NTNU

Department of Computer Science

Examination paper for TDT4215 Web Intelligence

Academic contact during examination: Prof. Jon Atle Gulla

Phone: 73591847

Examination date: 6 June 2017

Examination time: 09:00 – 13:00

Permitted examination support material: D (No printed or hand-written support material is allowed. A specific basic calculator is allowed)

Other information:

Language: English

Number of pages (front page excluded): 4

Number of pages enclosed: 5

Informasjon om trykking av eksamensoppgave
Originalen er:
1-sidig ☐ **2-sidig** ☐
sort/hvit ☐ **farger** ☐
skal ha flervalgskjema ☐

Checked by:

_____

Date Signature

Students will find the examination results in Studentweb. Please contact the department if you have questions about your results. The Examinations Office will not be able to answer this.

Question 1. Neighborhood-Based Collaborative Filtering

From Aggarwal: Recommender Systems, Chapter 2.

(a) Explain the "*long tail property*"?

(b) Given the following User-Item table, estimate (using a *user-based approach*) the rating of *Item1* for *User3* and the rating of *Item4* for *User5*.

	Item1	*Item2*	*Item3*	*Item4*	*Item5*
User1	5	4	1	2	2
User2	3	4	2	1	3
User3	**?**	5	3	2	1
User4	1	2	5	3	4
User5	2	3	5	**?**	3

(c) What are the main strengths and weaknesses of the Neighborhood-based methods?

Question 2. Model-Based Collaborative Filtering

From Aggarwal: Recommender Systems, Chapter 3.

(a) What are the main advantages of model-based methods in comparison to NN-based methods?

(b) Given the following binary ratings matrix, estimate the ratings of *Item3* for *User4* and the rating of *Item5* for *User5* *using the Bayes method.*

	Item1	*Item2*	*Item3*	*Item4*	*Item5*
User1	1	1	-1	-1	1
User2	-1	1	1	1	-1
User3	1	-1	-1	1	-1
User4	1	-1	**?**	1	1
User5	1	-1	-1	1	**?**

(c) What is the main intuition behind the Matrix Factorization approach? What are the main methods?

Question 3. Content-Based Recommender Systems

From Aggarwal: Recommender Systems, Chapter 4.

(a) The Gini Index is used to assess a word's discriminative properties on a rating. Assume the following table that shows the presence of three words (i.e. three football players) in news stories that have been rated as Interesting or not interesting by the user:

News story	*"Messi"*	*"Ronaldo"*	*"Zlatan"*	*Rating*
1	Yes	Yes	Yes	Interesting
2			Yes	Interesting
3	Yes	Yes	Yes	Not interesting
4		Yes		Not interesting
5	Yes		Yes	Interesting
6		Yes	Yes	Not interesting
7	Yes		Yes	Interesting
8		Yes	Yes	Not interesting
9		Yes	Yes	Interesting
10	Yes			Interesting

Compute the *Gini index* for the three words. Which football player seems to the most interesting one to this user?

(b) We use nearest neighbor (k-NN) classification to recommend news stories to users. How do you think stop word removal, stemming or phrase extraction would affect the results from k-NN? Justify your answer. How can you reduce the computational complexity of k-NN?

Question 4. Ensemble-Based and Hybrid Recommender Systems

From Aggarwal: Recommender Systems, Chapter 6.

Explain the main differences between the *feature combination* and the *feature augmentation* hybrids methods.

Question 5. Evaluating Recommender Systems

From Aggarwal: Recommender Systems, Chapter 7.

Given the following dataset with book ratings from users and predicted ratings from our recommender system:

Nr	User ID	Book ID	User rating (r_i)	Predicted rating (p_i)
1	1	2-140	2	2.5
2	1	2-90	1	3
3	2	1-120	5	4
4	2	2-140	3	2.6
5	2	1-55	3	3.2
6	3	3-80	5	5
7	4	1-120	4	3.6
8	4	3-10	2	2.2
9	4	2-140	3	4
10	5	3-10	2	1.8

(a) Compute the *Mean Absolute Error (MAE)* of the predicted ratings.
(b) Make your own assumptions and compute a *precision* value for the predicted ratings. Explain the assumptions you make.
(c) Define *user-space coverage* and *item-space coverage*. Why is *catalog coverage* often a more useful measure than normal item-space coverage?

Question 6. Time- and Location-Sensitive Recommender Systems

From Aggarwal: Recommender Systems, Chapter 9.

Explain the main idea behind the *recency-based models* in temporal collaborative filtering.

Question 7. Querying the Semantic Web

From Antoniou et al., A Semantic Web Primer, Chapter 2 & 3

You want to build and RDF model (RDF triples) of the 100 biggest mountains in Norway. For each mountain the model should specify the mountain's Norwegian name, how high it is, and in which municipality it is located. Some (but not all) mountains also have a Sami name.

(a) Explain how you would build this model with RDF triples and show an example of how a particular mountain will be described in RDF.
(b) Formulate a query in SPARQL that lists the names (Norwegian name and Sami name if available) of all mountains in the database.
(c) Formulate a query in SPARQL that lists all municipalities that have at least two mountains higher than 1,500 m.

Question 8. Personalized Information Access using Semantic Knowledge

(a) The User Behaviour Ontology (UBO) describes all events relevant for modeling user behavior such as user clicks or mouse-over events. What are the two main goals that UBO serves?

(b) In the paper link prediction is used for the enrichment of user profiles. It aims to find important related items from a semantic dataset and to infer missing links in an observed graph that are likely to exist. What are the main advantages of enriching user profiles?