Recommending Concepts to Experts

An Exploration of Recommender Techniques for Collaborative Ontology Engineering Environments in the Biomedical Domain.

Simon Walk[1], Markus Strohmaier[1], Tania Tudorache[2], Natalya F. Noy[2], Csongor Nyulas[2] and Mark A. Musen[2]

[1]Graz University of Technology

[2]Stanford Center for Biomedical Informatics Research

Background

- International Classification of Diseases 11 (ICD-11)
  - Consists of (ca.):
    - 33.000 concepts.
    - 38.000 „is-a“ relations.
    - 200 active users.
  - Collaboratively engineered
    - Wikipedia like approach
    - Including the public (e.g. everybody can contribute).

- Possible problems of this approach:
  - Stagnation or decline in user participation.
  - Specific areas of ICD-11 are neglected by users.

Goals

Find and implement means to:

1. increase participation/activity of users
2. Influence/direct user activity

Approach:

- Recommending Concepts to Experts
  - Implementation of Recommender Systems into Collaborative Ontology Engineering Environments.
  - Objective: Identify and suggest concepts (work) of interest for every user to increase participation.

Types of Recommender Systems

1. Content Based Recommender Systems
   – Suggest similar concepts by calculating and comparing similarity between content-related features of each concept.
   – Content: Features/Textual properties of concepts, notes and discussions.
2. Knowledge Based Recommender Systems
   – Suggest similar concepts based on specific domain knowledge.
   – Domain Knowledge: Structural Information, Reasoning, Relationships, Links to other ontologies.
3. Collaborative Filtering
   – Identify concepts or items to suggest based on similar user behavior by identifying and calculating similarity between behavioral patterns or usage patterns of users.
   – Usage Patterns: adding, editing, moving or deleting concepts, properties or individuals.

Content Based Recommendations

F00

F00 ~~'~~Dementia ~~in~~ Alzheimer diseas~~e'~~

Alzheimer diseas~~e is a~~ primari~~y~~ degen~~erative~~ cerebr~~al~~ diseas~~e of~~ unknown etiolog~~y with~~ characterist~~ic~~ neuropatholog~~ical and~~ neurochem~~ical~~ featur~~es. The~~ disord~~er is~~ usual~~ly~~ insidi~~ous in~~ onset ~~and~~ develop~~s~~ slowli~~y but~~ steadili~~y over a~~ period ~~of~~ sever~~al~~ year~~s.~~

Create Word-Count list for user

Word	Count
diseas	3
Alzheimer	2
Dementia	1
cerebr	1
…	…

Calculate Cosine-Similarity

F01

F01 ~~'~~Vascular dementia~~'~~

Vascular dementia ~~is the~~ result ~~of~~ infarct~~ion of the~~ brain due ~~to~~ vascular diseas~~e,~~ includ~~ing~~ hypertens~~ive~~ cerebrovascular diseas~~e. The~~ infarct~~s are~~ usual~~ly~~ small ~~but~~ cumul~~ative in their~~ effect~~.~~ Onset ~~is~~ usual~~ly in~~ later life~~.~~

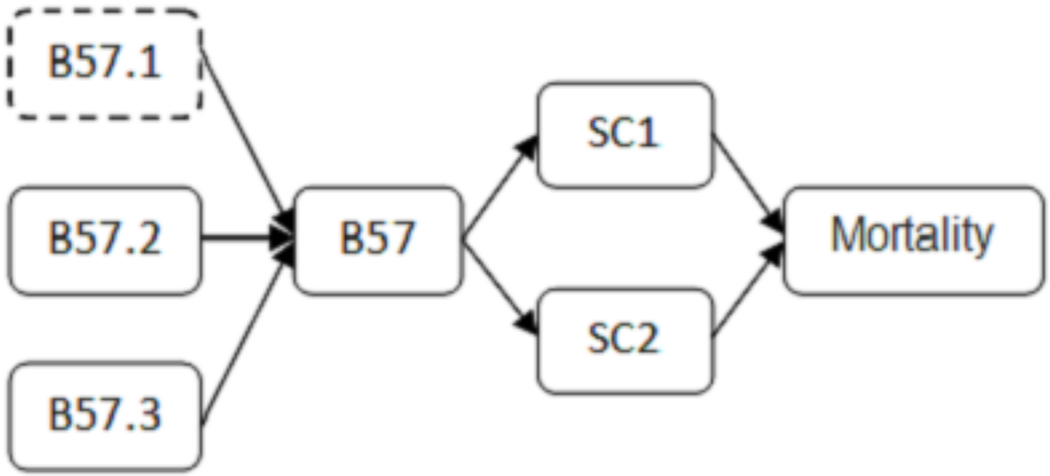
Create Word-Count lists for every concept

Word	Count
diseas	2
Infarct	2
Vascular	2
dementia	2
…	…

Rank Concepts

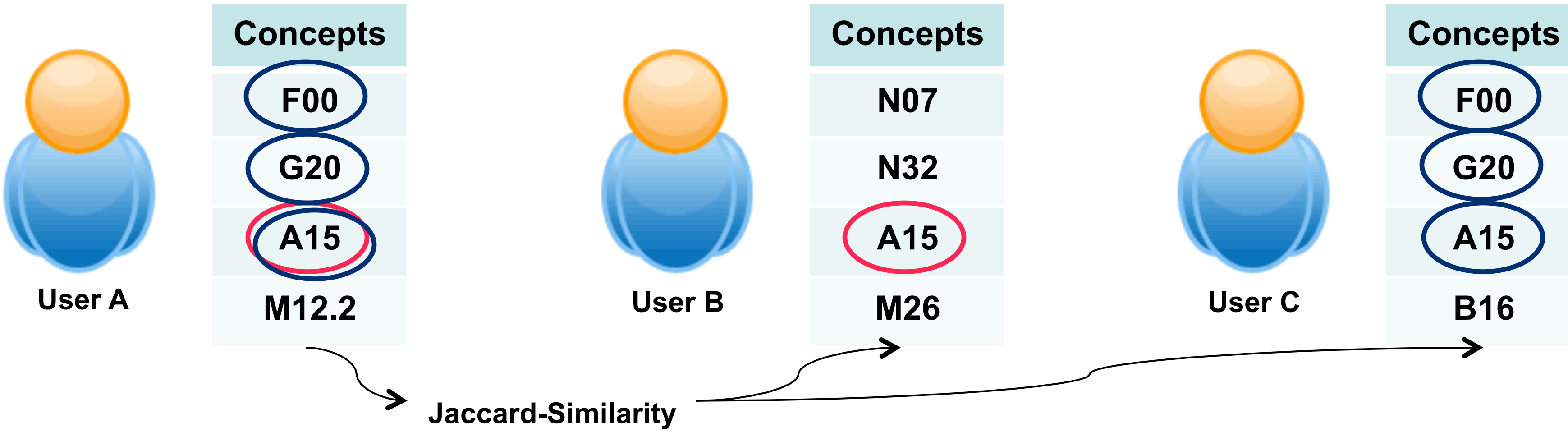
Concept	Rank
F01	1
G30	2
A10	3
M15	4
…	…

Knowledge Based Recommendations



Depth	B57.1	B57.2	B57.3	B57	SC1	SC2	Mortality
1	1	0	0	1	0	0	0
2	1	0	0	2	1	1	0
3	1	0	0	2	2	2	2
4	1	0	0	2	2	2	3

Collaborative Filtering Recommendations



User C is more similar to User A than User B,

thus rank B16 before N07, N32 and M26

Concepts	Rank
B16	1
N07	2
N32	3
…	…

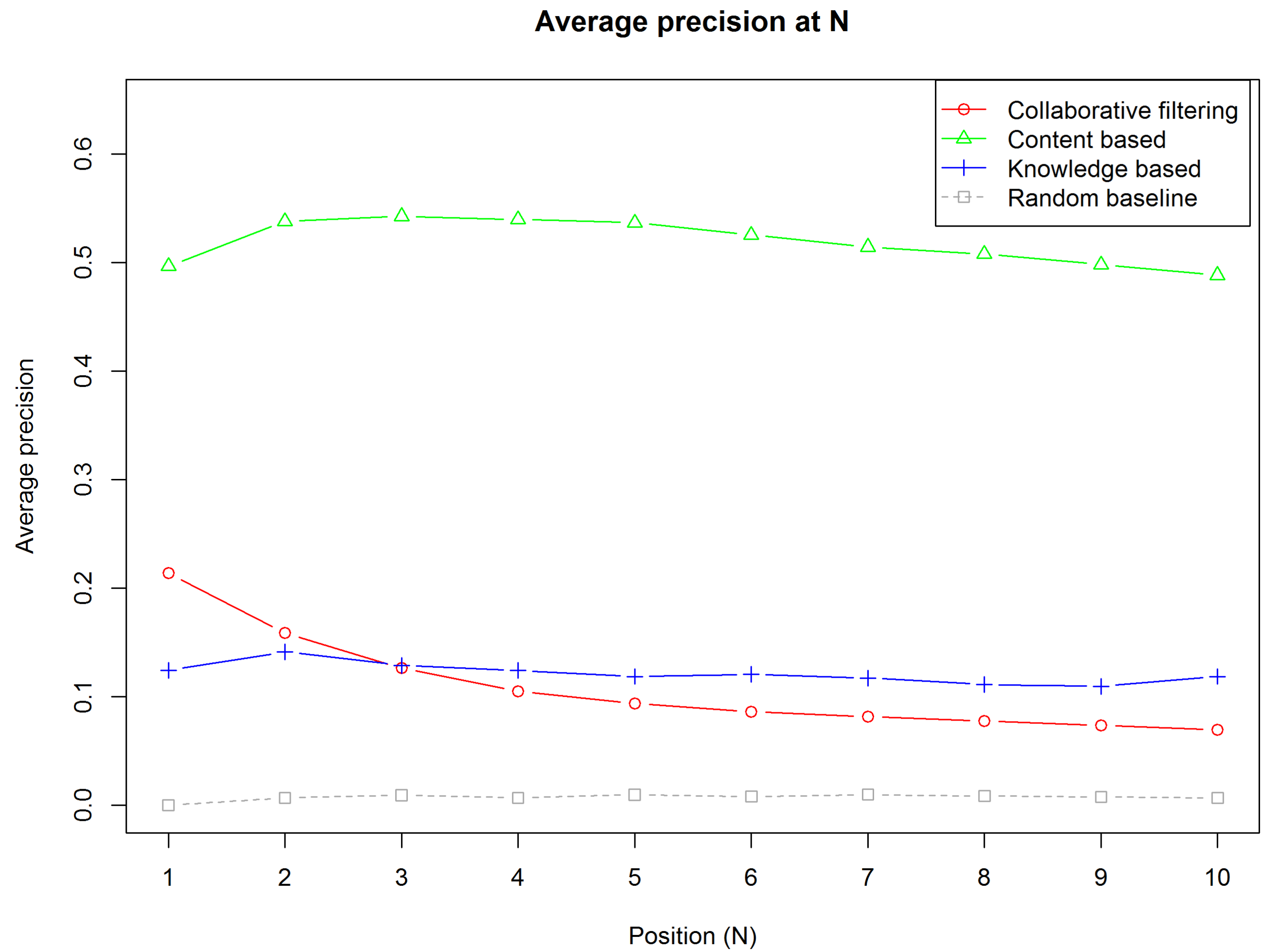
5-Fold Cross Validation Evaluation

1. ICD-11 change data for each user is split into 5 equally (subsequent) sized parts.
   - 4 parts Training
   - 1 part Validation
2. Training Set is used to calculate recommendations
3. **Top N** recommendations (and Validation Set) are used to calculate **Precision at N**.

$$Precision = \frac{|\text{found relevant documents}|}{|\text{found documents}|}$$

4. Found relevant documents represents the number of concepts within the calculated Top N recommendations, which are **also in the validation set**.
5. Found documents is incremented from 1 to 10 according to current position N.

Results


Average precision at N
Collaborative filtering
Content based
Knowledge based
Random baseline
Average precision
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
7
8
9
10
Position (N)

Conclusions

1. Recommender systems perform better than randomly suggesting concepts to work on.
2. By introducing a (wanted) bias, recommender systems could be used to distribute participation across an ontology.
3. Traditional recommender systems have to be adapted to work in Collaborative Ontology Engineering Environments.
4. In the case of ICD-11, Content-Based Recommendations clearly outperform other approaches (yet).

Future Work

1. Refine approaches to calculate recommendations (e.g. taking the amount of changes performed on each concept into account).
2. Re-Evaluate Collaborative Filtering Recommendations after ICD-11 goes public
3. Re-Evaluate (and identify) different types of domain knowledge.
4. Further evaluations (e.g. by asking contributors if the generated recommendations are useful) are planned.

Thank you for your attention

Content Based Recommendations

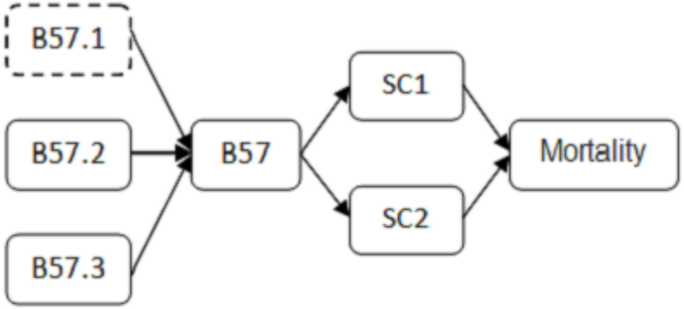
1. Get the set of all previously changed concepts for every user.
2. Create word count list of all previously edited concepts for every author by processing title and definition of these concepts.
   1. Remove stop-words (e.g. is, as, and, so etc.)
   2. Stemming (reduce words to their stem)
   3. Remove special characters (e.g. *, #, ‘ etc.)
   4. Aggregate and count words
3. Create a word count list (using the same four steps) for every concept in the ontology.
4. Calculate cosine similarity between both word count lists and rank concepts accordingly.

This approach identifies concepts, which have a similar vocabulary as previously changed concepts of a user.

Knowledge Based Recommendations

1. Get the set of all previously changed concepts for every user
2. Traverse along all neighboring concepts (“is-a”) until a predefined threshold is reached.
3. Store how often each concept was encountered.
4. The concepts with the highest amount of encounters are stored and suggested to a user.

Depth	B57.1	B57.2	B57.3	B57	SC1	SC2	Mortality
1	1	0	0	1	0	0	0
2	1	0	0	2	1	1	0
3	1	0	0	2	2	2	2
4	1	0	0	2	2	2	**3**



This approach identifies concepts that are highly related to previously changed concepts.

Collaborative Filtering Recommendations

1. Get the set of all previously changed concepts for every user.
2. Calculate Jaccard similarity between all sets of edited concepts (users).
3. Identify “most similar” users for a specific user ranked according to Jaccard similarity.
4. Store/Recommend all changed concepts of similar authors, which have not been changed by that specific user before.

This approach identifies concepts that are of interest to users that were interested (e.g. have edited) in the same concepts.