DAY 11
IR-related technologies

Several information processing techniques related to IR will be explained one by one, starting with standard IR. Certain aspects, including purpose, user behaviour and underlying ideas have gradually shifted. So what makes these techniques different, or in some cases, similar?

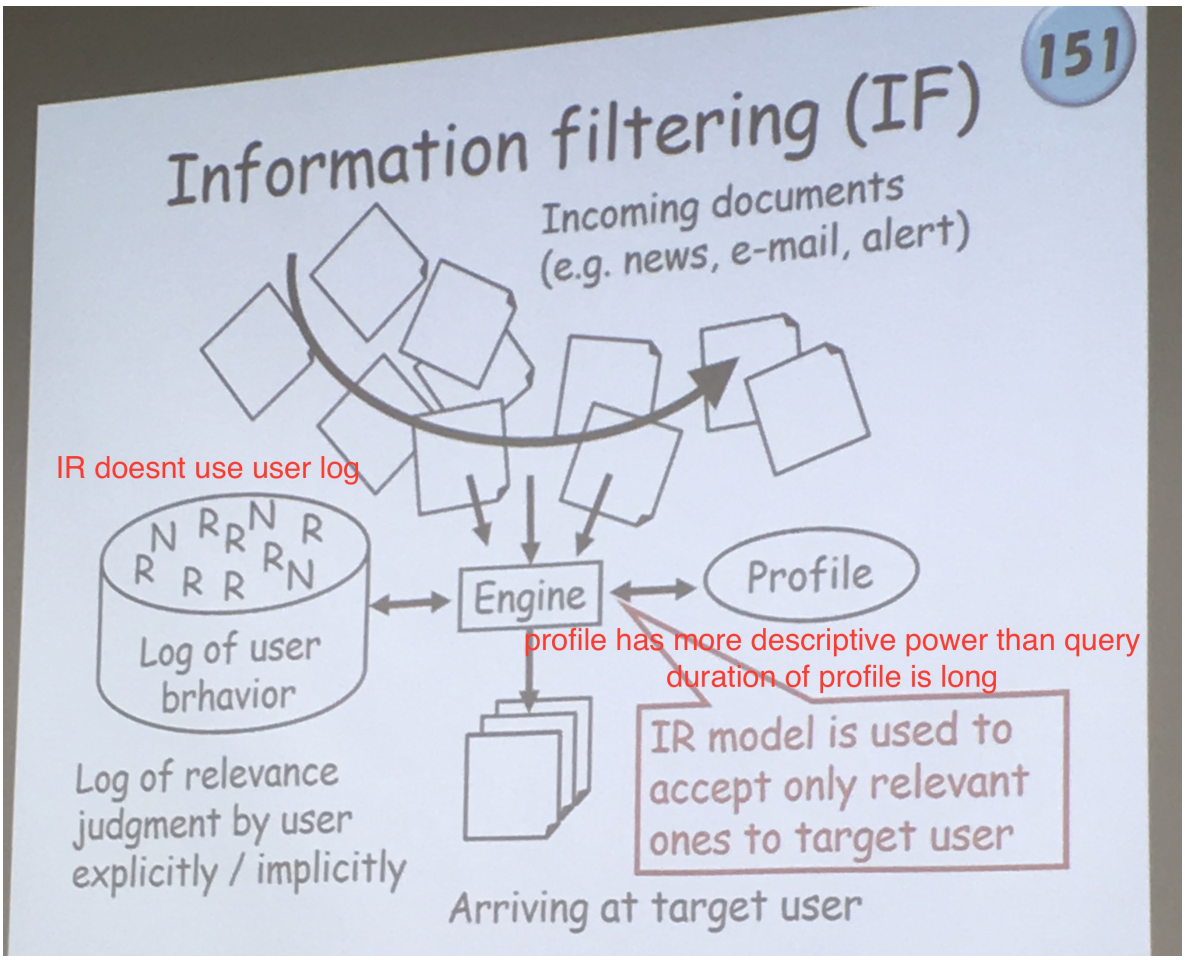
Information Filtering (IF)

So there's a huge flow of incoming documents (i.e. news, e-mails, alerts) and these all go into the engine. This IR model, however, only accepts those relevant to the target user. So the engine filters it, using a log of user behaviour as well and a profile. It uses the profile to check. The log of relevance is judgements done by the user implicitly.

Compare this to Information Retrieval:

They are not the same when it comes to need and document. They can be seen as the other side of the same coin.

So in IR when information need (dis)appears arbitrarily, it's called ad hoc retrieval.
In IF, because the duration of need is long, it makes sense to adapt system to user model.
IR and IF use the same IR model, but differently. Hence the coin.



Comparing IR and IF

- Contrastive in duration of need and document
  - In IR, information need (dis-)appears arbitrarily -> IR is called “ad hoc retrieval”
  - In IF, because duration of need is long, it makes sense to adapt system to user model
- IR and IF use same IR model differently
- Approximately by query and profile, in IR and IF, respectively

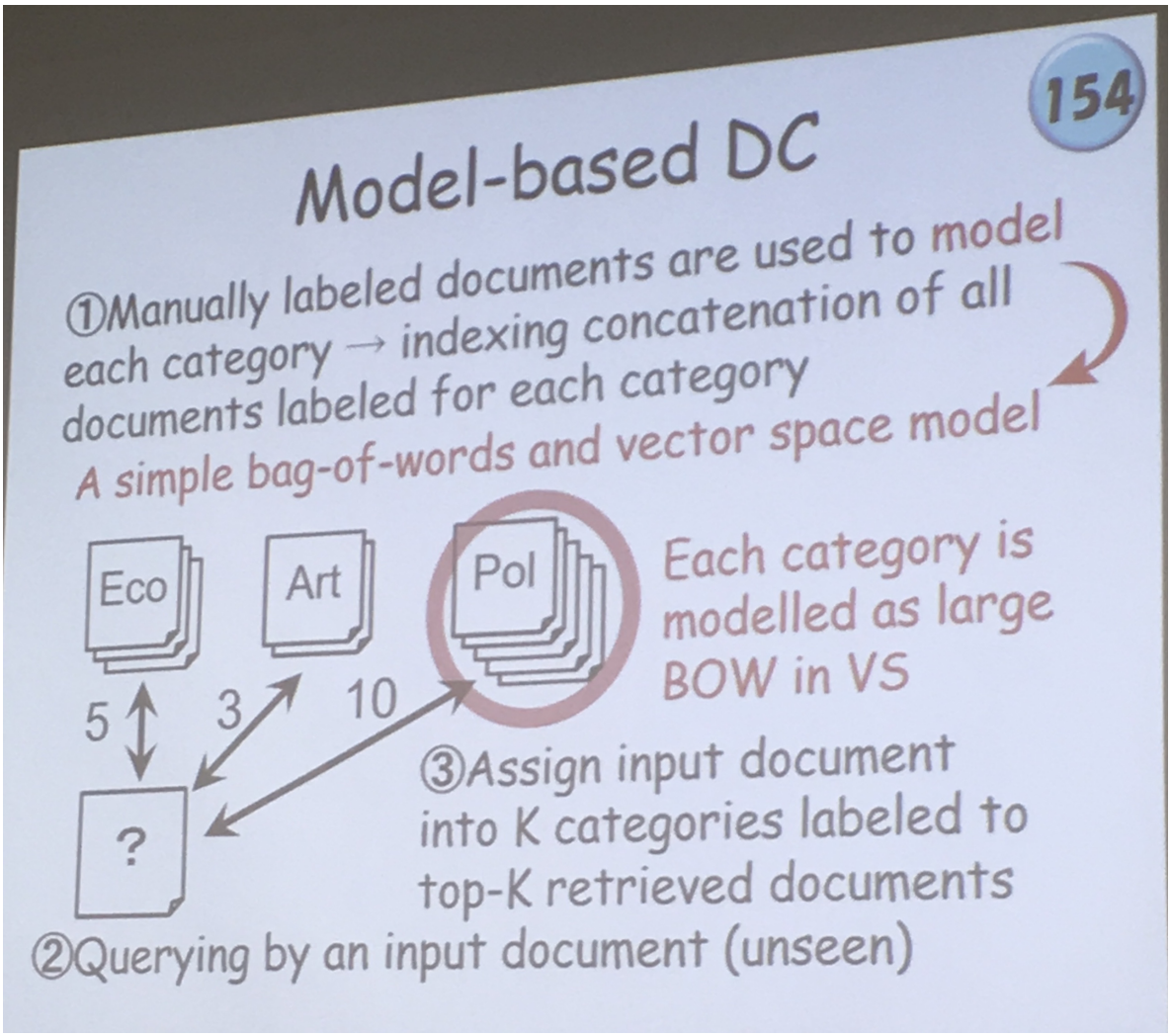
	IR	IF
Need	short	long
Doc	long	Short

Document categorization (DC)

- Categorization
  - Assign a document to predefined categories, such as politics, sports, and science
  - Same document can be categorized into more than one category
  - Eg. Article of olympics -> politics +sports
- Clustering
  - Divide a set of documents into predefined number of subsets
  - Criteria for division is dependent of the set

(DC:Association among documents. IR: association between information and documents)

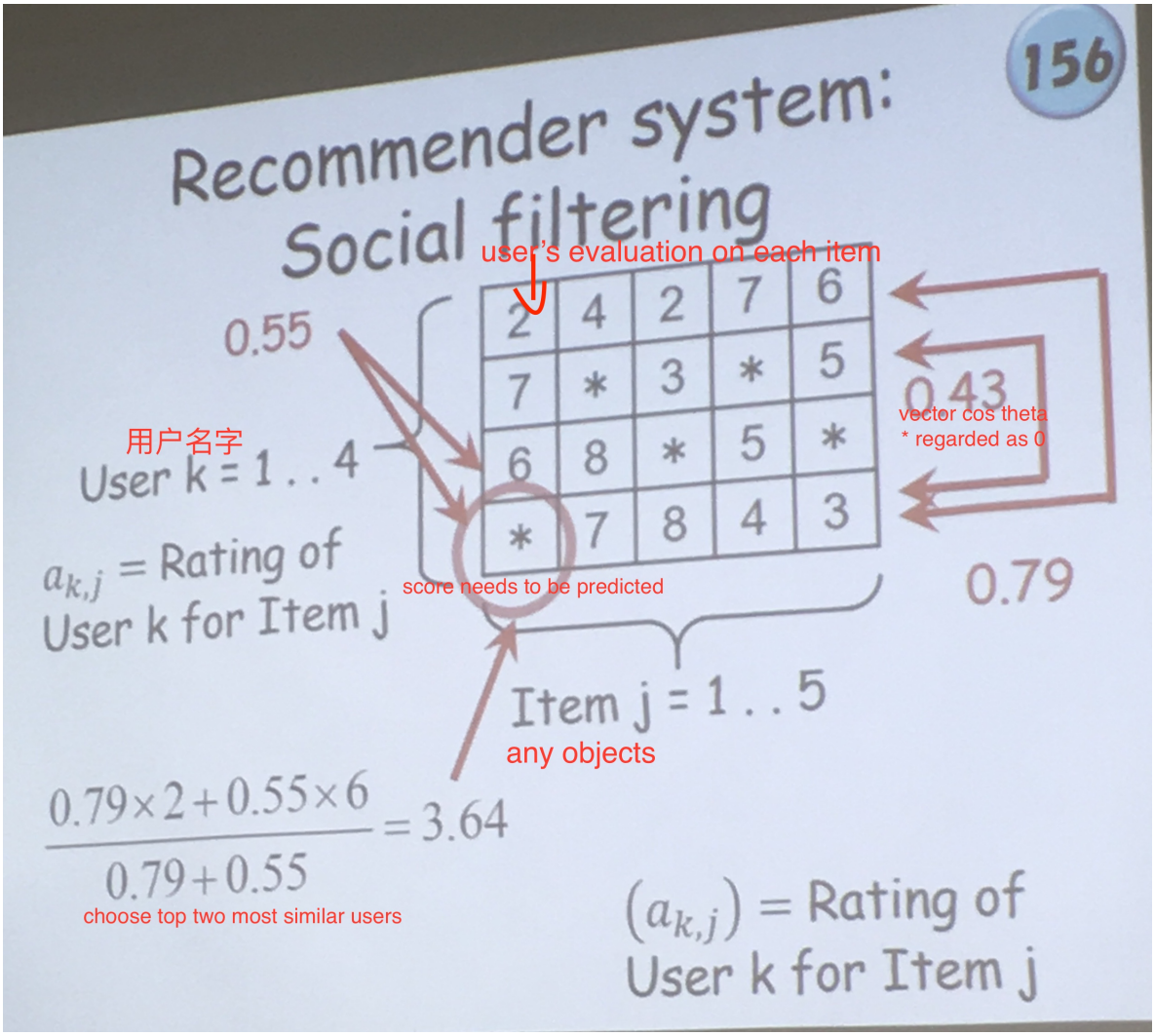
Model-based DC



Example-based DC (memory-based)

- KNN
  1. Indexing all manually labeled documents
  2. Querying by an input document (unseen)
  3. Top-k documents vote for their category -> assign input doc into majority category
- Unlike model-based DC, overgeneralization can be avoided
- K = 5

?—querying—>1. Eco 2. Eco 3. Art 4. Eco 5. Pol


156
Recommender system:
Social filtering
user's evaluation on each item
0.55
0.43
vector cos theta
* regarded as 0
用户名字
User k = 1 . . 4
$a_{k,j}$ = Rating of
User k for Item j
score needs to be predicted
0.79
Item j = 1 . . 5
any objects
$\frac{0.79\times 2+0.55\times 6}{0.79+0.55} = 3.64$
choose top two most similar users
$(a_{k,j})$ = Rating of
User k for Item j