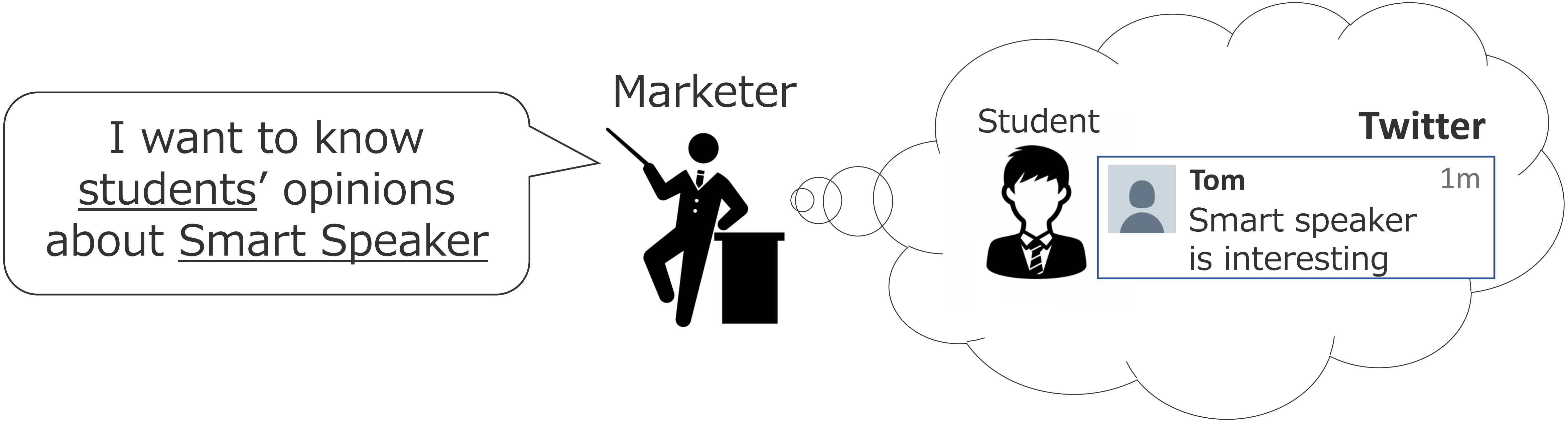
SNS Retrieval Based On User Profile Estimation Using Transfer Learning From Web Search

Daisuke Kataoka and Keishi Tajima

Kyoto University, Japan

Background

- Twitter is useful for opinion mining
  Ex) Target marketing of new products

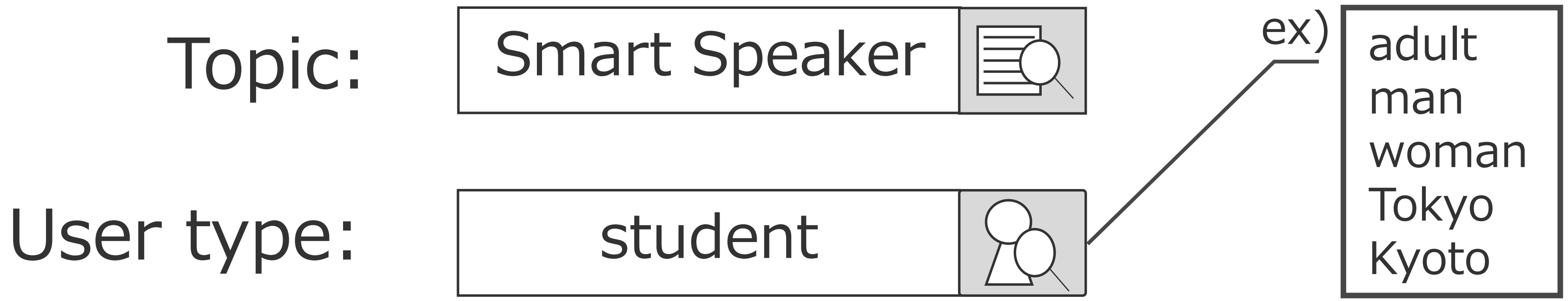


- To know the opinions from students,
  the marketer search Twitter with [Smart Speaker student]

Purpose

Tweet Retrieval specifying **topics** and **user types**

Ex) posts about Smart Speaker posted by students

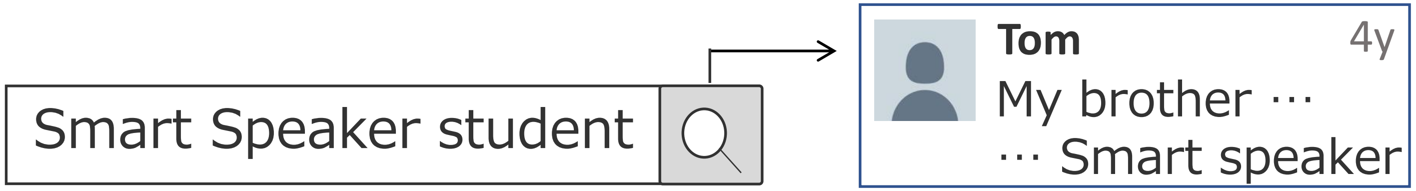


- In this way, search by these two conditions

Problems on Twitter search

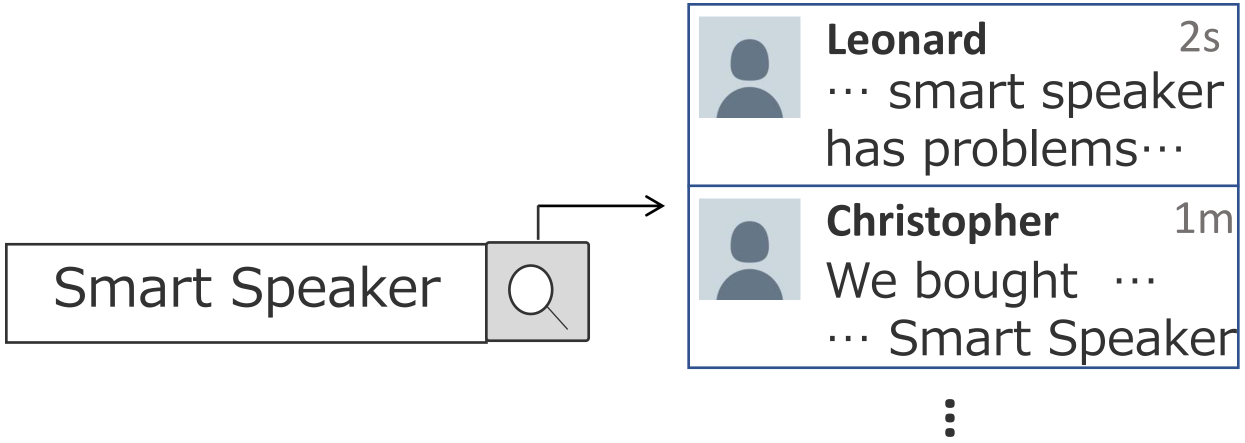
➢ However, a user type do not appear in tweet itself

① search with topic and user types → low recall



- Users do not write their affiliation on each tweet

② only topic keywords → low precision

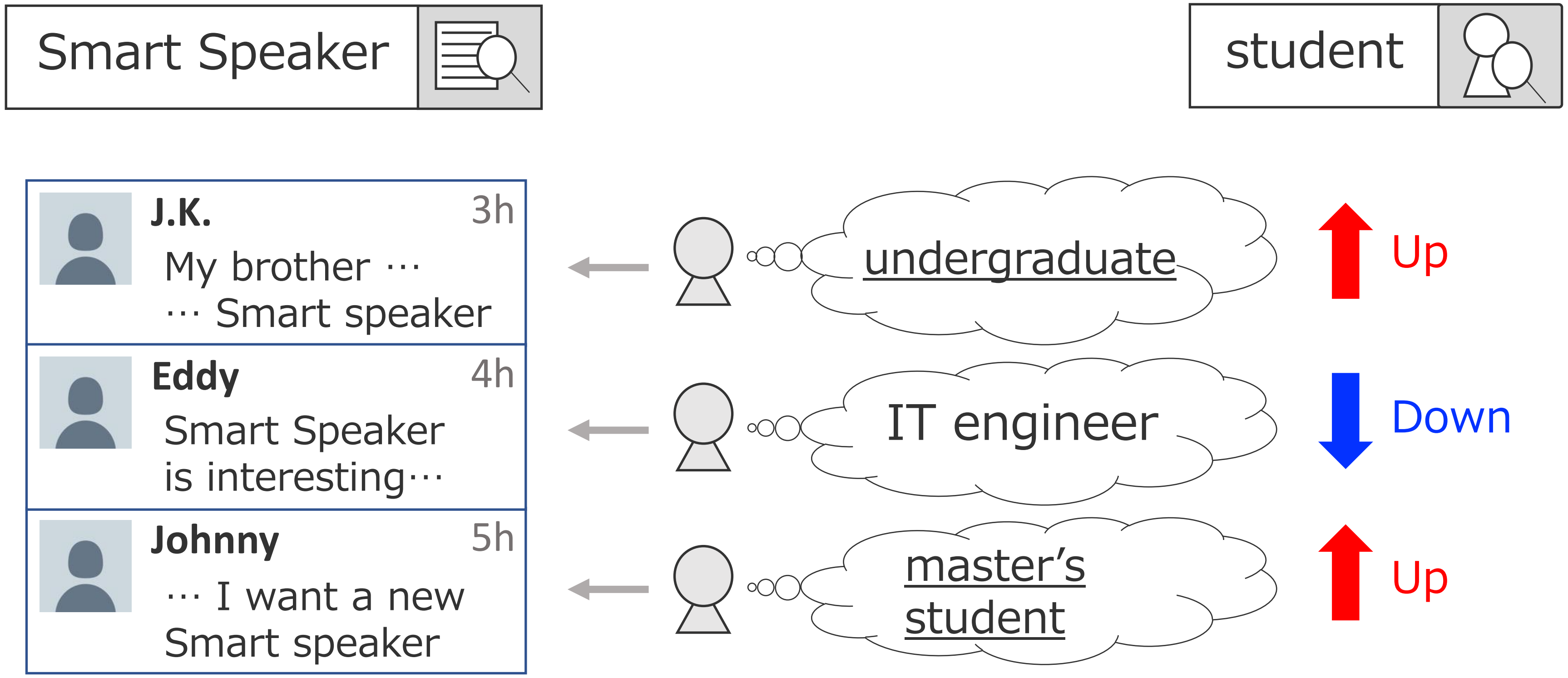


- including posts by non-student users

Proposal

- We propose Re-Ranking method integrated with **user profile estimation**

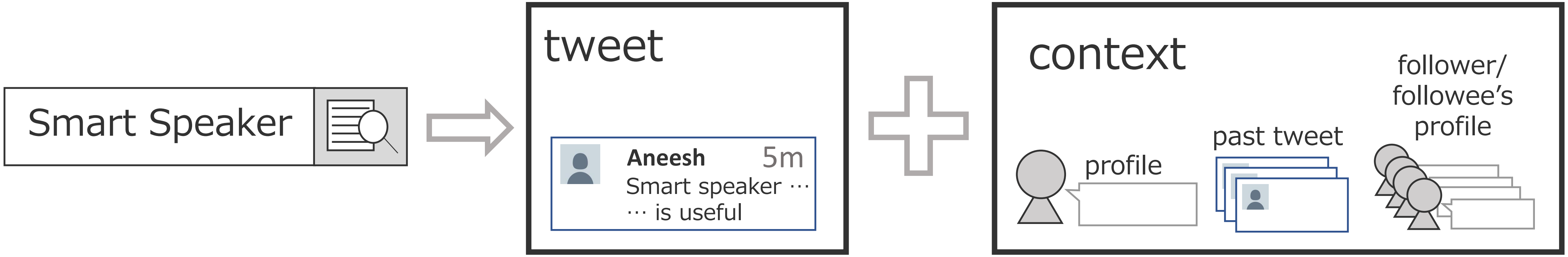
- Search by **Topic query**
- Re-rank based on **Profile query**



Approach

- Regard user profile estimation as a **classification** problem
  - Estimate a likelihood of having the target user profile
  - Give higher ranks to the tweets with higher likelihood

- Judge from the tweets and their context †



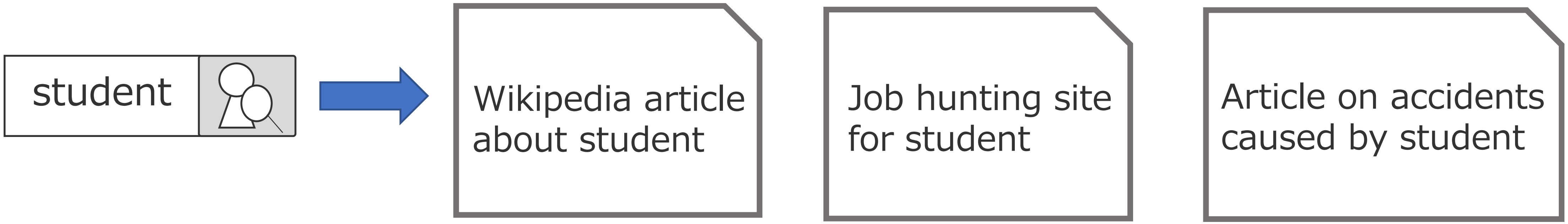
† D. Kataoka et al., Context-aware Relevance Feedback over SNS Graph Data, WI 2017

Hypothesis 1 on user profile estimation

Search Engine is useful for
obtaining vocabulary about user profiles

- Search engine returns documents relevant to query

→ By retrieving by a keyword representing the type of users, documents describing the target user type are obtained



Hypothesis 2 on user profile estimation

Web search includes more vocabulary
for identifying the target user types
than tweets retrieved through **Twitter search**

Twitter student

- Relevant but less vocabulary

A
Students are jealous

B
Adults are busier
than students

C
I'd like to go abroad
while I'm a student

student **Web**

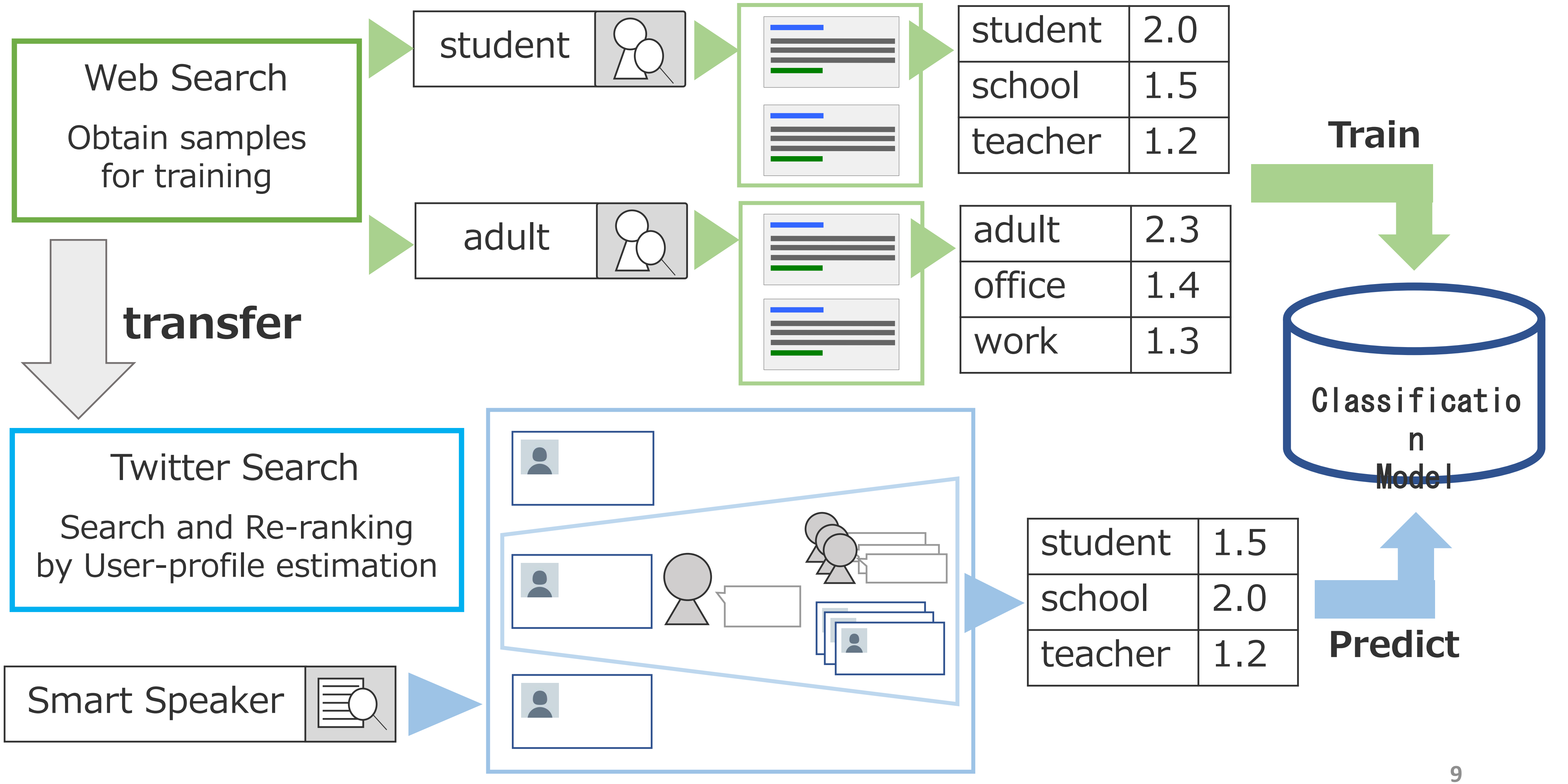
Wikipedia:STUDENT - Wikipedia
A student is primarily a person
enrolled in school l ...

Student Training Modules
Student Training Modules.
This set of modules helps ...

NI LabVIEW Student Edition
The LabVIEW Student Edition
contains the following soft ...

- much more vocabulary
- The more documents describing the user types

Transfer Learning

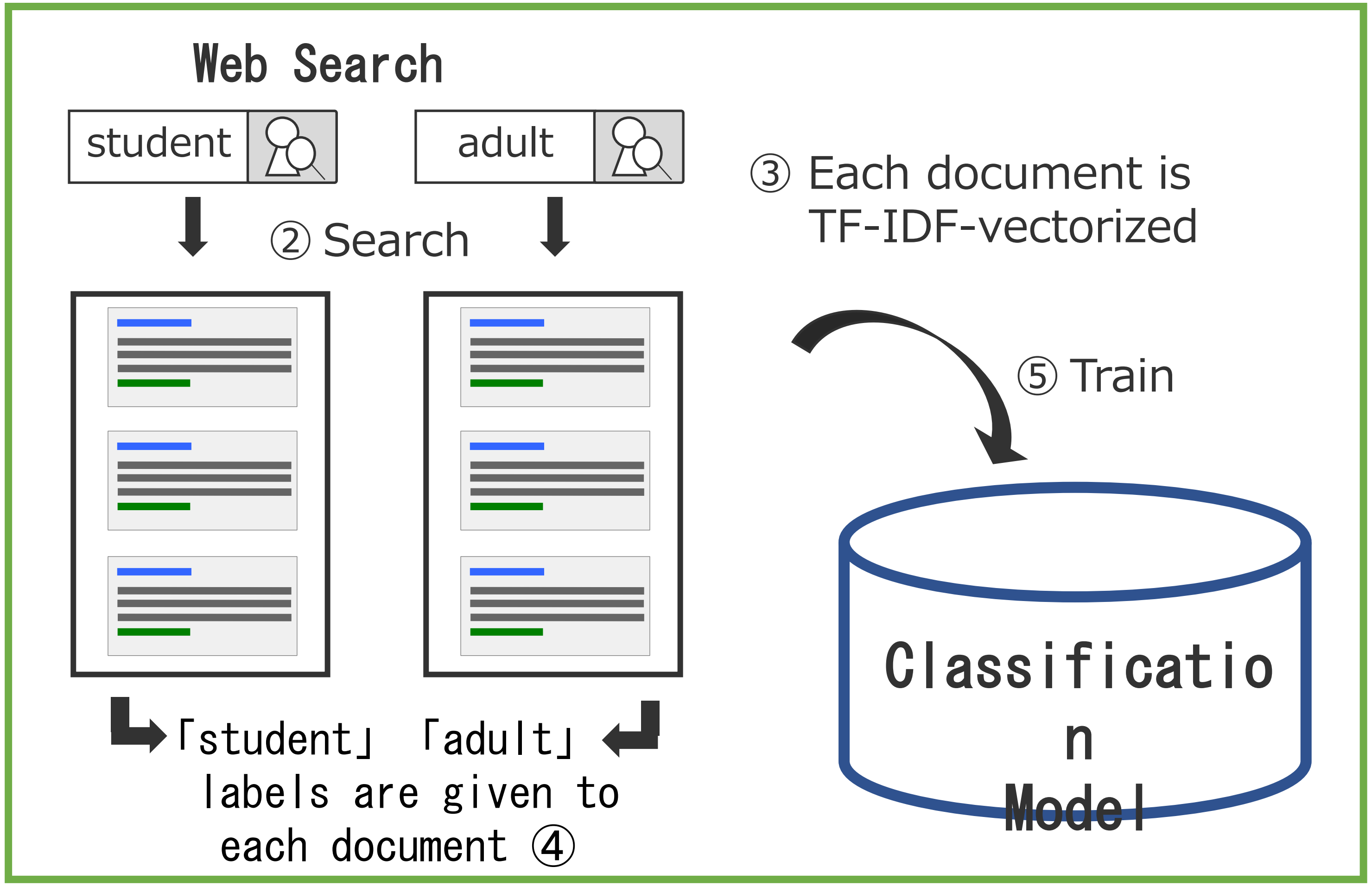


Training phase

➢ Collect positive and negative samples as datasets and train a classification model

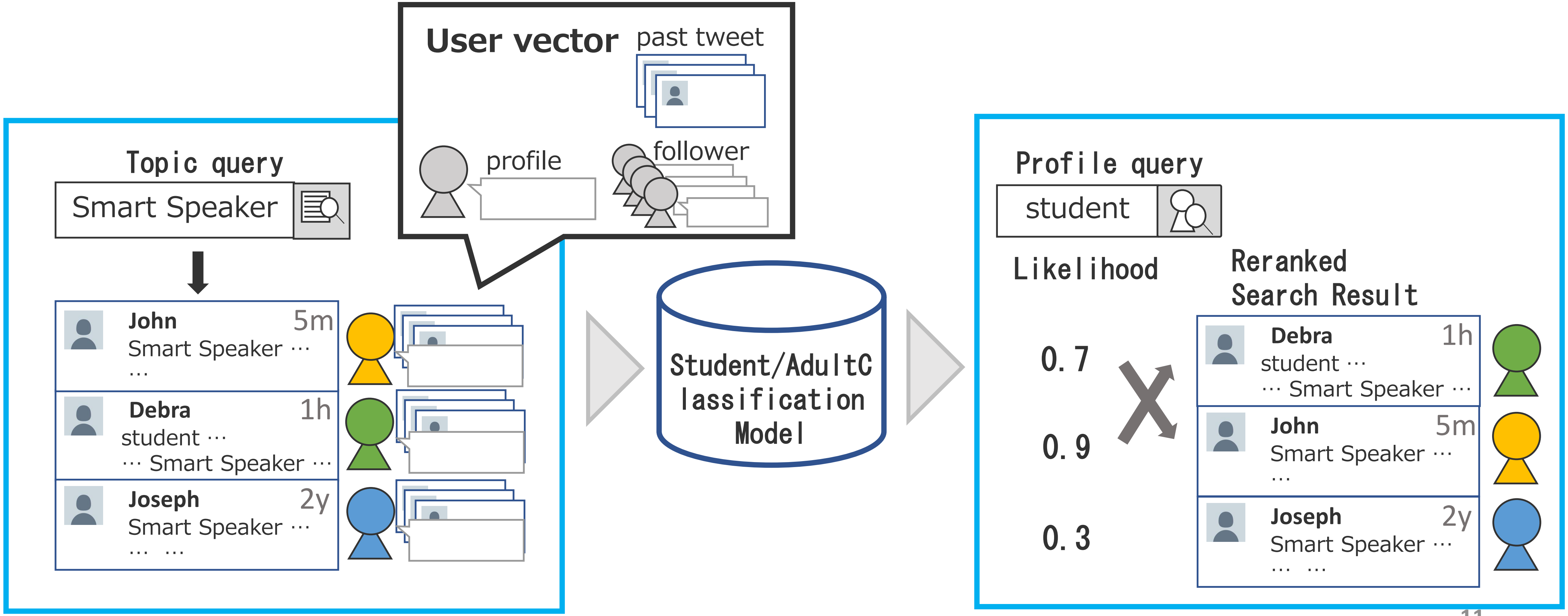
① **Choose words describing the target user type and its opposite**

(student, adult)
(man, woman)
(Kyoto, Tokyo, …)
(soccer, baseball, …)
⋮



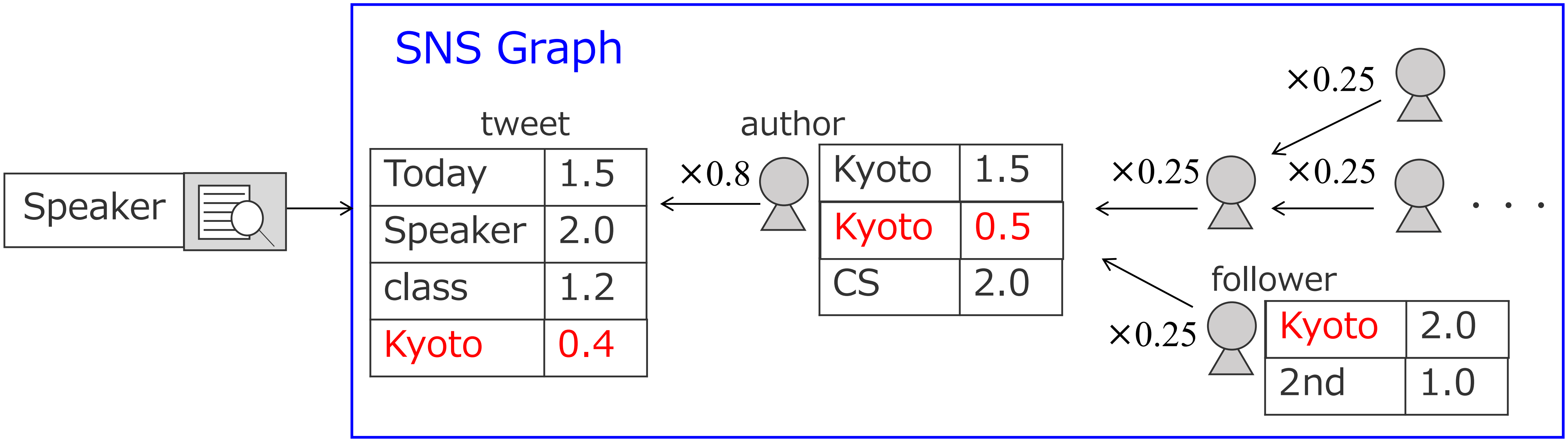
Applying phase

- Apply the trained model on Tweet Retrieval
  - Input each user vector and obtain prediction score
  - Rank tweets by the likelihood of the target user type



How to make a user vector on Twitter †

- Represent Twitter data as a graph
- Propagate the word weight
- Weights for user vector decay through the graph
  - Low weight is assigned to the context far from a tweet



■ The weight of「Kyoto」is computed as $2.0 \times 0.25 \times 0.8 = 0.4$

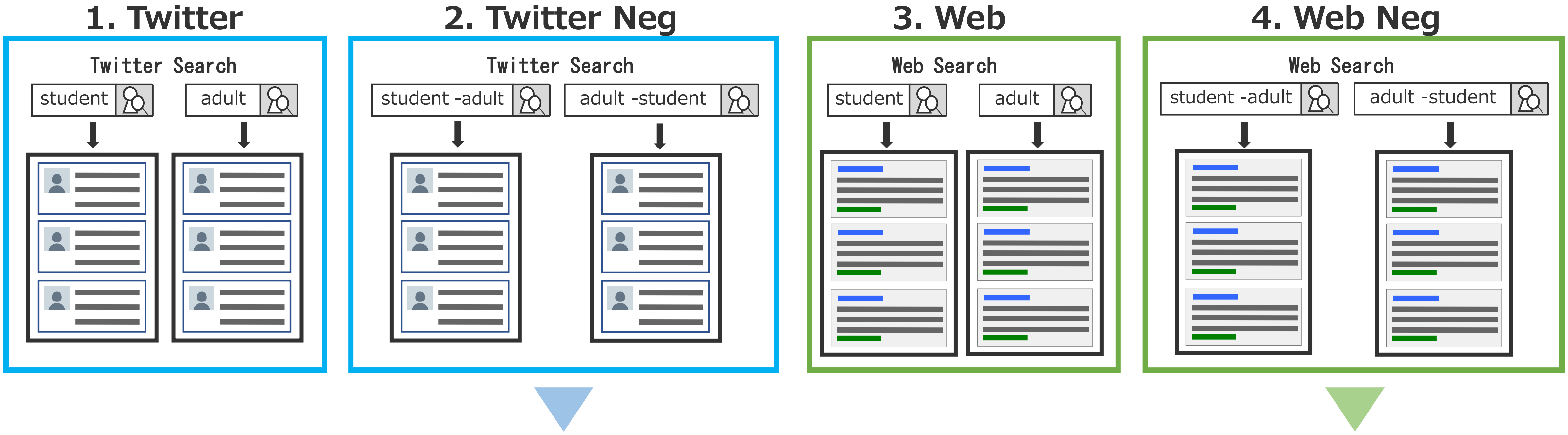
† D. Kataoka et al., Context-aware Relevance Feedback over SNS Graph Data, WI 2017

Experiments

➢ Comparison with four methods

1. Twitter (Baseline 1)
2. Twitter Neg (Baseline 2)
3. Web (Proposed method)
4. Web Neg (Proposed method)

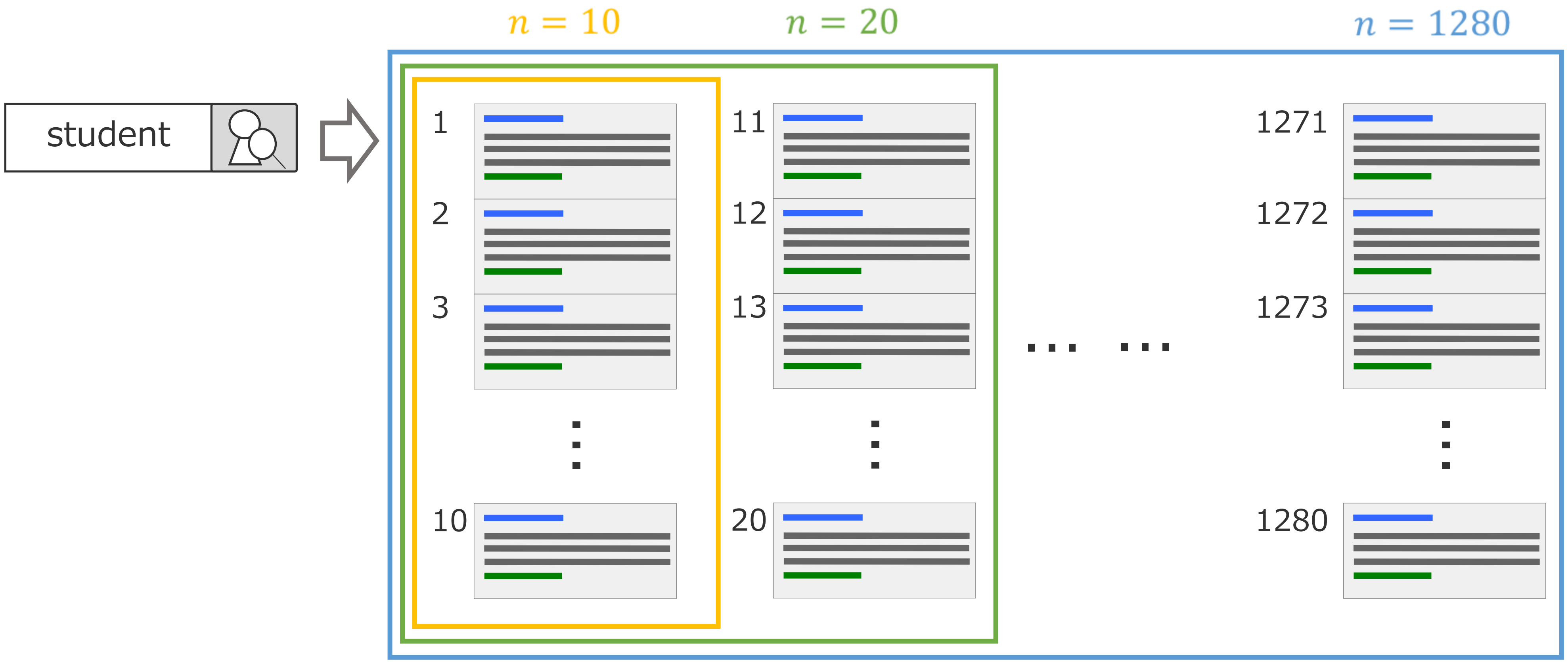
* In Neg method, negation operator is used on Web search Ex) student -adult



➢ The search results would **not** contain the documents about the opposite user type

Experiments

- Random Forest classification model is used for training
- **Top n** Web/Twitter **search results** are used for training models
  - examine the difference when changing the size of datasets
  - $n = 10, 20, 40, 80, 160, 320, 640, 1280$



Experiments

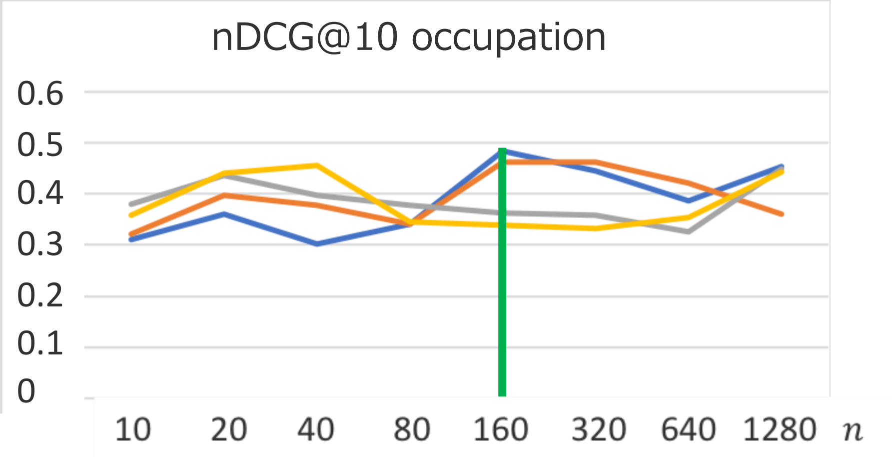
- Queries: 5 (topic query) * 6 (profile query) = 30 queries

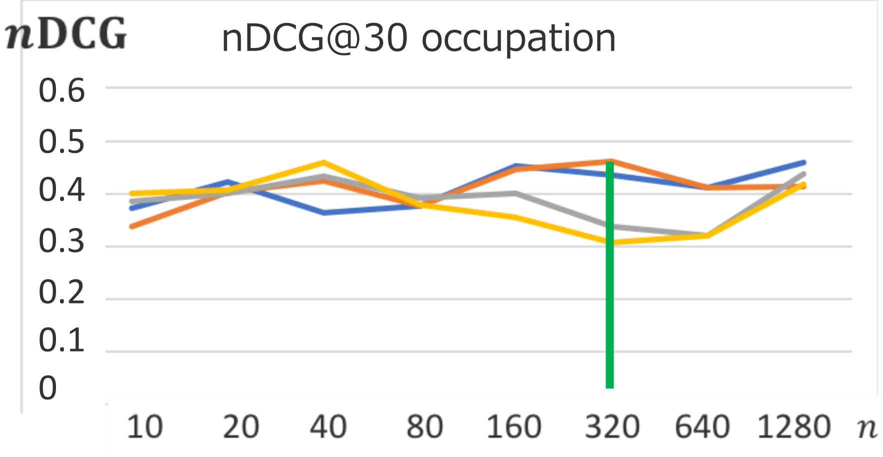
Occupation	Region	Gender
student/adult	Osaka/Tokyo	man/woman

	Topic Query	Profile Query	Category	Intention
1	Smart Speaker	student	Occupation	posts about Smart Speaker by students
2	Osaka Castle	Osaka	Region	posts about Osaka Castle by the persons from Osaka
3	Star Wars	man	Gender	posts about Star Wars by men

Experimental Results

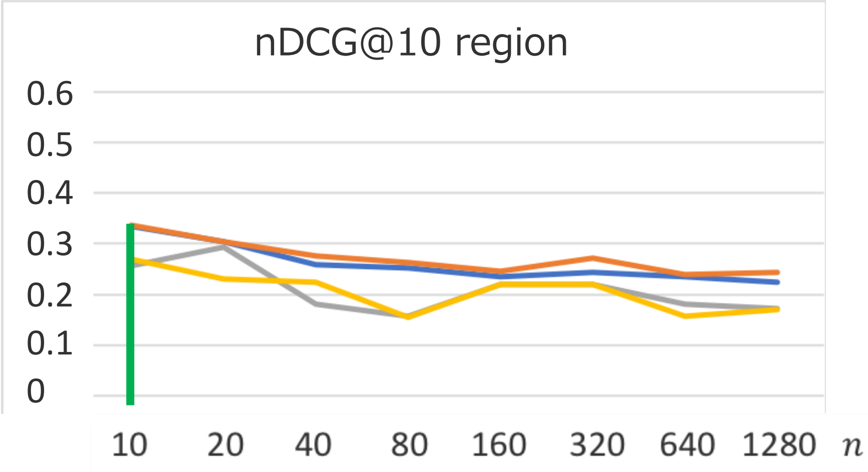
- In the **occupation**, **Web Neg** method showed highest nDCG@10,30

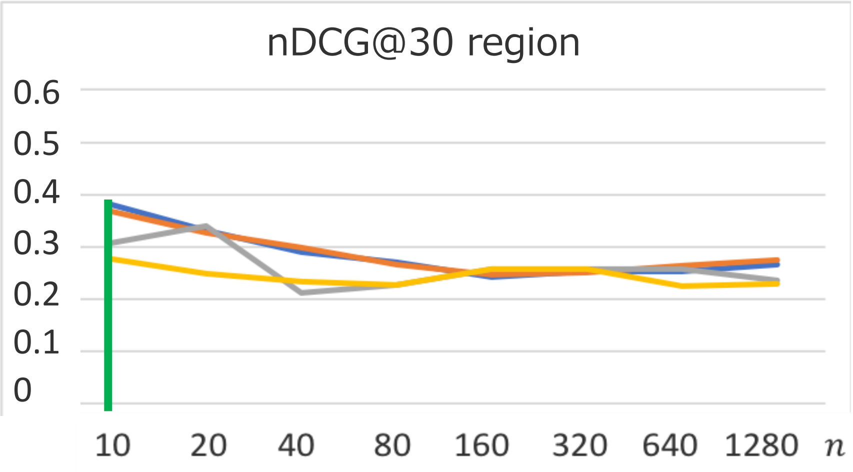




< **Top-n results**

- In the **region**, **Web Neg** method showed highest nDCG@10,30

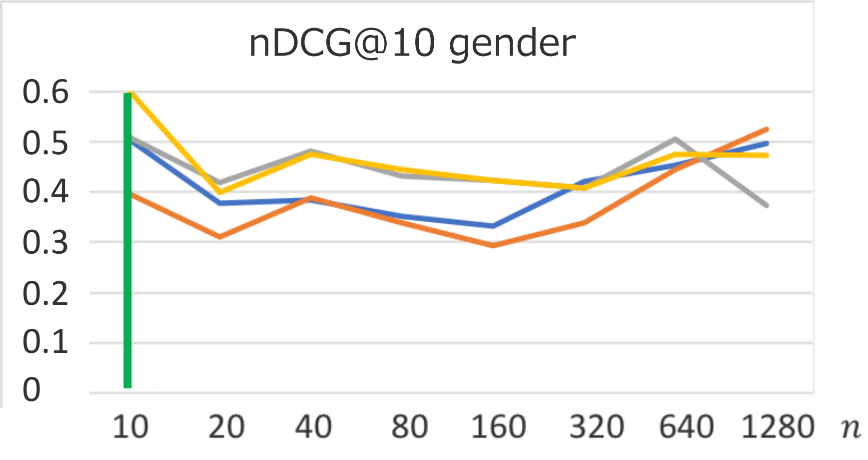


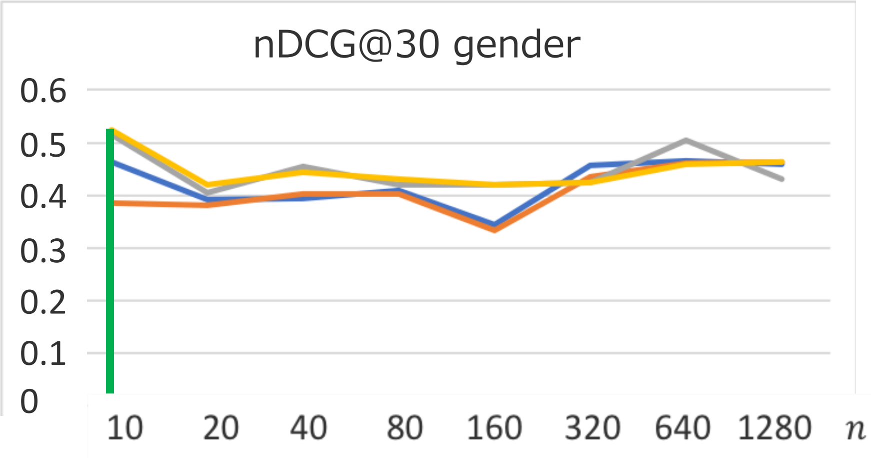


Experimental Results

— : Twitter — : Twitter Neg — : Web — : Web Neg

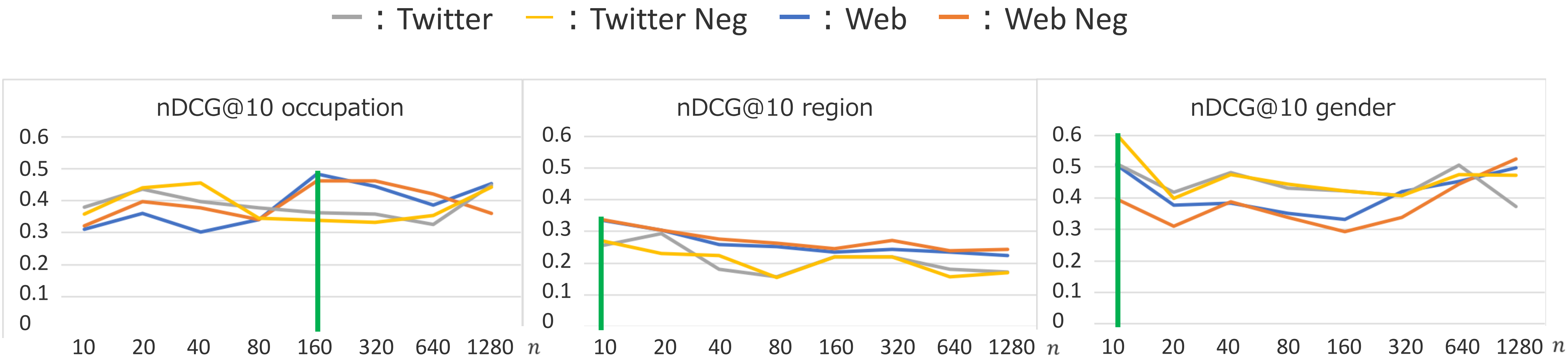
- In the **gender**, **Twitter Neg** method showed highest nDCG@10,30





- In the three categories,
  - the best methods are different
  - the best n of top-n results is different

Discussions about n



- The best n is different

- A bigger n does **not necessarily** produce a higher nDCG

← because search quality becomes lower at low rank

→ Not all documents are useful for training

Discussions about effective features

➢ Examine the effective features of the Random Forest based on the feature importance calculated by "gini importance" †

Occupation Models

Rank	Twitter 320	score	Twitter Neg 320	score	Web 320	score	Web Neg 320	score
1	do	0.952	given	0.418	**society**	0.538	**society**	0.589
2	behind	0.0281	(	0.2196	**student**	0.00918	**student**	0.0557
3	woman	0.0056	like	0.146	reserved	0.0388	©□	0.0539
4	3	0.00429	.	0.0268	©□	0.0287	reserved	0.0412
5	)	0.00244	center	0.023	copylight	0.0273	copylight	0.0327
6	exists	0.00147	3	0.0171	**person**	0.0127	**person**	0.0113
7	**person**	0.00146	/	0.0882	**work**	0.00833	**public**	0.00907
8	.	0.00123	**person**	0.00455	include	0.00739	woman	0.00707
9	become	0.00123	do	0.0045	**age**	0.00614	**drink**	0.00596
10	/	0.000742	Dazai	0.0043	**public**	0.00545	all	0.00502

➢ There are more effective features on **Web** and **Web Neg**

† Breiman, Friedman, "Classification and regression trees", 1984.

Summary & Future works

Summary

- We proposed a new tweet ranking method using transfer learning from Web domain
- Our proposed method achieved higher nDCG in the occupation and region categories, but low in the gender category

Future works

- determines n of the top-n results automatically and dynamically
- adopt domain adaptation approaches