course:

Searching the Web (NDBI038)

Searching the Web and Multimedia Databases (BI-VWM)

lecture 7:

Personalized search and social context

prof. RNDr. Tomáš Skopal, Ph.D.
Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague
Department of Software Engineering, Faculty of Information Technology, Czech Technical University in Prague

Today's lecture outline

- recommender systems
  - collaborative filtering
  - content recommendation
  - knowledge-based recommendation
  - social recommender systems
- social information retrieval
  - social network = graph
  - social pagerank

Web 2.0

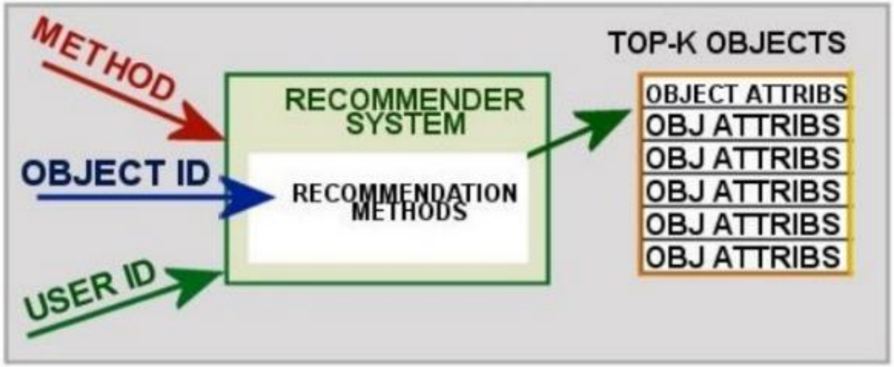
- basic web infrastructure is application-agnostic
- personalization in (web) applications
  - blogs, forums, e-commerce, social media, hosting services, e-banking, e-gov
  - cloud services (office, collaboration, video, etc)
- people part of web creation
  - user-generated content
  - user profiles
  - online communities

Social media

- social media
  - allow people to share content and interact
  - social networks (explicit or inferred)
- information overload
  - blogs, youtube, news, forums, wiki, social networks
- interaction overload
  - friends, followers, comments, sharing, „likers"
- content-based search not sufficient

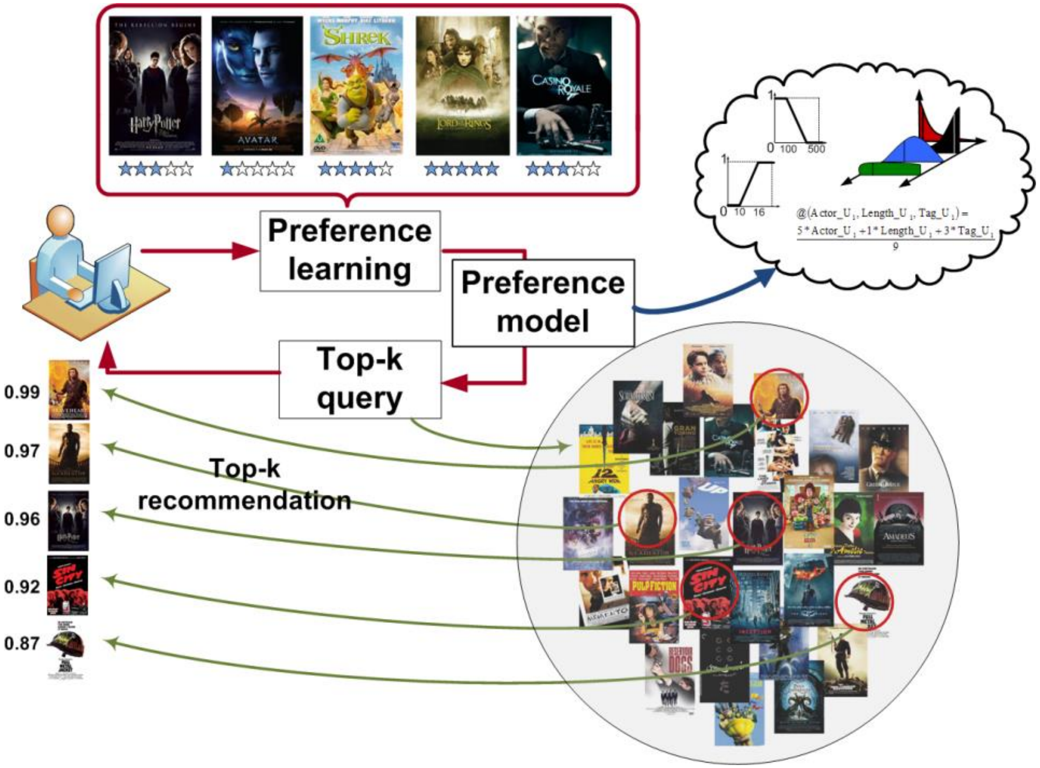
Recommender systems

- subsystems within applications registering users (and items)
- input
  - user + history
  - actual context
    - web page
    - locality
    - time, season
    - actual item displayed
  - item/user attributes (profiles)
- output
  - top-k list of items
  - item ratings
  - presentation and recommendation of top-ranked items



Recommender systems

1. collect user feedback
2. learn preferences
3. recommend top-k items



Moment of recommendation
- front page, sidebar
- follow-up
- on demand

Recommender systems

- user feedback
  - explicit (e.g., ratings) 
    - like annotation
      - \+ precisely stated preference
      - \- not guaranteed, not complete, subjective, inconsistent, etc.,
    - problem in e-commerce – need to buy/try item before
  - implicit (tracking user behavior)
    - noisy, hard to interpret
    - problem – what to track?
      - time on page/object, number of visits, display zoom, mouse clicks/hovers, etc.
    - for free and objective (user cannot fake like by the explicit)

Recommender approaches

- **collaborative filtering** (CF)
  - recommendation based on inter-user similarity
- **content-based recommendation**
  - based on matching features of items to user ratings
- **knowledge-based recommendation**
  - based on matching items to user needs
- **demographic recommendation**
  - based on demographic classes (age, gender, ...)
- hybrid
  - combinations of the above

Recommender approaches

Technique	Model	Source	Method
Collaborative filtering	user-item matrix	rating users x items	identify similar users, extrapolate from their ratings
Content-based	item features	rating users x items	generate a classifier based on user ratings
Knowledge-based	item features	user needs	infer a match between items and user’s needs
Demographic	demographic attributes	demographic attributes	identify similar users, extrapolate from their ratings

Collaborative filtering

- implements recommendation using advice from trusted people (friends, colleagues, etc.)
  - similarities of all users to the active user are computed
  - a subset of most similar users is selected
  - prediction of active user's rating of specific items is computed based on users' rankings
  - recommendation of items with maximum prediction

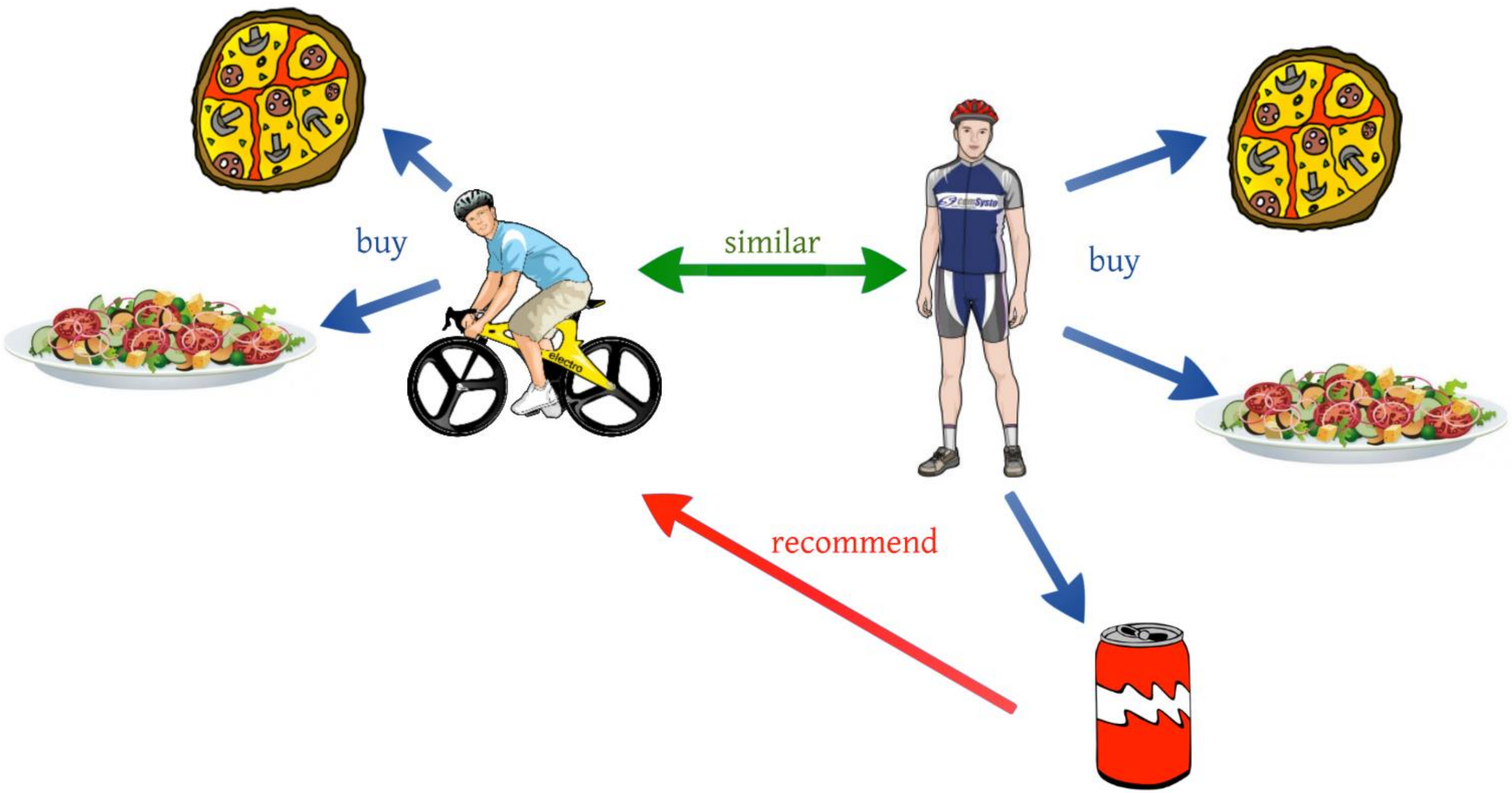
Collaborative filtering

user	Inglourious basterds	Hunger games	Alien
Peter	like	like	dislike
Dana	n/a	dislike	like
Michael	like	like	dislike
Bob	like	like	n/a

Shall we recommend Alien to Bob?

Bob taste is similar to both Peter and Michael, so do **not recommend** Alien to Bob.

Collaborative filtering



Collaborative filtering

- user-based nearest neighbor CF algorithm
  - the mean vote for user i: $$\overline{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}$$

    $v_{i,j}$ = vote (ranking) of user i on item j
    I_i = items on which user i has voted (ranked)
  - predicted vote (rank) of active user a:

$$p_{a,j} = \overline{v}_a + \underbrace{\kappa}_{\text{normalizer}} \sum_{i=1}^{n} \underbrace{w(a,i)}_{\text{weights of } n \text{ similar users}} (v_{i,j} - \overline{v}_i)$$

Collaborative filtering

- cosine similarity

$$\text{simil}(x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{||\vec{x}|| \times ||\vec{y}||} = \frac{\sum\limits_{i\in I_{xy}} r_{x,i} r_{y,i}}{\sqrt{\sum\limits_{i\in I_x} r_{x,i}^2}\sqrt{\sum\limits_{i\in I_y} r_{y,i}^2}}$$

- Pearson similarity

$$\text{simil}(x,y) = \frac{\sum\limits_{i\in I_{xy}} (r_{x,i} - \bar{r_x})(r_{y,i} - \bar{r_y})}{\sqrt{\sum\limits_{i\in I_{xy}} (r_{x,i} - \bar{r_x})^2 \sum\limits_{i\in I_{xy}} (r_{y,i} - \bar{r_y})^2}}$$

Collaborative filtering

- item-based nearest neighbor CF algorithm
  - similarity of items
  - prediction for an item based on the user's ratings for similar items

user	Inglourious basterds	Hunger games	Alien
Peter	like	like	dislike
Dana	n/a	dislike	like
Michael	like	like	dislike
Bob	like	like	n/a

Dana dislikes Hunger games (similar to Ingl.bast.) $\rightarrow$ do not recommend Hunger games to Dana

Collaborative filtering

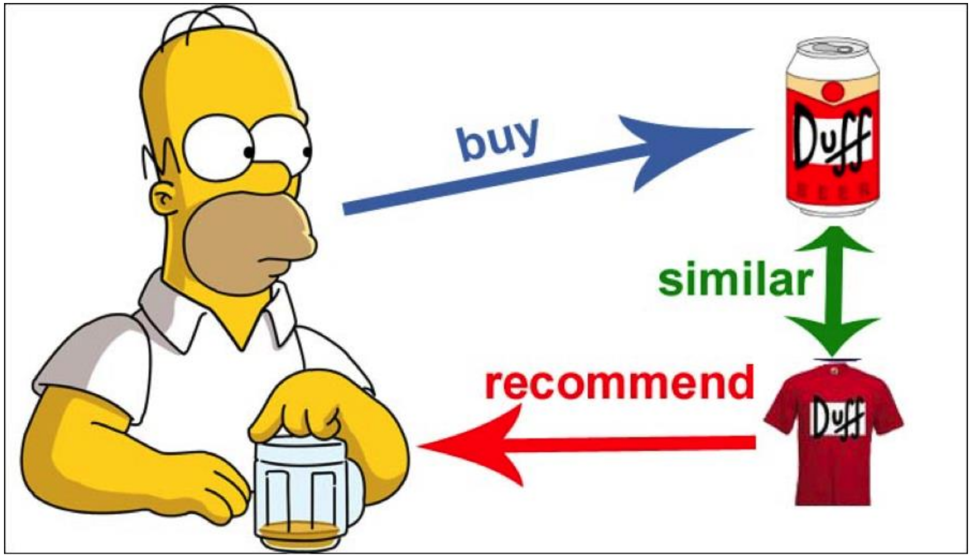
- problems
  - ratings data sparse – prone to skewed correlations
    - nearest neighbor actually not near at all
      - random recommending
  - some users have only single session
    - need to learn fast
  - fails to incorporate agreement about an item in the population
    - controversial vs. loved item
  - expensive calculation of similar users
    - sampling
    - clustering/indexing

Collaborative filtering

- reduce the domain complexity by dimensionality reduction
  - latent topics/tastes/categories in items
  - improve accuracy in predicting
  - reduction of runtime
- SVD and PCA
- recently machine learning
  - e.g., simulated users/ratings to cope with the data sparsity problem

Content-based recommendation

- instead of user-item ratings, we need content information
  - about items (annotation, content-based descriptor)
  - about user profiles (preferences, content-based descriptor)



Content-based recommendation

- CB nearest neighbor recommendation
  - k-nearest neighbors
  - predicting rating for not-yet-seen item
    - find k most similar items (based on features), already rated
    - predict rating based on these
- content-based similarity
  - vector model (bag of words)
    - full-text content
  - other similarity models (also see next lectures)
    - multimedia content
      - metric distances
    - keyword sets A, B
      - Dice coefficient $2*|A \cap B|/(|A|+|B|)$ Jaccard coefficient $|A \cap B|/|A \cup B|$

Content-based recommendation

- classification
  - features $\rightarrow$ rating (like/dislike)
  - machine learning techniques
    - Naive Bayes
    - linear classifiers
    - decision trees
    - neural networks
    - SVM
    - ...

Content-based recommendation

- advantages
  - user independence – does not depend on other users (and so possible sparsity)
  - transparency – explanations, understandable
  - new items can be easily included – no cold start like in CF
- limitations
  - limited content analysis
    - some content not easily extractable (multimedia)
    - domain dependent
    - annotation insufficient/missing
  - overspecialization – too similar items
  - new user – user features missing at the beginning, must be collected

Knowledge-based recommendation

- some application domains
  - not frequently purchased items, few ratings
    - expensive items (supersport car, house, yacht)
    - professional items (surgeon instruments, tools)
  - time span important (tech products)
  - explicit requirements of user (vacation)
- collaborative filtering not applicable
- content-based filtering – insufficient similarity

→ knowledge-based recommendations

Knowledge-based recommendation

- constraint-based
  - explicitly defined conditions
  - constraint satisfaction problem
- case-based
  - similarity to specified requirements

Knowledge-based recommendation

- knowledge base
  - customer (user) properties V_C
  - product (item) features V_{PROD}
  - constraints (on customer properties) C_R
  - filter conditions C_F – relationship between customer and product
  - products C_{PROD} – possible instantiation

$V_C = \{kl_c$: [expert, average, beginner] .... /* level of expertise */
wr_c: [low, medium, high] .... /* willingness to take risks */
id_c: [shortterm, mediumterm, longterm] .... /* duration of investment */
aw_c: [yes, no] .... /* advisory wanted ? */
ds_c: [savings, bonds, stockfunds, singleshares] ...... /* direct product search */
sl_c: [savings, bonds] .... /* type of low-risk investment */
av_c: [yes, no] .... /* availability of funds */
sh_c: [stockfunds, singlshares] .... /* type of high-risk investment */ }

$V_{PROD} = \{name_p$: [text] .... /* name of the product */
er_p: [1..40] .... /* expected return rate */
ri_p: [low, medium, high] .... /* risk level */
$mniv_p$: [1..14] .... /* minimum investment period of product in years */
$inst_p$: [text] .... /* financial institute */ }

$$C_R = \{CR_1: wr_c = high \rightarrow id_c \neq shortterm,$$
$$CR_2: kl_c = beginner \rightarrow wr_c \neq high\}$$

$$C_F = \{CF_1: id_c = shortterm \rightarrow mniv_p < 3,$$
$$CF_2: id_c = mediumterm \rightarrow mniv_p \geq 3 \wedge mniv_p < 6,$$
$$CF_3: id_c = longterm \rightarrow mniv_p \geq 6,$$
$$CF_4: wr_c = low \rightarrow ri_p = low,$$
$$CF_5: wr_c = medium \rightarrow ri_p = low \vee ri_p = medium,$$
$$CF_6: wr_c = high \rightarrow ri_p = low \vee ri_p = medium \vee ri_p = high,$$
$$CF_7: kl_c = beginner \rightarrow ri_p \neq high,$$
$$CF_8: sl_c = savings \rightarrow name_p = savings,$$
$$CF_9: sl_c = bonds \rightarrow name_p = bonds\ \}$$

$$C_{PROD} = \{CPROD_1: name_p = savings \wedge er_p = 3 \wedge ri_p = low \wedge mniv_p = 1 \wedge inst_p = A;$$
$$CPROD_2: name_p = bonds \wedge er_p = 5 \wedge ri_p = medium \wedge mniv_p = 5 \wedge inst_p = B;$$
$$CPROD_3: name_p = equity \wedge er_p = 9 \wedge ri_p = high \wedge mniv_p = 10 \wedge inst_p = B\}$$

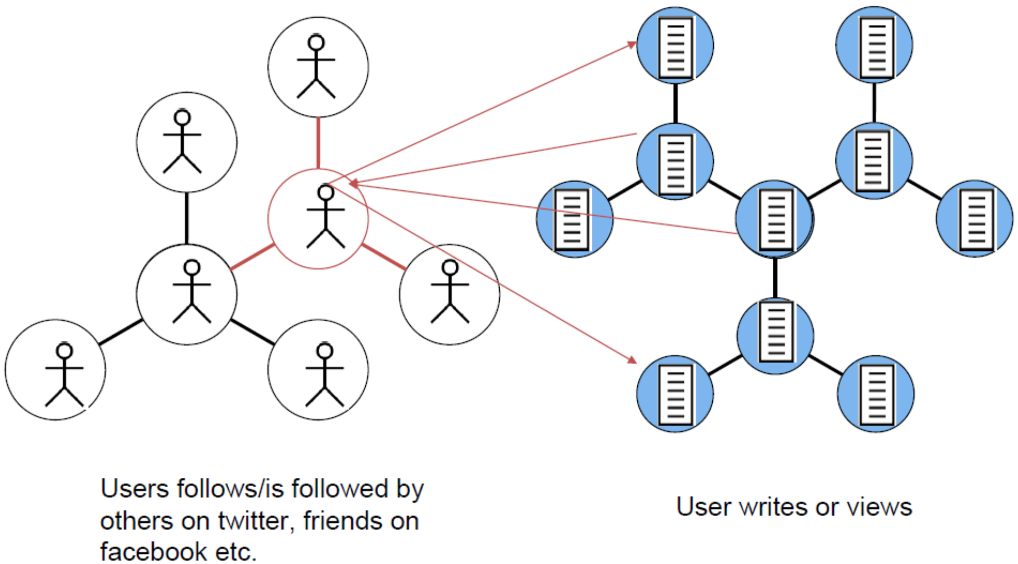
Recommender Systems Handbook; Developing Constraint-based Recommenders

Knowledge-based recommendation

- development of knowledge bases
  - difficult, expensive
  - specialized tools
  - methodology
    - rapid prototyping, detection of invalid constraints
  - unsatisfied constraints
    - problem of empty result – similar to restrictive Boolean query
    - constraint relaxation, hints to the relaxation
  - user guidance
    - elicitation process – conversational dialogs, user profile

Social information retrieval

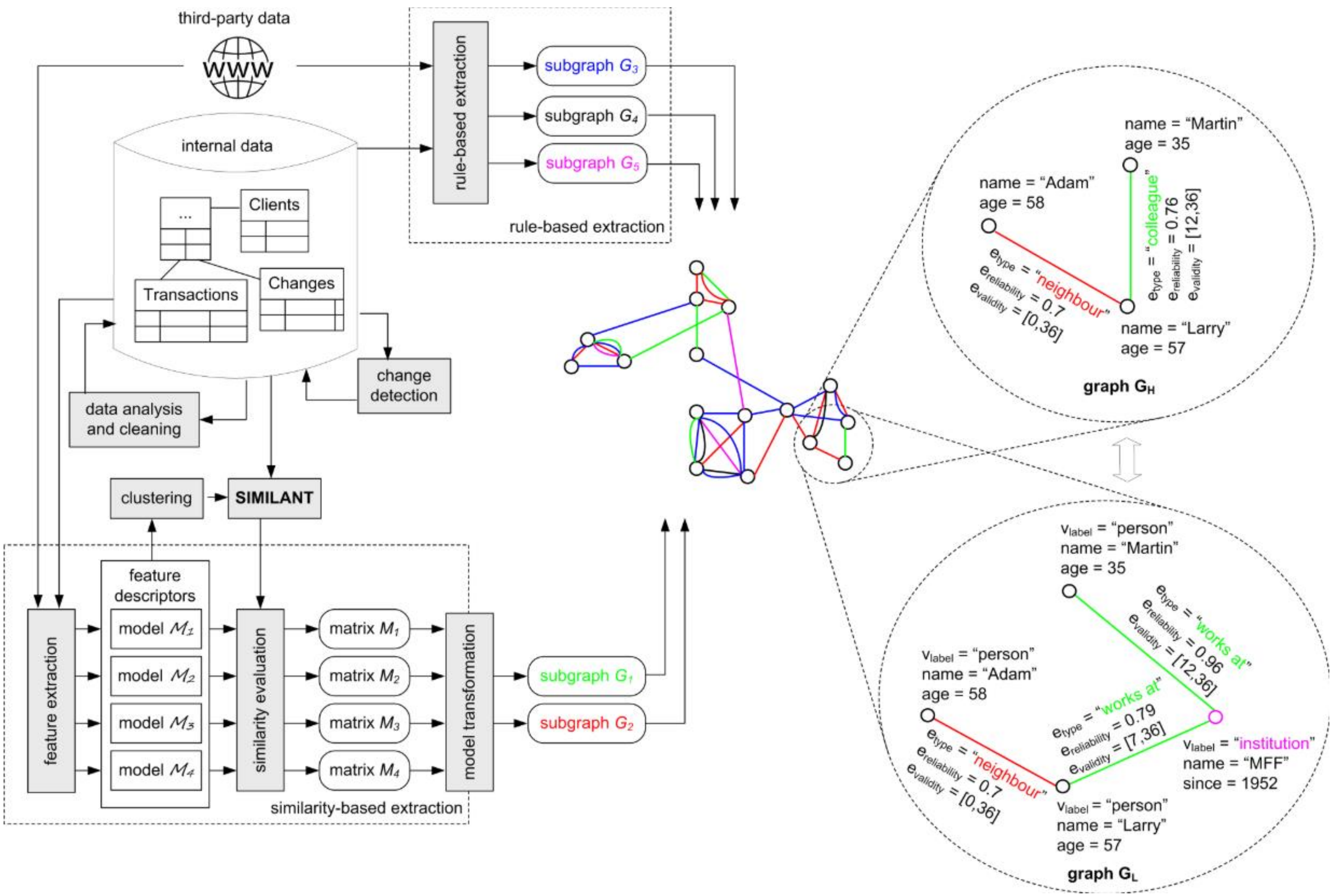
- social network platform
- users and documents (posts)
- multi-graph



Social information retrieval

- inferred social networks
  - constructed from other data by spatial/similarity joins (see later lectures)
  - domain-specific construction, legal issues
- examples
  - network of bank clients
    - similarity join on client profiles
  - COVID-19 suspects
    - spatial join on phone-localized data

Social information retrieval



I. Holubová, M. Svoboda, T. Skopal, D. Bernhauer, L. Peška. **Advanced Behavioral Analyses Using Inferred Social Networks: a Vision**, *TIR 2019*, Springer, 2019

Social information retrieval

- social PageRank
  - different graph, the same purpose as original PageRank
  - a tweeter is reliable/popular if retweeted by other popular tweeters
  - could be fine-grained to some domains (e.g., #hashtags)
    - tweeter reliable in technological topics but not in general topics

Social information retrieval

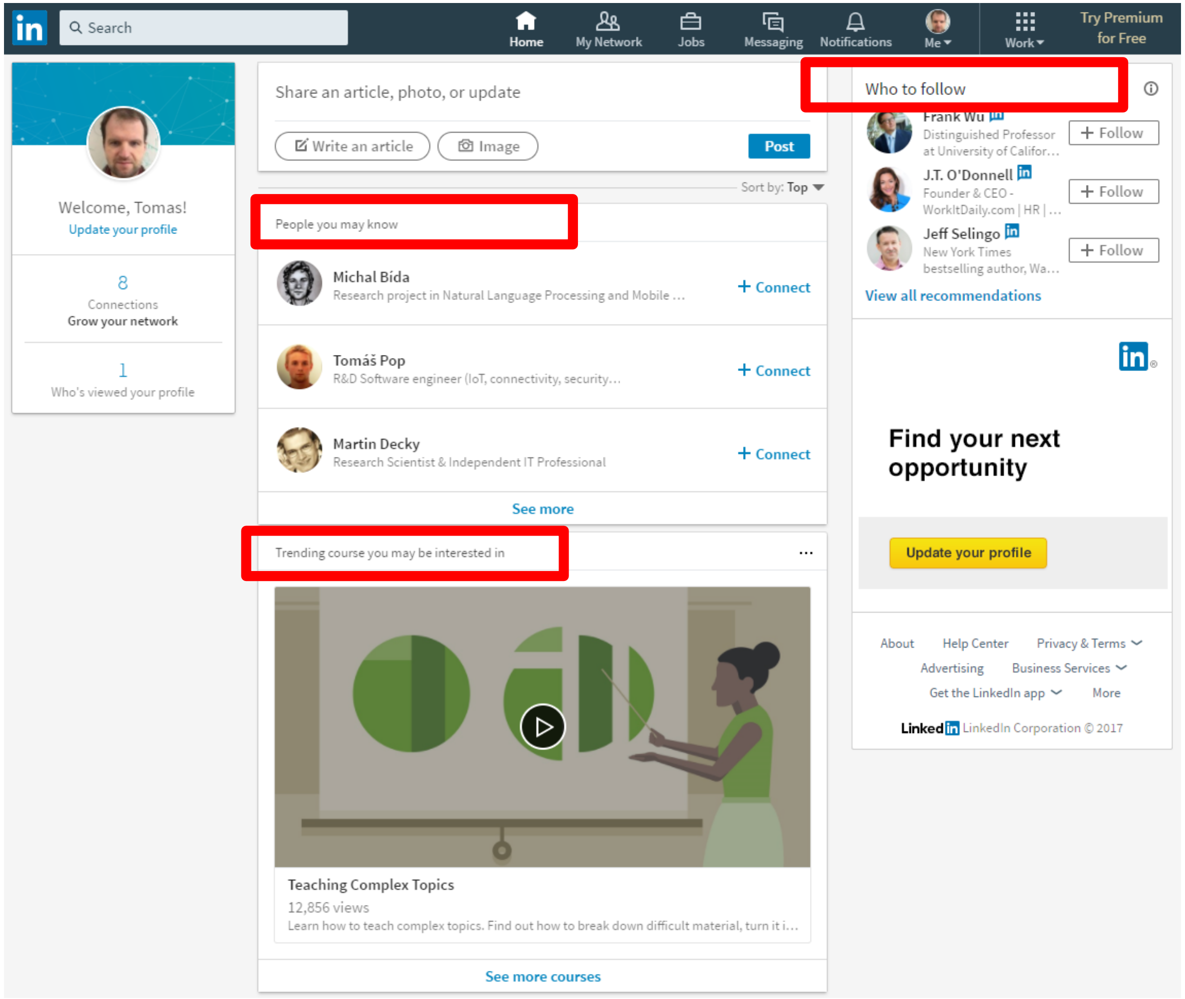
- social recommender systems
  - ratings observes from “likes”, “retweets”, “follows”, etc.
  - set relevant users for recommendation reduced to “friends”
    - collaboration filtering
    - content-based
  - additional ranking of items by “friends”
    - top-k
    - re-ranking
  - global/local trends

Social recommender systems

- recommender systems (RS) for social media domain
- battle the social overload
- offer personalization techniques
  - content
  - ads
- social media and recommender systems benefit from each other

recommender systems	→ RS impact the success of social media as each user is present with most relevant items suiting her needs → ← social media provide new types of data and metadata to be leveraged by RS (tags, comments, likes, relationships) ←	social media

Example (LinkedIn homepage)



Example (YouTube homepage)