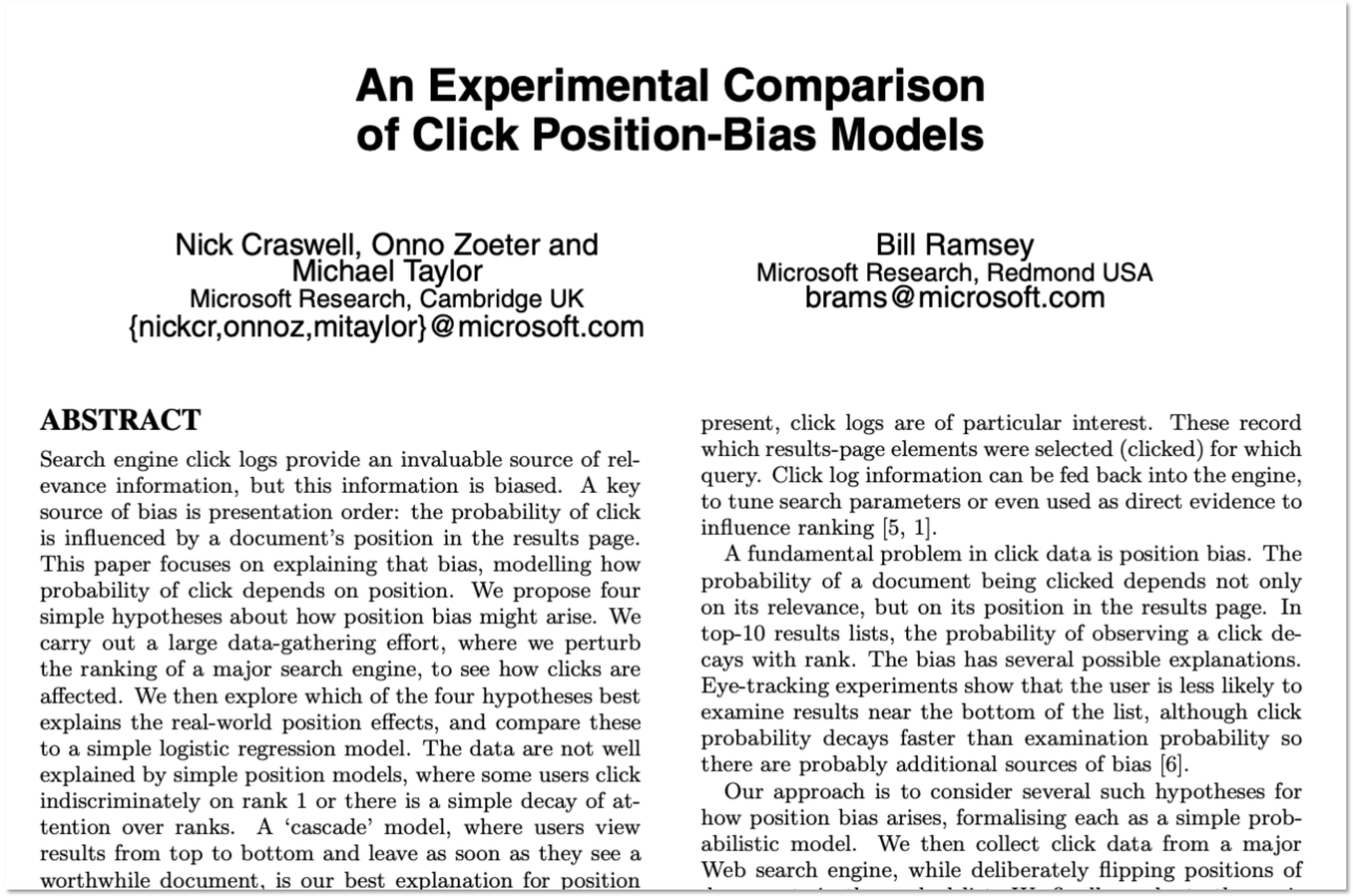
- 사용자 행위를 PGM 기반으로 정의
- 이를 활용해, 사용자의 여러 행위를 계산하는 확률 모델

PGM Click Model 소개

- PBM (Position Based Model, WSDM 2008)
- UBM (User Browsing Model, SIGIR 2009)
- DBN (Dynamic Bayesian Network Model, WWW 2009)

2.4.1. PGM Click Model : PBM

PBM은 2008년 WSDM에서 아래 논문에 의해 제안됨



An Experimental Comparison of Click Position-Bias Models

Nick Craswell, Onno Zoeter and Michael Taylor
Microsoft Research, Cambridge UK
{nickcr,onnoz,mitaylor}@microsoft.com

Bill Ramsey
Microsoft Research, Redmond USA
brams@microsoft.com

ABSTRACT

Search engine click logs provide an invaluable source of relevance information, but this information is biased. A key source of bias is presentation order: the probability of click is influenced by a document's position in the results page. This paper focuses on explaining that bias, modelling how probability of click depends on position. We propose four simple hypotheses about how position bias might arise. We carry out a large data-gathering effort, where we perturb the ranking of a major search engine, to see how clicks are affected. We then explore which of the four hypotheses best explains the real-world position effects, and compare these to a simple logistic regression model. The data are not well explained by simple position models, where some users click indiscriminately on rank 1 or there is a simple decay of attention over ranks. A 'cascade' model, where users view results from top to bottom and leave as soon as they see a worthwhile document, is our best explanation for position

present, click logs are of particular interest. These record which results-page elements were selected (clicked) for which query. Click log information can be fed back into the engine, to tune search parameters or even used as direct evidence to influence ranking [5, 1].

A fundamental problem in click data is position bias. The probability of a document being clicked depends not only on its relevance, but on its position in the results page. In top-10 results lists, the probability of observing a click decays with rank. The bias has several possible explanations. Eye-tracking experiments show that the user is less likely to examine results near the bottom of the list, although click probability decays faster than examination probability so there are probably additional sources of bias [6].

Our approach is to consider several such hypotheses for how position bias arises, formalising each as a simple probabilistic model. We then collect click data from a major Web search engine, while deliberately flipping positions of

논문출처
An experimental comparison of click position-bias models, WSDM 2008

2.4.1. PGM Click Model : PBM

PBM(Position Based Model) 이란?

- 사용자가 문서를 확인하고 그 문서가 매력적일 때,
  사용자는 해당 문서를 클릭 한다는 가정에서 출발
- 사용자가 문서를 확인하는 행동은
  해당 문서의 위치 (랭킹)에 영향을 받는다는 정의 포함
- 위의 가정을 PGM 문제로 정의, 해결하는 모델

2.4.1. PGM Click Model : PBM

PBM Notations

- u : Url
- r : Rank
- q : Query
- $P(C)$: Click Probability
- $P(E)$: Examination Probability
- $P(A)$: Attractiveness Probability

자료출처: Click Models For Web Search
(https://clickmodels.weebly.com/the-book.html)

2.4.1. PGM Click Model : PBM

PBM Definition

문서를 클릭할 확률 문서를 확인할 확률 문서가 매력적일 확률

$$P(C_u = 1) = P(E_u = 1) \cdot P(A_u = 1)$$

문서가 매력적일 확률

$$P(A_u = 1) = \alpha_{uq}$$

문서를 확인할 확률

$$P(E_u = 1) = \gamma_{r_u}$$

● 학습 파라미터

[PBM Definition of parameters]

자료출처: Click Models For Web Search
(https://clickmodels.weebly.com/the-book.html)

2.4.1. PGM Click Model : PBM

EM 알고리즘을 활용한 PBM 파라미터 학습 (1/3)

- EM (Expectation-maximization)은

  관측되지 않는 변수에 의존하는 확률모델에서,

  MLE(Maximum-likelihood)를 갖는 모수의 추정값을 찾는 알고리즘

  (from 위키피디아)

2.4.1. PGM Click Model : PBM

EM 알고리즘을 활용한 PBM 파라미터 학습 (2/3)

- PBM에서 관측되지 않는 변수란?
  - 검색로그에서 바로 구해낼 수 없는 변수 `$\alpha_{uq}, \gamma_{r_u}$`

- PBM에서 MLE(Maximum Likelihood)를 갖는 모수의 추정값이란?
  - 클릭할 확률 $P(C_u = 1) = P(E_u = 1) \cdot P(A_u = 1)$ 을
    잘 예측할 수 있는 적절한 `$\alpha_{uq}, \gamma_{r_u}$`

2.4.1. PGM Click Model : PBM

EM 알고리즘을 활용한 PBM 파라미터 학습 (3/3)

- EM 알고리즘은 구현 복잡도가 상당히 높음
- 이를 비교적 단순하며, 비슷한 성능인 Online-EM으로 대체 구현

Expectation Step

$$
\begin{aligned}
Q(\theta_c) &= \sum_{s \in \mathcal{S}} \mathbb{E}_{\mathbf{X}|\mathbf{C}^{(s)},\mathbf{\Psi}} \left[\log P\left(\mathbf{X}, \mathbf{C}^{(s)} \mid \mathbf{\Psi}\right) \right] \\
&= \sum_{s \in \mathcal{S}} \mathbb{E}_{\mathbf{X}|\mathbf{C}^{(s)},\mathbf{\Psi}} \left[\sum_{c_i \in s} \left(\mathcal{I}\left(X_{c_i}^{(s)} = 1, \mathcal{P}(X_{c_i}^{(s)}) = \mathbf{p}\right) \log(\theta_c) + \right. \right. \\
&\qquad \left. \left. \mathcal{I}\left(X_{c_i}^{(s)} = 0, \mathcal{P}(X_{c_i}^{(s)}) = \mathbf{p}\right) \log(1 - \theta_c) \right) + \mathcal{Z} \right] \\
&= \sum_{s \in \mathcal{S}} \sum_{c_i \in s} \left(P\left(X_{c_i}^{(s)} = 1, \mathcal{P}(X_{c_i}^{(s)}) = \mathbf{p} \mid \mathbf{C}^{(s)}, \mathbf{\Psi}\right) \log(\theta_c) + \right. \\
&\qquad \left. P\left(X_{c_i}^{(s)} = 0, \mathcal{P}(X_{c_i}^{(s)}) = \mathbf{p} \mid \mathbf{C}^{(s)}, \mathbf{\Psi}\right) \log(1 - \theta_c) \right) + \mathcal{Z},
\end{aligned}
$$

Maximization Step

$$
\begin{aligned}
&\frac{\partial Q(\theta_c)}{\partial \theta_c} \\
&= \sum_{s \in \mathcal{S}} \sum_{c_i \in s} \left(\frac{P\left(X_{c_i}^{(s)} = 1, \mathcal{P}(X_{c_i}^{(s)}) = \mathbf{p} \mid \mathbf{C}^{(s)}, \mathbf{\Psi}\right)}{\theta_c} - \frac{P\left(X_{c_i}^{(s)} = 0, \mathcal{P}(X_{c_i}^{(s)}) = \mathbf{p} \mid \mathbf{C}^{(s)}, \mathbf{\Psi}\right)}{1 - \theta_c} \right)
\end{aligned}
$$

Update parameter θ
Iteratively

$$\theta^{(k+1)} = \frac{1}{|\mathcal{S}_X|} \sum_{s \in \mathcal{S}_X} P(X^{(s)} = 1 \mid \mathbf{C}^{(s)}, \mathbf{\Psi}^{(k)})$$

논문출처
Online Expectation-Maximization for Click Models, CIKM 2017

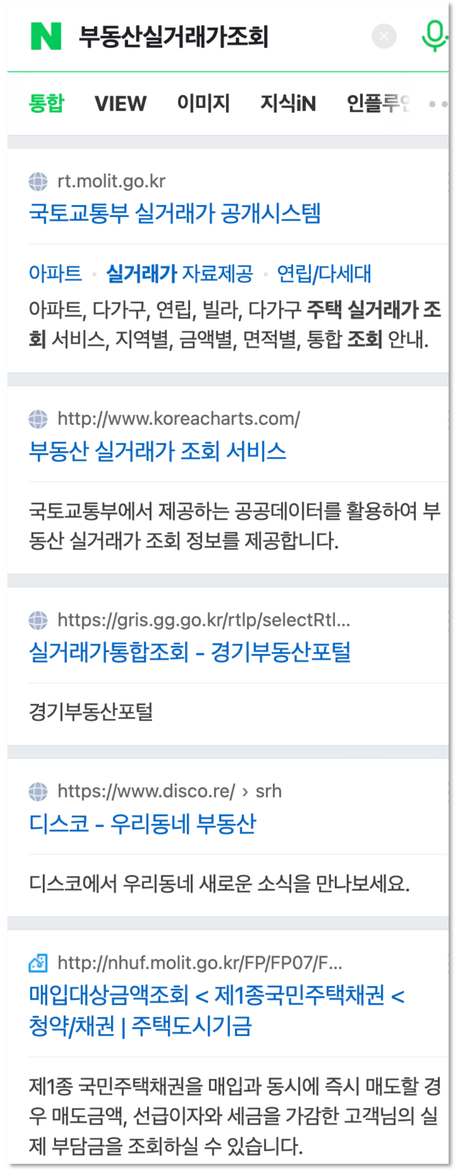
2.4.1. PGM Click Model : PBM

PBM 예시

- 학습된 파라미터를 검색어-문서기준
  추출한 샘플결과 (검색어 : 부동산실거래가조회)

- 검색결과 위치에 의존적임

- 주로 1등(혹은 상위)문서의 경우,
  높은 상황이 많이 존재



PBM
0.7116
0.0781
0.0049
0.0014
0.0006

2.4.2. PGM Click Model : UBM

UBM은 2008년 SIGIR에서 아래 논문에 의해 제안됨

A User Browsing Model to Predict Search Engine Click Data from Past Observations.

Georges Dupret
Yahoo! Research Latin America
gdupret@yahoo-inc.com

Benjamin Piwowarski
Yahoo! Research Latin America
bpiwowar@yahoo-inc.com

ABSTRACT

Search engine click logs provide an invaluable source of relevance information but this information is biased because we ignore which documents from the result list the users have actually seen before and after they clicked. Otherwise, we could estimate document relevance by simple counting. In this paper, we propose a set of assumptions on user browsing behavior that allows the estimation of the probability that a document is seen, thereby providing an unbiased estimate of document relevance. To train, test and compare our model to the best alternatives described in the Literature, we gather a large set of real data and proceed to an extensive cross-validation experiment. Our solution outperforms very significantly all previous models. As a side effect, we gain insight into the browsing behavior of users and we can compare it to the conclusions of an eye-tracking experiments by Joachims et al. [12]. In particular, our findings

are increasingly understood to be the driving force of the Internet and many initiatives are aimed at empowering them. Arguably, this is a long term trend that started with Kleinberg idea of Hubs and Authorities, which proposed that a hyperlink from one document to another was a vote in favor of the document linked to, an idea in practice exploited in the Pagerank algorithm.

Social search, as its name implies, supposes participation from users who tag, bookmark, and comment their search results. In addition to this information explicitly provided by users, there is a much larger source of implicit data which is collected by search engines. This feedback provides detailed and valuable information about users interactions with the system as the issued query, the presented URLs, the selected documents and their ranking. It is a poll of millions of users over an enormous variety of topics. It has been used in many ways to mine user interests and preferences. Examples of applications include Web personalization, Web spam

논문출처
A User Browsing Model to Predict Search Engine Click Data from Past Observations, sigir 2008

2.4.2. PGM Click Model : UBM

DEVIEW 2021

UBM(User Browsing Model) 이란?

- 기본적인 정의는 PBM과 동일함

- 다만, 사용자가 문서를 확인하는 행동이
  해당 문서 이전에 클릭한 위치(랭킹)에도 영향을
  받는다는 정의 포함

- 위의 가정을 PGM 문제로 정의, 해결하는 모델

2.4.2. PGM Click Model : UBM

UBM 정의

문서를 클릭할 확률 문서를 확인할 확률 문서가 매력적일 확률

$$P(C_u = 1) = P(E_u = 1) \cdot P(A_u = 1)$$

문서가 매력적일 확률

$$P(A_u = 1) = \alpha_{uq}$$

문서를 확인하고 이전의 클릭한 가장 높은 랭킹이 r′ 등일 때

$$P(E_r = 1 \mid C_1 = c_1, \ldots, C_{r-1} = c_{r-1}) = \gamma_{rr'}$$

$$r' = \max\{k \in \{0, \ldots, r-1\} : c_k = 1\}$$

● 학습 파라미터

[UBM Definition of parameters]

자료출처: Click Models For Web Search
(https://clickmodels.weebly.com/the-book.html)

2.4.2. PGM Click Model : UBM

UBM Parameter 학습

- Online-EM을 활용하여 진행

$$\theta^{(k+1)} = \frac{1}{|\mathcal{S}_X|} \sum_{s \in \mathcal{S}_X} P(X^{(s)} = 1 \mid \mathbf{C}^{(s)}, \mathbf{\Psi}^{(k)})$$

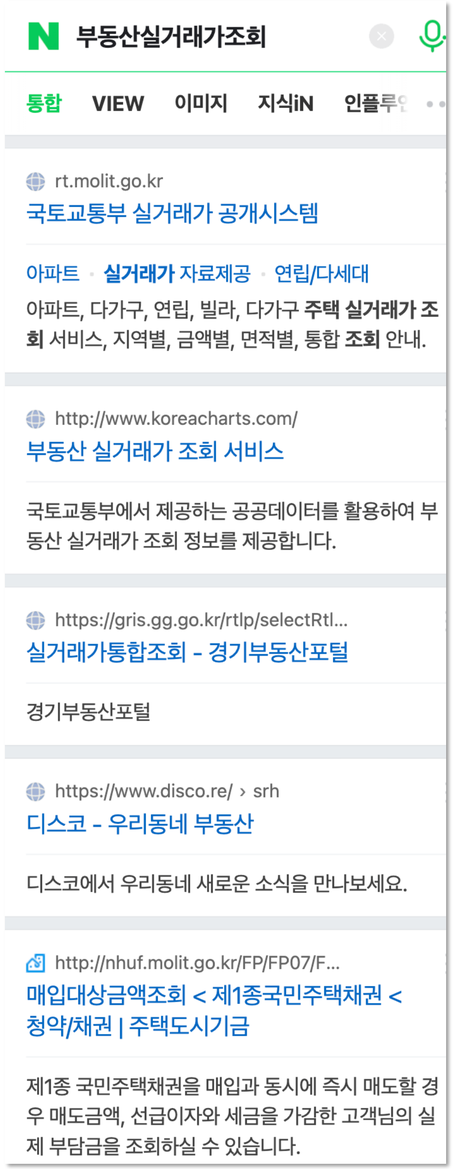
논문출처
Online Expectation-Maximization for Click Models,
CIKM 2017

2.4.2. PGM Click Model : UBM

DEVIEW 2021

UBM 예시

- 학습된 파라미터를 검색어-문서기준 추출한 샘플결과 (검색어 : 부동산실거래가조회)
- 1등문서 뿐 아니라, 사용자가 연속적으로 클릭한 관련 여러 문서들에도 영향
- 이전 클릭문서의 랭킹도 함께 고려하여 학습한 결과



	UBM	PBM
	0.8101	0.7116
	0.5342	0.0781
	0.0539	0.0049
	0.0232	0.0014
	0.0098	0.0006

2.4.3. PGM Click Model : DBN

DBN은 2009년 WWW에서 아래 논문에서 제안됨

A Dynamic Bayesian Network Click Model for Web Search Ranking

Olivier Chapelle
Yahoo! Labs
Santa Clara, CA
chap@yahoo-inc.com

Ya Zhang
Yahoo! Labs
Santa Clara, CA
yazhang@yahoo-inc.com

ABSTRACT

As with any application of machine learning, web search ranking requires labeled data. The labels usually come in the form of relevance assessments made by editors. Click logs can also provide an important source of implicit feedback and can be used as a cheap proxy for editorial labels. The main difficulty however comes from the so called *position bias* — urls appearing in lower positions are less likely to be clicked even if they are relevant. In this paper, we propose a Dynamic Bayesian Network which aims at providing us with unbiased estimation of the relevance from the click logs. Experiments show that the proposed click model outperforms other existing click models in predicting both click-through rate and relevance.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]; H.3.5 [**Online Information Services**]; I.2.6 [**ARTIFICIAL INTELLI-**

most up-to-date documents/sites to the users. However, it would be prohibitive to keep all the relevance labels up to date. Click logs embed important information about user satisfaction with a search engine and can provide a highly valuable source of relevance information. Compare to editorial labels, clicks are much cheaper to obtain and always reflect current relevance.

Clicks have been used in multiple ways by a search engine: to tune search parameters, to evaluate different ranking functions [7, 13, 14, 15], or as signals to directly influence ranking [1, 13]. However, clicks are known to be biased, by the presentation order, the appearance (e.g. title and abstract) of the documents, and the reputation of individual sites. Many studies [8, 10] have attempted to account the position-bias of click. Carterette and Jones [7] proposed to model the relationship between clicks and relevance so that clicks can be used to unbiasedly evaluate search engine when lack of editorial relevance judgment. Other research [10, 21, 16] attempted to model user click behavior during search

논문출처
A dynamic bayesian network click model for web search ranking, WWW 2009

2.4.3. PGM Click Model : DBN

DEVIEW 2021

DBN(Dynamic Bayesian Network Model) 이란?

- 다른 모델과 동일하게 사용자가 문서를 확인하고
  그 문서가 매력적일 때, 해당 문서를 클릭 한다는 가정에서 출발
- 다만, 사용자가 해당 문서를 얼마나 만족했는지에 따라
  지속적으로 검색을 수행/중단할지 여부 결정
  - 직전 검색결과 만족 : 검색을 중단함
  - 직전 검색결과 불만족 : 검색을 중단 혹은 지속함
- 위의 가정을 PGM 문제로 정의, 해결하는 모델

2.4.3. PGM Click Model : DBN

DEVIEW 2021

DBN Notations

- u : Url
- r : Rank
- q : Query
- $P(C)$: Click Probability
- $P(E)$: Examination Probability
- $P(A)$: Attractiveness Probability
- $P(S)$: **Satisfaction Probability**

자료출처: Click Models For Web Search
(https://clickmodels.weebly.com/the-book.html)

2.4.3. PGM Click Model : DBN

DBN Definition (1/2)

현재(r등)문서를 클릭한다 / 현재(r등) 문서를 확인하고 / 현재(r등) 문서가 매력적이면

$$C_r = 1 \Leftrightarrow E_r = 1 \text{ and } A_r = 1$$

현재(r등) 문서가 매력적일

$$P(A_r = 1) = \alpha_{u_r q}$$

1등 문서는 무조건 확인

$$P(E_1 = 1) = 1$$

이전(r-1등) 문서를 확인하지 않고, 현재(r등) 문서를 확인 X

$$P(E_r = 1 \mid E_{r-1} = 0) = 0$$

● **학습 파라미터**

[DBN Definition of parameters]

자료출처: Click Models For Web Search
(https://clickmodels.weebly.com/the-book.html)

2.4.3. PGM Click Model : DBN

DBN Definition (2/2)

현재(r등)문서를 만족할 확률 / 현재(r등) 문서를 클릭할 때

$$P(S_r = 1 \mid C_r = 1) = \sigma_{u_r q}$$

현재(r등) 문서를 확인 X / 이전(r-1) 문서가 만족스러웠을 때

$$P(E_r = 1 \mid S_{r-1} = 1) = 0$$

현재(r등) 문서를 확인할 확률 / 위의(r-1등) 문서를 확인했고, / 위의(r-1등) 문서가 만족스럽지 않았을 때

$$P(E_r = 1 \mid E_{r-1} = 1, S_{r-1} = 0) = \gamma$$

● **학습 파라미터**

[DBN Definition of parameters]

자료출처: Click Models For Web Search
(https://clickmodels.weebly.com/the-book.html)

2.4.3. PGM Click Model : DBN

DBN Parameter 학습

- Online-EM을 활용하여 진행

$$\theta^{(k+1)} = \frac{1}{|\mathcal{S}_X|} \sum_{s \in \mathcal{S}_X} P(X^{(s)} = 1 \mid \mathbf{C}^{(s)}, \mathbf{\Psi}^{(k)})$$

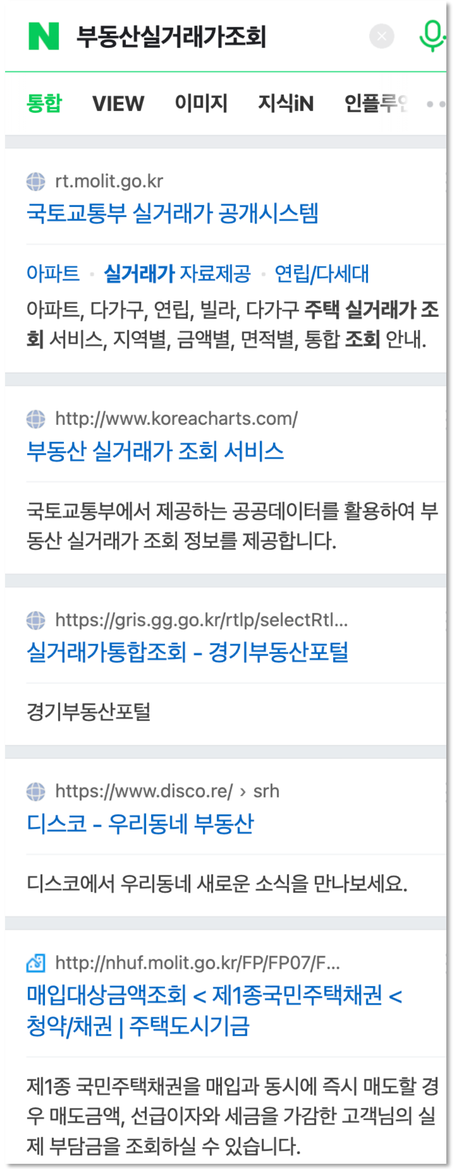
논문출처
Online Expectation-Maximization for Click Models,
CIKM 2017

2.4.3. PGM Click Model : DBN

DEVIEW 2021

DBN 예시

- 학습된 파라미터를 검색어-문서기준 추출한 샘플결과 (검색어 : 부동산실거래가조회)
- 사용자가 어느 문서에서 만족했는지를 잘 보여주는 예시
- 만족/불만족의 경계를 어느정도 확인할 수 있는 특징 존재



DBN	UBM	PBM
0.2401	0.8101	0.7116
0.0278	0.5342	0.0781
0.0017	0.0539	0.0049
0.0005	0.0232	0.0014
0.0001	0.0098	0.0006

2.5 Advanced PGM Click Model

PGM Click Model이 갖는 한계

- 클릭정보가 부족한 문서의 경우 모델의 학습/예측에 활용되지 못함
- 블로그/카페/웹등이 혼재된 통합검색의 특징을 잘 담아내지 못함
- 사용자의 행동에 포함된 노이즈값의 고려 필요

이를 해결하기 위해, Advanced PGM Model 등장

2.5 Advanced PGM Click Model

Advanced PGM 모델 소개

- 각 모델의 구현체, 실제 서비스에 미치는 성능을 연구 중

특징	Advanced PGM Model	Published
사용자 행동에 포함된 노이즈를 고려한 모델	GCM	WSDM, 2010 (Microsoft)
블로그/카페/이미지 등 통합검색의 특징을 고려한 모델	FCM, VCM	WSDM, 2012 (Microsoft)
클릭로그가 부족한 쿼리/문서를 고려한 모델	MFCM	WSDM, 2012 (Microsoft)
쿼리의 Refomulation등을 고려한 모델	TCM, SCM	Yandex

2.6 Neural Click Model

PGM/Advanced PGM 모델의 한계

- PGM 모델은 사람이 정의한 행위를 모델이 학습함
- 이에 따라, 사용자 행위 자체를 모델에게 학습시키는 연구가 진행

Neural Network 기반 Click Model 등장

- RNN 활용 첫 모델
  : NCM (Neural Click Model, WWW 2016)
- 검색어 / 문서 / 클릭행동간 Context를 학습
  : CACM (Context Aware Click Model, WSDM 2020)

2.6.1. Neural Click Model : NCM

NCM은 2016년 WWW에서 아래 논문에 의해 제안됨

A Neural Click Model for Web Search

Alexey Borisov[†, ‡] alborisov@yandex-team.ru
Ilya Markov[‡] i.markov@uva.nl
Maarten de Rijke[‡] derijke@uva.nl
Pavel Serdyukov[†] pavser@yandex-team.ru

[†] Yandex, Moscow, Russia
[‡] University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Understanding user browsing behavior in web search is key to improving web search effectiveness. Many click models have been proposed to explain or predict user clicks on search engine results. They are based on the *probabilistic graphical model* (PGM) framework, in which user behavior is represented as a sequence of observable and hidden events. The PGM framework provides a mathematically solid way to reason about a set of events given some information about other events. But the structure of the dependencies between the events has to be set manually. Different click models use different hand-crafted sets of dependencies.

We propose an alternative based on the idea of distributed representations: to represent the user's information need and the information available to the user with a *vector state*. The components of the vector state are learned to represent concepts that are useful for modeling user behavior. And user behavior is modeled as a sequence of vector states associated with a query session: the vector state is initialized with a query, and then iteratively updated based on information about interactions with the search engine re-

1. INTRODUCTION

Understanding users' interaction behavior with a complex Information Retrieval (IR) system is key to improving its quality. In web search, the ability to accurately predict the behavior of a particular user with a certain information need, formulated as a query, in response to a search engine result page allows search engines to construct result pages that minimize the time that it takes users to satisfy their information needs, or increase the probability that users click on sponsors' advertisements.

Recently, many models have been proposed to explain or predict user behavior in web search; see [10] for an overview. These models, also called *click models* as the main observed user interaction with a search system concerns clicks, are used for click prediction and they may help in cases where we do not have real users to experiment with, or prefer not to experiment with real users for fear of hurting the user experience. Click models are also used to improve document ranking (i.e., infer document relevance from clicks predicted by a click model) [4, 12], improve evaluation metrics (e.g., model-based metrics) [5, 8, 35] and to better understand a user by

논문출처
A Neural Click Model for Web Search, WWW 2016

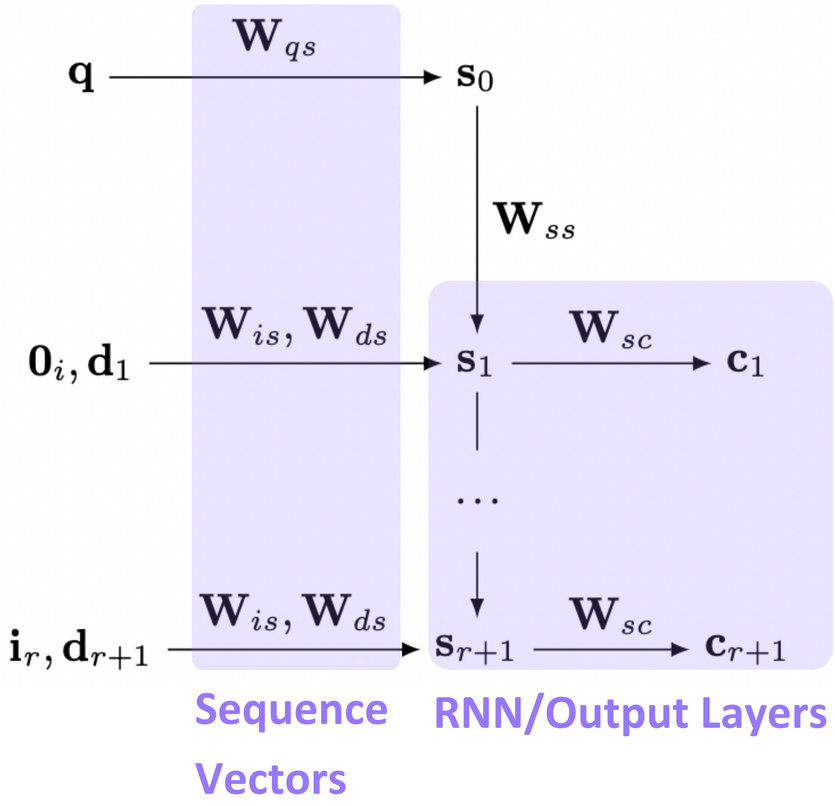
2.6.1. Neural Click Model : NCM

NCM(Neural Click Model) 이란?

- NCM은 사용자 행동을 Sequence Model로 정의 / 해결한 첫 모델
- 검색어 / 문서 / 클릭행동을 Sequence Vector로 변환 및 RNN으로 학습
- 비교적 단순한 Network 구조에 비해, 좋은 성능을 냄

2.6.1. Neural Click Model : NCM

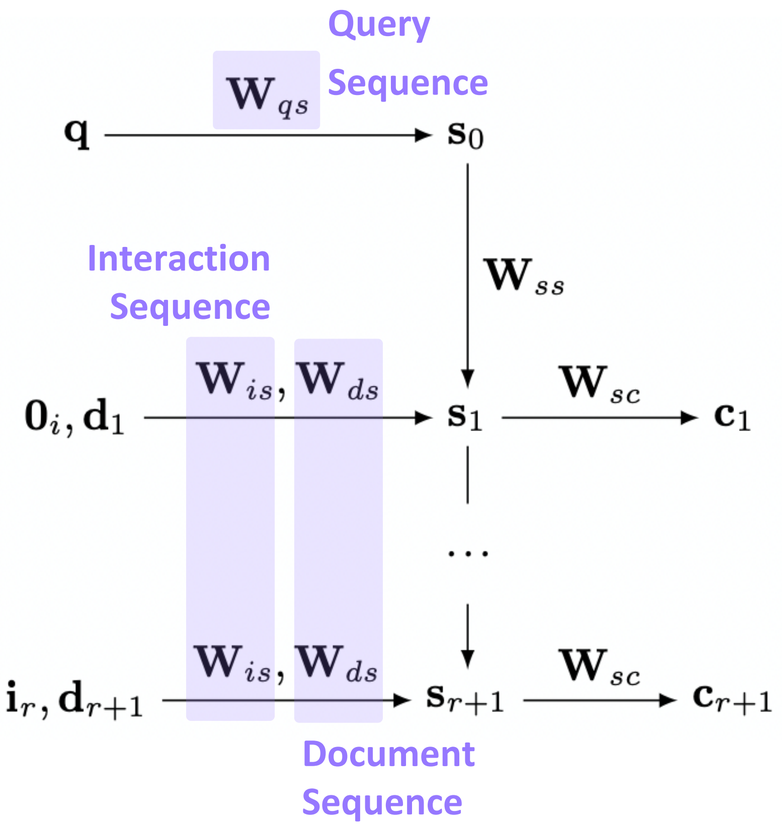
NCM 논문에 소개된 Network architecture



2.6.1. Neural Click Model : NCM

NCM Sequence Vectors

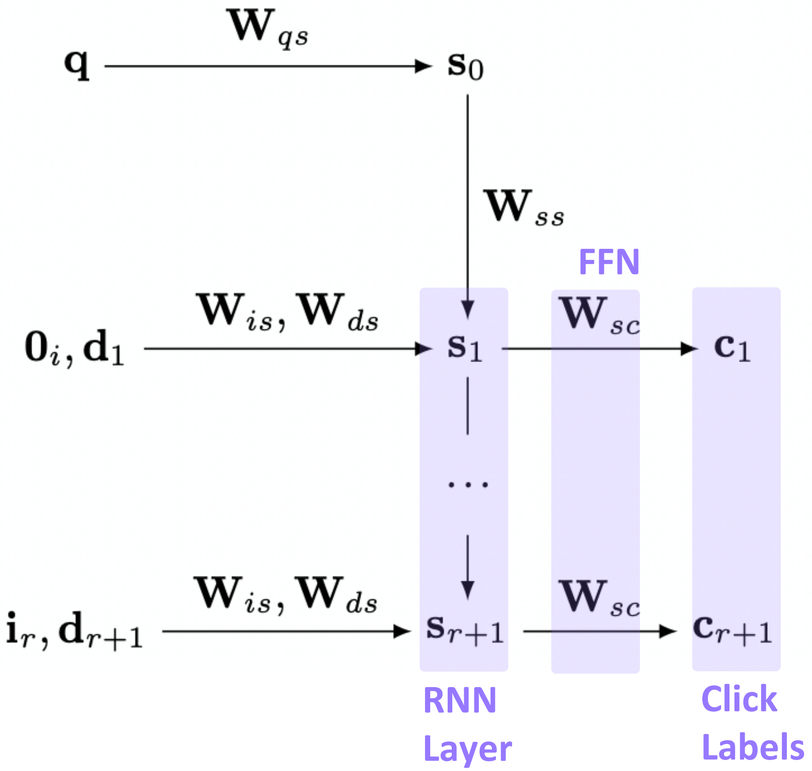
- Query Sequence
  - 검색 세션 내, 검색한 쿼리를 Embedding을 통과시킨 값
- Document Sequence
  - 랭킹순으로 정렬된 검색결과 문서들을 Embedding을 통과시킨 값
- Interaction Sequence
  - 검색결과가 클릭되었는지 여부들을 Embedding을 통과시킨 값



2.6.1. Neural Click Model : NCM

NCM RNN/Output Layers

- RNN layer
  - Sequence vector들을 RNN을 통과하여 학습
- Feedforward Network
  - RNN의 결과를 최종 FFN를 통과,
    최종값 Click probability 생성



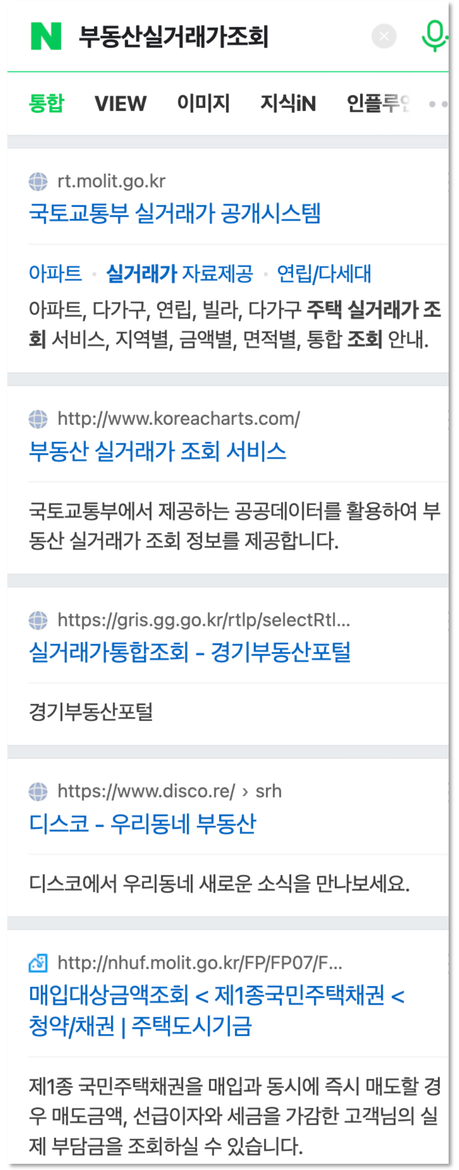
2.6.1. Neural Click Model : NCM

NCM 예시

- 학습된 파라미터를 검색어-문서기준
  추출한 샘플결과
  (검색어 : 부동산실거래가조회)

- UBM+DBN의 전반적인 특징과 비슷하게,
  검색행위의 연속적인 특징을 잘 담아낸 느낌

- 랭킹위치까지 함께 학습하여,
  어느정도 랭킹위치에 bias한 느낌을 줌



NCM
0.3370
0.2113
0.0119
0.0186
0.0001

2.6.2. Neural Click Model : CACM

CACM은 2020년 WSDM에서 아래 논문에 의해 제안됨

A Context-Aware Click Model for Web Search

Jia Chen, Jiaxin Mao, Yiqun Liu*, Min Zhang, Shaoping Ma
Department of Computer Science and Technology, Institute for Artificial Intelligence
Beijing National Research Center for Information Science and Technology
Tsinghua University, Beijing 100084, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

To better exploit the search logs, various click models have been proposed to extract implicit relevance feedback from user clicks. Most traditional click models are based on probability graphical models (PGMs) with manually designed dependencies. Recently, some researchers also adopt neural-based methods to improve the accuracy of click prediction. However, most of the existing click models only model user behavior in query level. As the previous iterations within the session may have an impact on the current search round, we can leverage these behavior signals to better model user behaviors. In this paper, we propose a novel neural-based Context-Aware Click Model (CACM) for Web search. CACM consists of a context-aware relevance estimator and an examination predictor. The relevance estimator utilizes session context information, i.e., the query sequence and clickthrough data, as well as the pre-trained embeddings learned from a session-flow graph to estimate the context-aware relevance of each search result. The examination predictor estimates the examination probability of each result. We further investigate several combination functions to integrate the context-aware relevance and examination probability into click prediction. Experiment results on a public Web search dataset show that CACM outperforms existing click models in both relevance estimation and click prediction tasks.

CCS CONCEPTS

• **Information systems** → *Users and interactive retrieval*; Query

simulation may help search engines better fulfill users' information needs. To this end, numerous *click models* have been proposed for Web search [9]. These click models act as the click simulators in a virtual environment if no real users are available. While click signals are vulnerable due to behavioral biases, e.g., the position bias [24], click models also estimate the unbiased relevance scores for query-document pairs to facilitate document ranking.

Most existing click models represent user behaviors as a sequence of observable or hidden states. They are normally constructed based on the probabilistic graphical model (PGM) framework. Researchers may first analyze the search log data and then manually design the dependencies in the PGM framework, e.g. Position-biased Model (PBM) assumes the examination probability of a document depends heavily on its position in the Search Engine Result Page (SERP) [12]. These model-driven methods can reason about user behavior through the dependencies between the events in PGMs. However, these dependencies have to be set manually and are likely to miss key aspects of user behavior [3, 4].

To better capture users' behavior patterns, Borisov et al. [3] propose a neural click model (NCM) for Web search. Instead of the traditional PGM-based framework, they adopt the distributed representation (DR) approach for user behavior representation. In NCM, user interactions are represented as a vector sequence. Experiment results show that it outperforms traditional click models constructed on the PGM framework. Although NCM has utilized the query-level interaction information, they ignore the interaction effects between different search iterations within a session. It also

논문출처
A Context-Aware Click Model for Web Search, WSDM 2020

2.6.2. Neural Click Model : CACM

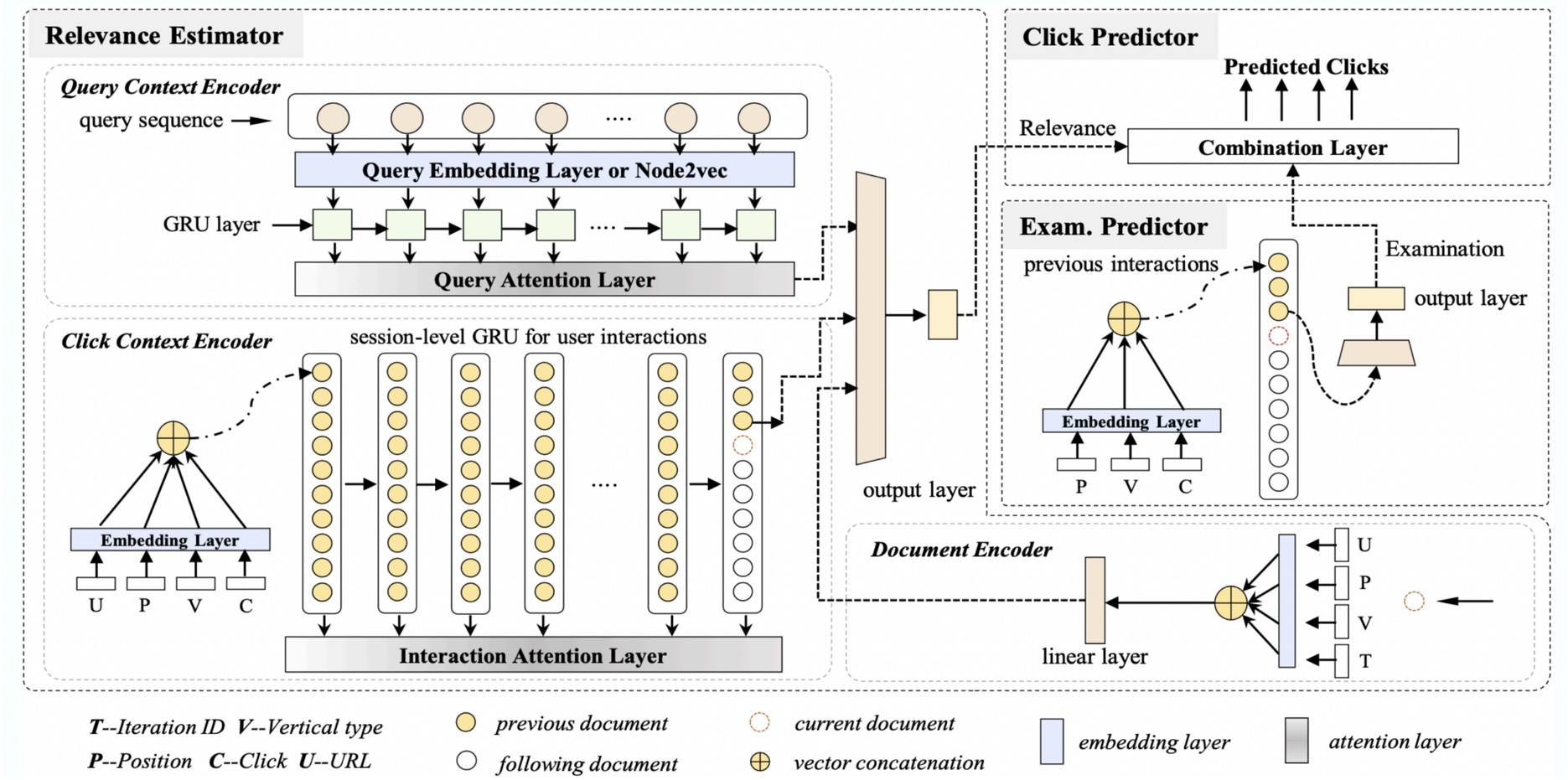
DEVIEW 2021

CACM(Context Aware Click Model) 이란?

- NLP가 발전하며, 주어진 상황의 이해를 위한 여러 연구들이 진행
- CACM은 이 중 Self-attention기법을 활용한 모델로,
  검색어 / 문서 / 클릭행동 간의 context를 학습하는 모델

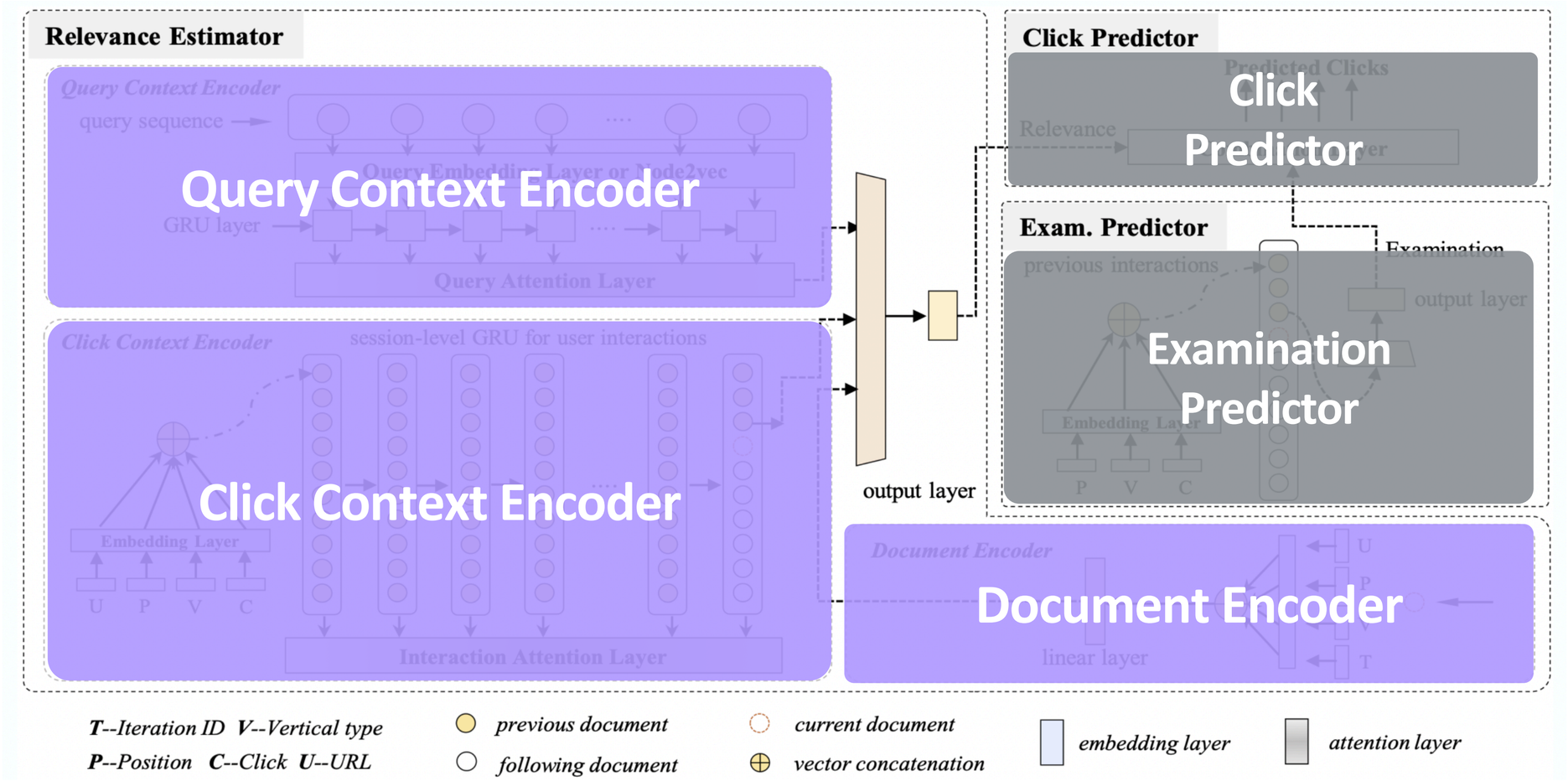
2.6.2. Neural Click Model : CACM

CACM 논문에 소개된 Network architecture



2.6.2. Neural Click Model : CACM

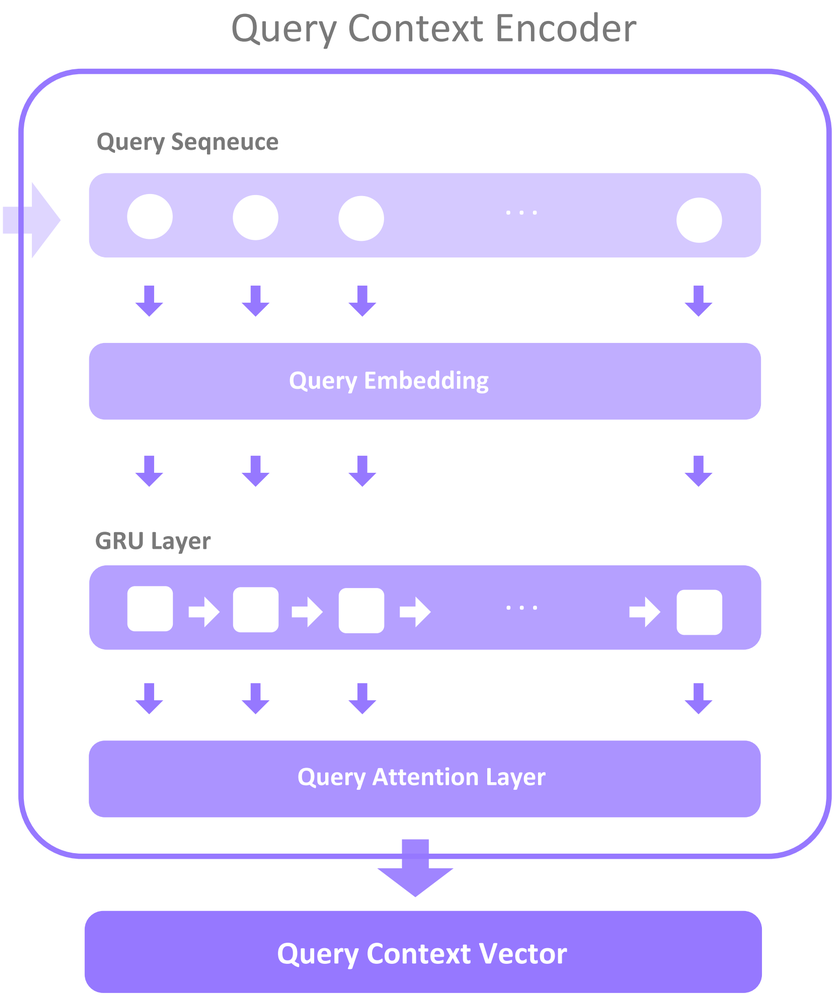
CACM 논문에 소개된 Network architecture



2.6.2. Neural Click Model : CACM

CACM Encoders (1/3)

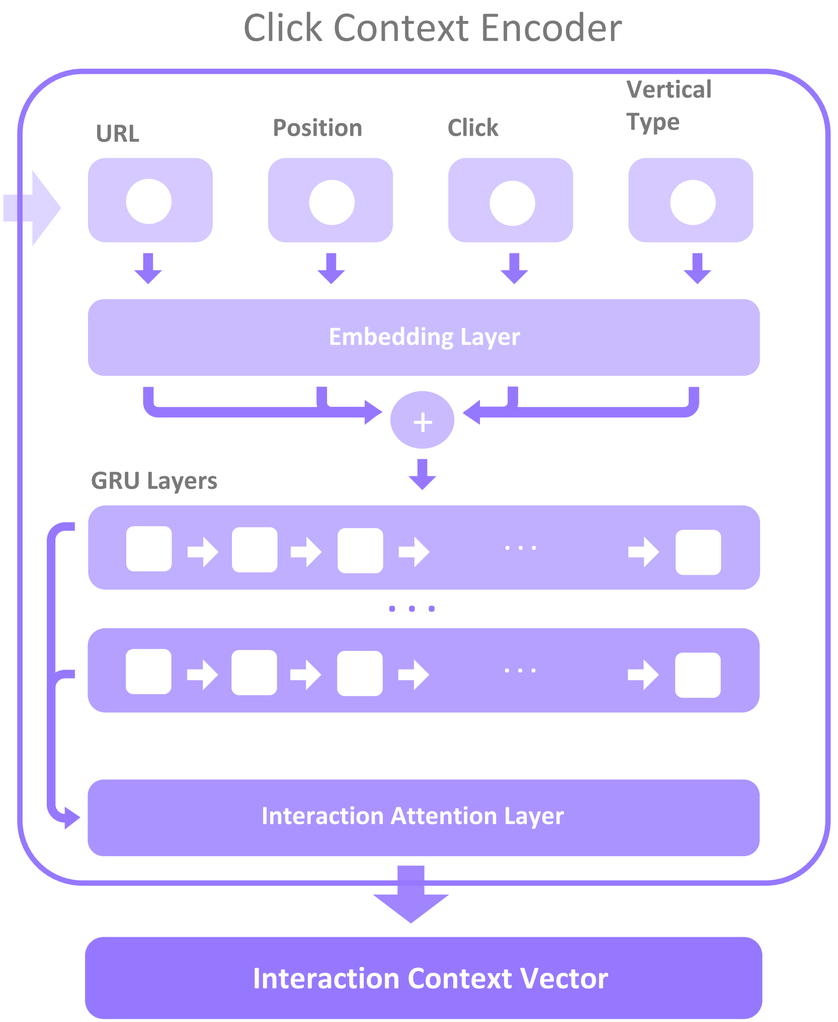
- Query Context Encoder
  - 세션의 검색쿼리를 sequence vector 변환
  - Embedding, GRU Layer,
    Attention Layer 를 차례로 통과
  - Query Context Vector 생성



2.6.2. Neural Click Model : CACM

CACM Encoders (2/3)

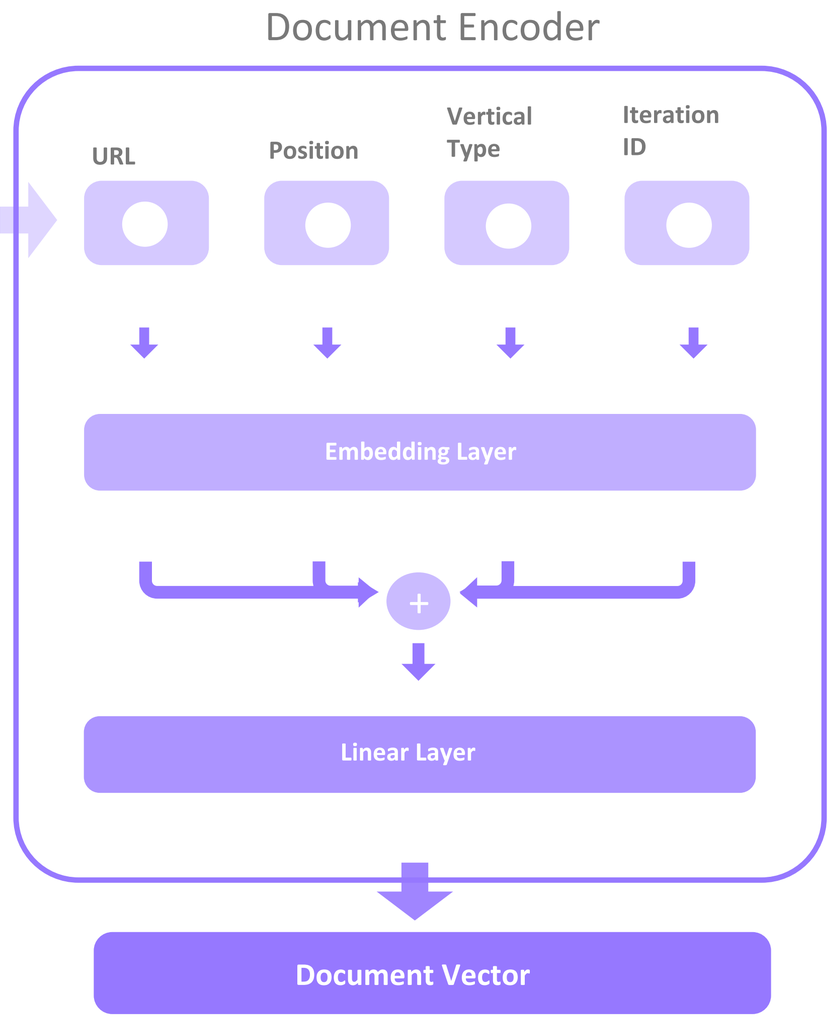
- Click Context Encoder
  - 문서, 랭킹위치, 클릭여부, 버티컬정보 활용
  - Embedding, GRU Layer,
    Attention Layer 를 차례로 통과
  - Interaction Context Vector 생성



2.6.2. Neural Click Model : CACM

CACM Encoders (3/3)

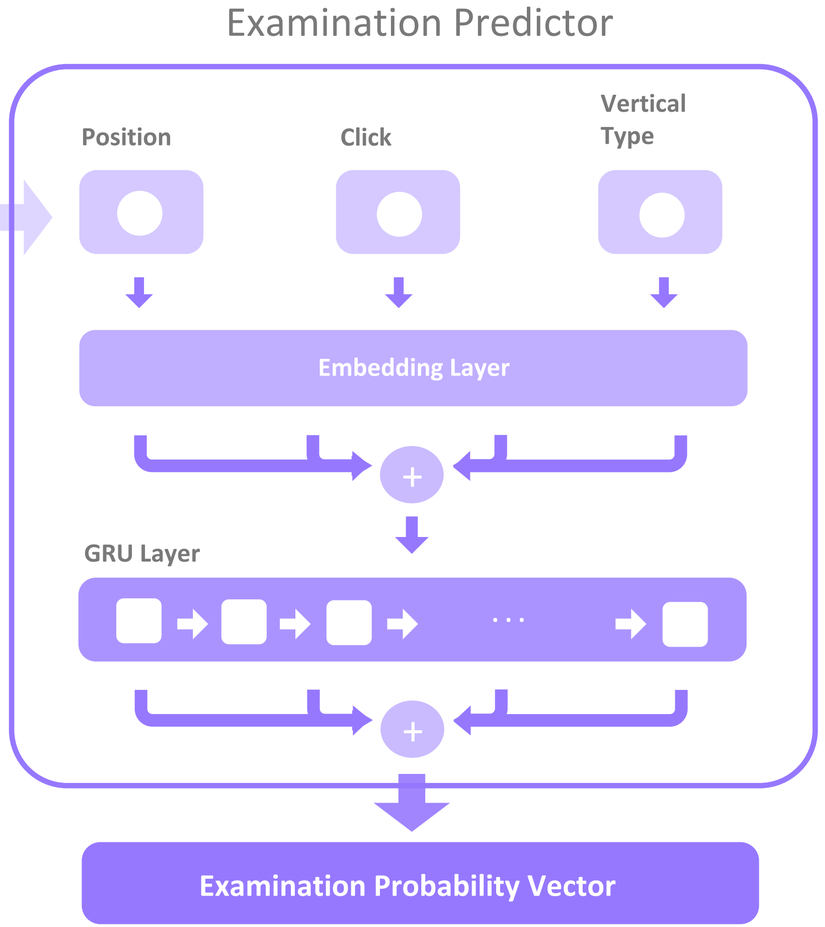
- Document Encoder
  - 문서, 랭킹위치, 버티컬정보, 쿼리순서정보 활용
  - Embedding, Linear Layer를 차례로 통과
  - Document Vector 생성



2.6.2. Neural Click Model : CACM

CACM Predictors (1/2)

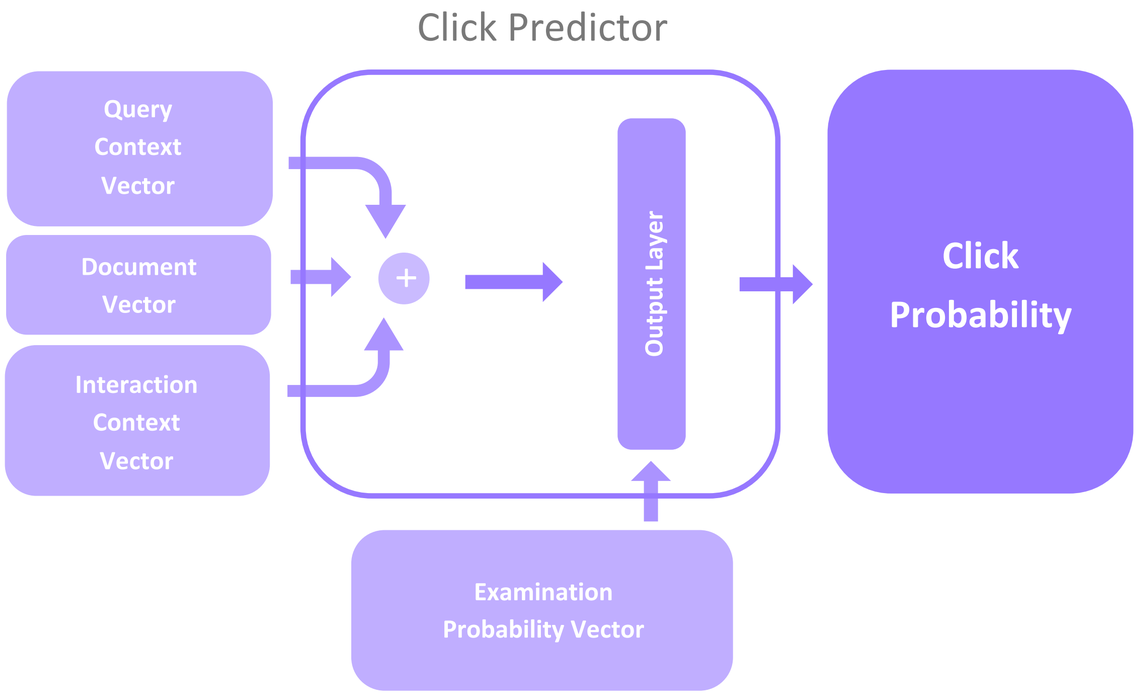
- Examination Predictor
  - 랭킹위치, 클릭정보, 버티컬정보 활용
  - Embedding, GRU Layer를 차례로 통과
  - Examination Probability Vector 생성



2.6.2. Neural Click Model : CACM

CACM Predictors (2/2)

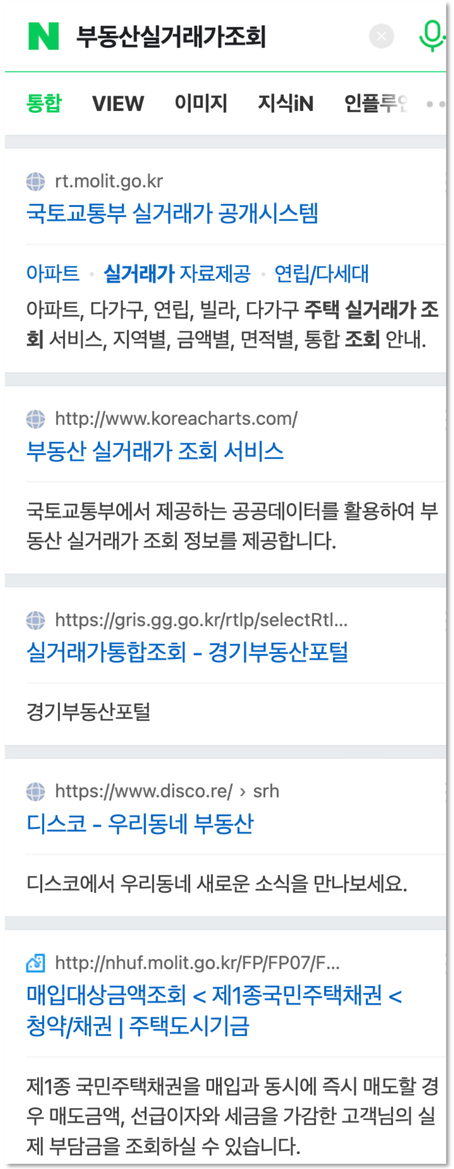
- Click Predictor
  - Encoder의 결과 Vector들,
    Examination Probability Vector를 활용
  - 최종값 Click probability 생성



2.6.2. Neural Click Model : CACM

CACM 예시

- 학습된 파라미터를 검색어-문서기준 추출한 샘플결과 (검색어 : 부동산실거래가조회)
- Attention vector를 사용하여, 랭킹위치에 bias하지 않음 (NCM과는 확연히 다른 상황)
- 다른 모델대비, 값의 분포 및 편차가 큼



	CACM	NCM
	0.9355	0.3370
	0.6525	0.2113
	0.1261	0.0119
	0.3429	0.0186
	0.0000	0.0001

2.7 Click Model Evaluation

DEVIEW 2021

그래서, 살펴본 PGM / NeuralNet 모델 중 가장 좋은 모델은?

- LogLikelihood
  - Measure the goodness of a model to a sample of the parameters. (from 위키피디아)
  - 학습한 모델이 실제 관측 데이터의 샘플과 일관되는 정도. 높을수록 좋음.
- Perplexity
  - How well a probability model predicts a sample. (from 위키피디아)
  - 학습한 확률 모델이 샘플을 얼마나 잘 예측하는지에 대한 값.
  - 1보다 갖거나 크며, 낮을수록 좋음

2.7 Click Model Evaluation

그래서, 어느 모델이 좋은걸까?

- 자체 평가셋 (Clara)을 활용한 평가 결과

	NDCG Avg (@1,3,5,10)	LogLikelihood	Perplexity
PBM	0.7870	-0.0515	1.0566
UBM	0.7720	-0.0497	1.0562
DBN (Simplified version)	0.8261	-0.0494	1.0556
NCM (논문의 80% 구현)	0.9581	-0.0516	1.0570
CACM (논문의 70% 구현)	0.9018	-0.0521	1.0590

2.8 Future work

목적에 맞는 모델을 잘 활용하기 위한 연구

- PGM/Advanced PGM 모델의 목적 및 특징에 따라,
  검색서비스에 존재하는 여러 문제점을 해결하기 위한 노력

- NCM, CACM 등 Neural Net 기반의 모델의 최적화 / 검색 서비스 반영 연구

- 사용자 행동과 관련한 여러 평가지표 고민

2.8 Future work

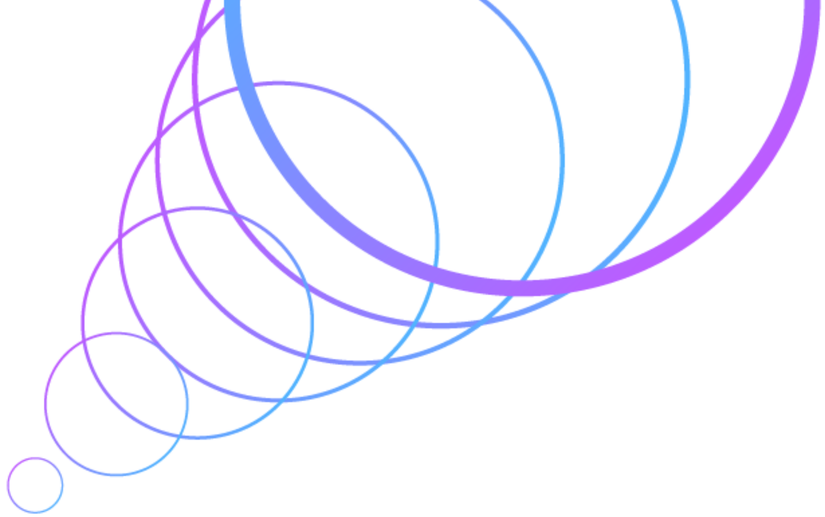
NAVER LABS EUROPE 연구 소개

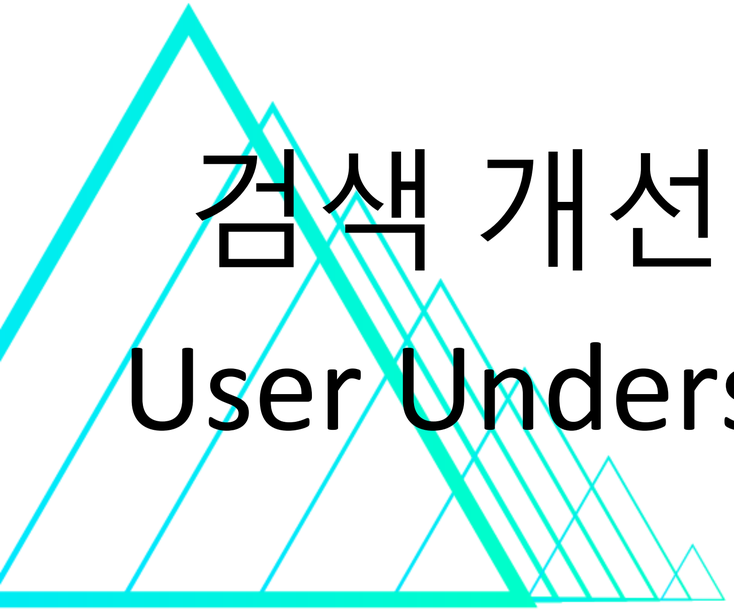
- Click Model은 클릭데이터의 Distributional Shift(분포의 변화)에 민감

- 기존 평가 방법인 Perplexity, NDCG는 Distributional Shift를 측정하기에 부족

- NAVER LABS EUROPE에서는 연구와 논문등을 통해,
  Distributional Shift 대응도 평가 및 Click Model 활용법에 대해 제안

- NAVER 자체 구축 데이터셋인 Clara에 대해 소개함





검색 개선을 위한 검색 사용자 이해하기
User Understanding for Search Enhancement

백준선, 백수현 NAVER

3.
네이버 클릭데이터셋 (CLARA)과 Pyclick 프레임워크

3.1 클릭 데이터셋에 대하여

DEVIEW 2021

클릭 데이터셋이란?

- 클릭 모델의 일관된 평가를 위해 정립된 데이터셋

클릭 데이터셋의 구성요소

- 검색어 (질의 또는 Query)
- 검색 결과 (Search Results 또는 SERP)
- 클릭 정보
- 검색어와 URL간의 관련 점수

3.2 CLARA 예시를 통한 클릭 데이터셋의 이해

항목	예시
Query	NAVER
검색 결과 및Relevance	1. **https://www.naver.com/** (5점) 2. https://www.navercorp.com/ (4점) 3. https://namu.wiki/w/%EB%84%A4%EC (5점) 4. **https://section.cafe.naver.com/ca-fe/** (4점) 5. https://map.naver.com/v5/ (4점)
클릭 정보	Click 1: **https://www.naver.com/** Click 2: **https://section.cafe.naver.com/ca-fe/**

3.3 새로워진 CLARA 데이터셋 : CLARA 2

CLARA 2 에서는 무엇이 달라졌을까?

- 더욱 검색 개선에 적합한 질의의 선정

- Query 및 URL 정보의 다양화 (Query Count, URL's Collection 등)

- 간편성, 단순성

3.4 CLARA 2의 생성과정

1. 질의 선정

- 검색 개선 있어서 좋은 질의는 무엇일까?
- 질의를 어떻게 다양하게 만들까?

2. 사용자 로그로부터 해당 Query에 해당하는 클릭 정보를 추출

3. Relevance 태그

3.5 Pyclick; 직접 Click Model을 만들어보자

DEVIEW 2021

examples	issue #3	3 days ago
pyclick	issue #3	3 days ago
.DS_Store	issue #3	3 days ago
.gitignore	Initial commit	last month
LICENSE	first commit from https://github.com/ghcho80/PyClick forked from h...	last month
README.md	first commit from https://github.com/ghcho80/PyClick forked from h...	last month
setup.py	first commit from https://github.com/ghcho80/PyClick forked from h...	last month
test_run.sh	first commit from https://github.com/ghcho80/PyClick forked from h...	last month

README.md

PyClick - Click Models for Web Search

PyClick is an open-source Python library of click models for web search. It implements all standard click models and most inference methods described in the following book:

```
Aleksandr Chuklin, Ilya Markov, Maarten de Rijke.
Click Models for Web Search.
Morgan & Claypool Publishers, 2015.
http://clickmodels.weebly.com/the-book.html
```

출처: https://github.com/markovi/PyClick

3.6 Our Contribution on Pyclick

1. Matrix Factorization Click Model (MFCM)

- Position Based Model에 Collaborative Filtering을 적용
- Tail Query에 대한 성능 향상에 좋음

2. Forgetting Rate (Online EM Algorithm)

- 실시간 학습에 효과적

3.6 Our Contribution on Pyclick

Matrix Factorization Click Model (MFCM)

Collaborative Filtering의
원리를 PBM에 적용

영화 추천

	Movie1	Movie2	Movie3	Movie4
User1	?	5	2	?
User2	1	3	?	2

Query-URL간의 PBM Attractiveness

	URL1	URL2	URL3	URL4
Query 1	?	1	0.4	?
Query 2	0.2	0.6	?	0.4

출처: Personalized click model through collaborative filtering (Shen et al., 201

3.6 Our Contribution on Pyclick

Forgetting Rate for Online EM

Online EM

θ: Parameter

$X \sim Bernoulli(\theta)$

$$\theta^{(new)} = \frac{P_X + P(X^{(s)} = 1 \mid \mathbf{C}^{(s)}, \Psi^{(k)})}{|\mathcal{S}_X| + 1}$$

과거 Session에 대한 x의 확률

현재 Session에 대한 x의 확률

Online EM with Forgetting Rate

$$\theta^{(new)} = \frac{P_X \cdot (1-\eta) + P(X^{(s)} = 1 \mid \mathbf{C}^{(s)}, \Psi^{(k)})}{|\mathcal{S}_X| \cdot (1-\eta) + 1}$$

Forggeting Rate 의적용

출처: Online Expectation-Maximization for Click Models (Markov et al., 2016)

3.7 PyClick Tutorial

3.7.1 Pyclick을 통한 모델의 학습

3.7.2 학습데이터 Search Session

3.7.3 Parser의 역할

3.7.3. PBM모델 예시를 통한 Click Model의 Class 구조

3.7 PyClick Tutorial

DEVIEW 2021

Pyclick을 통한 모델의 학습

모델 및 데이터셋 불러오기

Train, Test로 데이터 나누기

클릭모델의 학습 및 평가

```
42  click_model = globals()[sys.argv[1]]()
43
44  search_sessions_path = sys.argv[2]
45  search_sessions_num = int(sys.argv[3])
46  search_sessions = YandexRelPredChallengeParser().parse(search_sessions_path, search_sessions_num)
47
48  train_test_split = int(len(search_sessions) * 0.75)
49  train_sessions = search_sessions[:train_test_split]
50  train_queries = Utils.get_unique_queries(train_sessions)
51  test_sessions = Utils.filter_sessions(search_sessions[train_test_split:], train_queries)
52  test_queries = Utils.get_unique_queries(test_sessions)
53
54  click_model.train(train_sessions)
55
56  loglikelihood = LogLikelihood()
57  perplexity=Perplexity()
58  ll_value = loglikelihood.evaluate(click_model, test_sessions)
59  perp_value = perplexity.evaluate(click_model, test_sessions)[0]
```

parser

List of search_session

3.7 PyClick Tutorial

DEVIEW 2021

학습데이터 SearchSession

(TaskCentric) SearchSession	
field	**query** (type: str)
	task (type: str)
	web_results(type: list of **SearchResult**)

SearchResult	
field	click (type: int in {0,1})
	id (type: str)

3.7 PyClick Tutorial

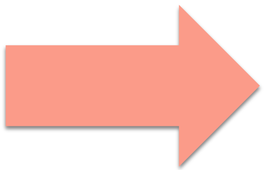
DEVIEW 2021

Parser의 기능

Raw Data

```
0 Q "NAVER"  2 3 4
0 C 3
0 C 4
```

Parser



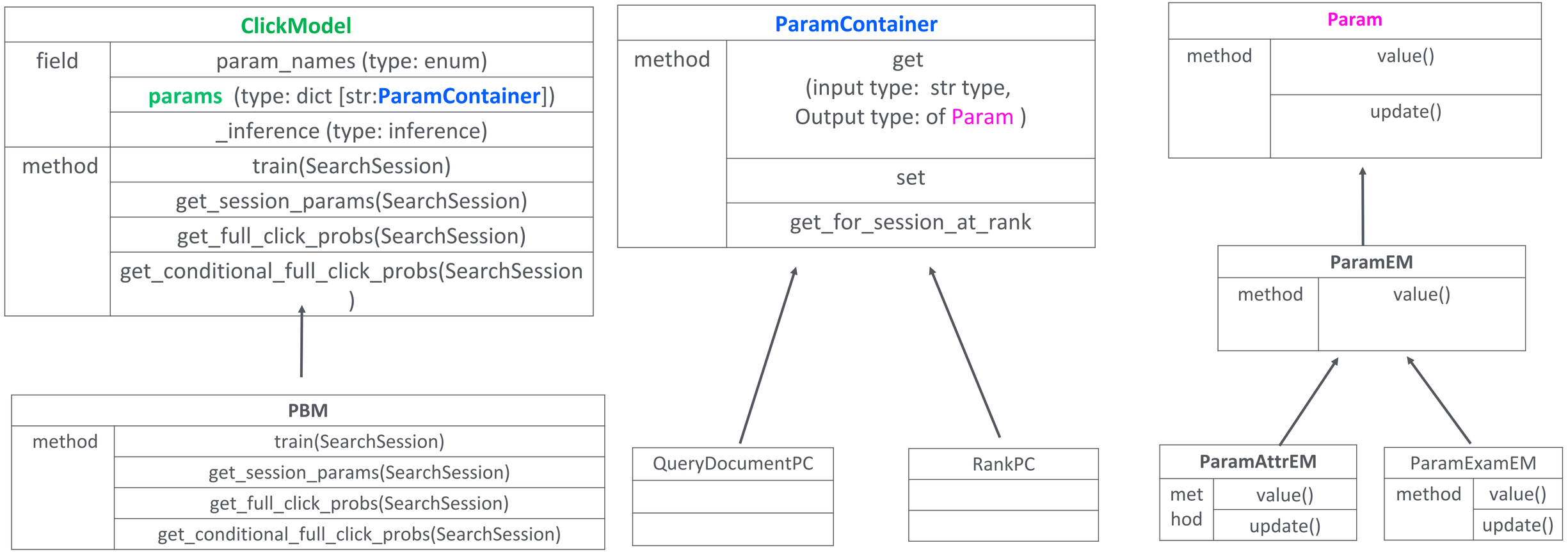
Search Session

```
{"query": "NAVER",
 "task": "0",
"web_results":[{"click": 0, "id": "2"},
               {"click": 1, "id": "3"},
               {"click": 1, "id": "4"}, ...]
```

3.7 PyClick Tutorial

DEVIEW 2021

PBM 모델을 통한 Click Model의 Class 구조의 이해



3.8 Pyclick을 활용한 MFCM의 구현

STEP 1. PBM 모델로 학습하기

STEP 2. PBM 모델에서 Attractiveness 파라미터를 추출하기

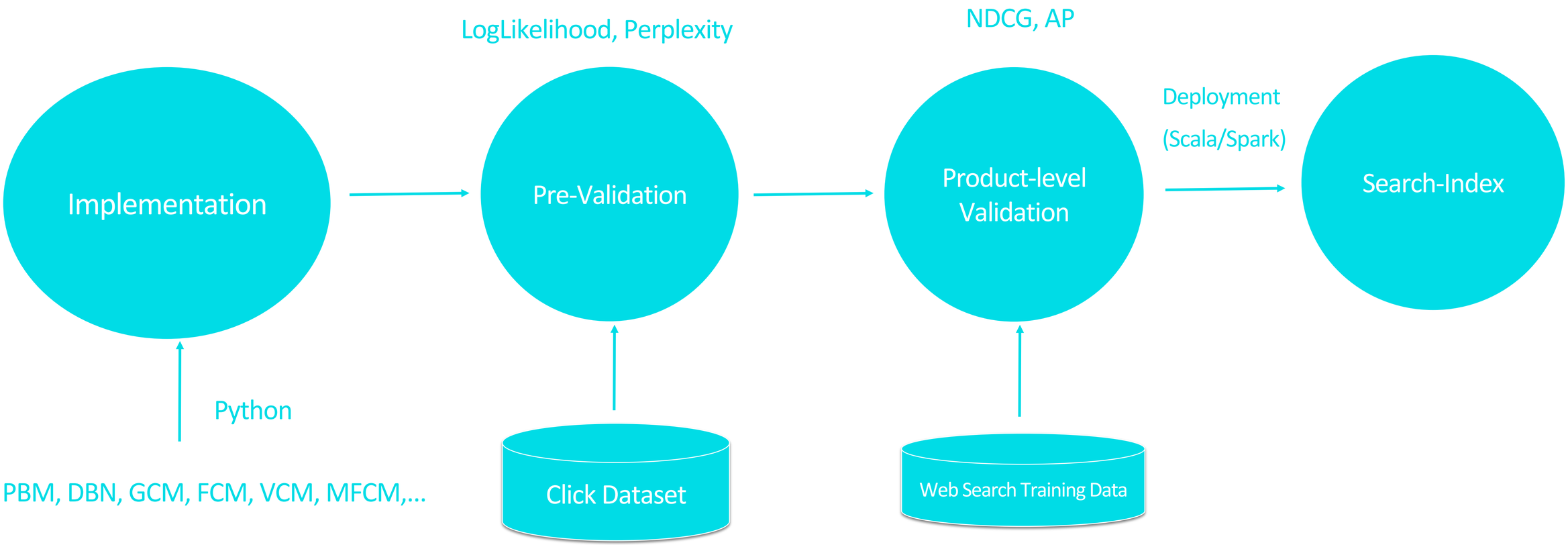
STEP 3. Attractiveness에 MFCM을 적용시켜 Query, URL에 대한 벡터를 얻어내기

STEP 4. STEP 3에서 얻어낸 벡터를 통해서, Attractiveness 파라미터를 재구성하기

STEP 5. 재구성된 Attractiveness 파라미터를 PBM 모델에 다시 재입력하기

4.검색에 활용하기

4.1 User Modeling Process



4.2 CLARA2를 통한 검색 개선에 참여하세요.

DEVIEW 2021

Naver github repository

- 네이버 깃허브 : https://github.com/orgs/naver
- 네이버 깃허브에서 **'Pyclick'** Repository 를 검색

데이터셋 경로

- example/data/clara2/*

Contributors

DEVIEW 2021

Members






QA/We are hiring!

- joonsun.baek@navercorp.com
- suhyeon.baek@navercorp.com
- geunhee.cho@navercorp.com

Thank You