web mining=

↳ web content mining = web page content mining + search result mining

↳ crawler, spider (sürünücü) = program that is used to search and automatically index website context and other information over the internet

↳ hub (merkez) pages = contain links to many other pages

↳ web structure mining

↳ pagerank = importance of a page is calculated based on number of pages which point to it
–weighted backlink–

↳ clever = identify authoritative and hub pages
–important–

↳ hits (hyperlink induces topic search) = based on set of keywords, find set of relevant pages

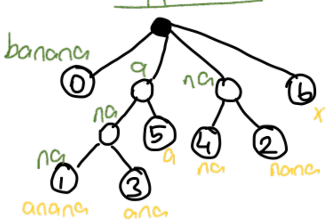
↳ web usage mining = general access pattern tracking + costumized usage tracking

- data structures used

↳ trie = rooted tree, path from root to leaf is pattern

↳ sufix tree = each suffix in the list is compressed and represented by a single node in tree

suffix tree



- episodes = partially ordered set of pages

text mining=

↳ text retrival measures = precision = $\frac{\text{relevant} \wedge \text{retreived}}{\text{retreived}}$ recall = $\frac{\text{relevant} \wedge \text{retreived}}{\text{relevant}}$

need labeled data to compute

- picky algorithm = precision high, recall low
- relaxed algorithm = precision low, recall high

* remove stopwords before mining → to reduce size and improve efficiency

- stemming = techniques used to find the root / stem of the word: gone, goes, going...
- cosine dist between two documents = 1 if a word exist, 0 else (to calculate the similarities between documents)

* if a term occurs frequently in many documents, less important

IDF (inverse document frequency) = $\log(N/N_j)$ N = total number of documents
N_j = number of documents that contain item N_j

Term Importance = TF (term frequency) × IDF

- LSI (latest semontic indexing) = trying to extract hidden semontic structure
  ↳ car and automobile are same cannot detect normally

recommender system=

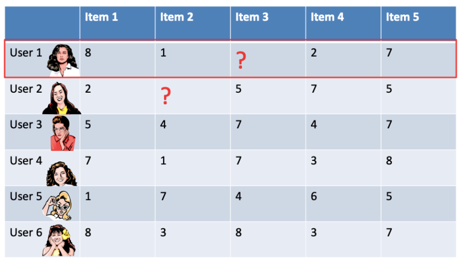
↳ content based recommender system = recommend items similar to those users preffered in the past
→item features are used

↳ collaborative recommender system = uses other users recommendations to recommend
→similar

↳ user-based collaborative filtering = people who agreed in the past are likely to agree again

↳ item-based collaborative filtering = a user is likely to have the same opinion for similar items

- difference from content based → similarity measure, here looking at how other users rated them



	Item 1	Item 2	Item 3	Item 4	Item 5
User 1	8	1	?	2	7
User 2	2	?	5	7	5
User 3	5	4	7	4	7
User 4	7	1	7	3	8
User 5	1	7	4	6	5
User 6	8	3	8	3	7

user based ⇒ first calculate the similarity between user 1 and rest.

↳ ex: user 1 and user 2 → (|8-2| + |2-7| + |7-5|) / 3 → distance

↳ use weighted sum $\frac{1}{dist}$ (item3 in user 2) to recommend, may consider only the K-nearest neighbor

item based ⇒ calculate how item 3 and item 4 are similar → (|7-5| + |4-7| + |3-7| .... / n

↳ recommend according to weighted sum and K-nn

↳ hybrid recommender system =

↳ weighted = several weighted recom. techniques

↳ switching = depening on the current situation

↳ mixed = recommendations from different recommenders are presented simultaneously

↳ cascade = $\boxed{\text{rec 1}} \longrightarrow \boxed{\text{rec2}} \longrightarrow$ recommend

↳ model based collaborative filtering