LINKÖPING UNIVERSITY
Institutionen för datavetenskap
Patrick Lambrix / Olaf Hartig / Valentina Ivanova / Zlatan Dragisic

Tentamen

DF22300 Advanced Data Models and Databases
2016

Grades:

For a pass grade all questions need to be answered and almost all correctly answered.

Instructions:

- This is an *individual* exam. If you have questions, ask the course leader.
- Send your answers to: Patrick Lambrix, Institutionen för datavetenskap, Linköpings universitet, 581 83 Linköping.
- Deadline: February 1, 2017.
- Write a contact e-mail address on your hand-in.
- If there are references to papers in the questions, then links to the papers will be on the course home page.

Question 1: Information Retrieval

Assume that we have two documents in our document base. Document 1 contains ’enzyme’ 5 times, ’gene’ 10 times, ’protein’ 0 times and ’signal’ 8 times. Document 2 contains ’enzyme’ 0 times, ’gene’ 0 times, ’protein’ 7 times and ’signal’ 1 time.

(i) Assume that we use the vector model for information retrieval. Assume that we are only interested in the words gene, enzyme, protein and signal.
1. Explain tf and idf in the vector model.
2. In which cases is a weight w$_{ij}$ in a document vector equal to 0?
3. Give the document representations for Document 1 and Document 2 according to the tf-idf model.

(ii) Assume that we use the boolean model for information retrieval. Assume that we are only interested in the words gene, enzyme, protein and signal.
1. Give the document representations for Document 1 and Document 2 according to the Boolean model.
2. Represent the query for all documents containing gene or protein, but not signal. Compute then the completed DNF (disjunctive normal form) of the query.

Question 2. Semi-structured data

(i) Below, a (simplified) description of the biological data source PIR-PSD, is given. This data source contains information about protein sequences. Each entry in the data source has information about the entry (identification numbers, dates of creation and update), names of the protein sequence and the parts that it contains, the organism from which the sequence was taken, references to the literature describing the sequence, references to other related databases and the protein sequence. The simplified PIR-PSD schema is given as a DTD. Draw a strong data guide for PIR-PSD. Use the OEM model.

```
<!-- ****************************************************************
PIR-International Protein Sequence Database (PSD)
*************************************************************** -->

<!-- ProteinEntry: the root element. -->
<!ELEMENT ProteinEntry (header,protein,organism,reference*,sequence)>

<!-- header: database information. -->
<!ELEMENT header (uid,accession*,created_date,seq-update_date)>

<!ELEMENT accession (#PCDATA)> <!-- accession number -->
<!ELEMENT created_date (#PCDATA)> <!-- date (DD-MMM-YYYY) -->
<!ELEMENT seq-update_date (#PCDATA)> <!-- date (DD-MMM-YYYY) -->

<!-- protein: the protein-names. -->
<!ELEMENT protein(name,alt-name*,contains*)>

<!ELEMENT name (#PCDATA)> <!-- protein name -->
<!ELEMENT alt-name (#PCDATA)> <!-- alternate protein name -->
<!ELEMENT contains (#PCDATA)> <!-- activity name -->

<!-- organism: identification of the biological source. -->
<!ELEMENT organism (source,common-name?,note*)>

<!ELEMENT source (#PCDATA)> <!-- source name -->
<!ELEMENT common-name(#PCDATA)> <!-- common name -->
```

```
<!-- reference -->
<!ELEMENT reference (refinfo,note*)>

<!-- refinfo: identification of the literature source. -->
<!ELEMENT refinfo (authors,citation,title?,xrefs?)>

<!ELEMENT authors (author+)> <!-- list of authors -->
<!ELEMENT author (#PCDATA)> <!-- author name -->
<!ELEMENT citation (#PCDATA)> <!-- citation name -->
<!ELEMENT title (#PCDATA)> <!-- title text -->

<!ELEMENT xrefs (xref+)> <!-- cross-references -->
<!ELEMENT xref (db,uid)> <!-- a cross-reference -->
<!ELEMENT db (#PCDATA)> <!-- database tag -->

<!-- sequence: the amino acid sequence. -->
<!ELEMENT sequence (#PCDATA)> <!-- amino acid symbols and
 punctuation -->

<!-- General elements. Elements that can be contained in several
 other elements. -->
<!ELEMENT note (#PCDATA)> <!-- note text -->
<!ELEMENT uid (#PCDATA)> <!-- entry identifier -->
```

(ii) In figures 1 and 2 (simplified and slightly modified) data from a SwissProt data source is given. (The dashed nodes in the figures refer to nodes in the other figure. Both figures together make up the data source.) Draw a strong data guide for this data. Use the OEM model.

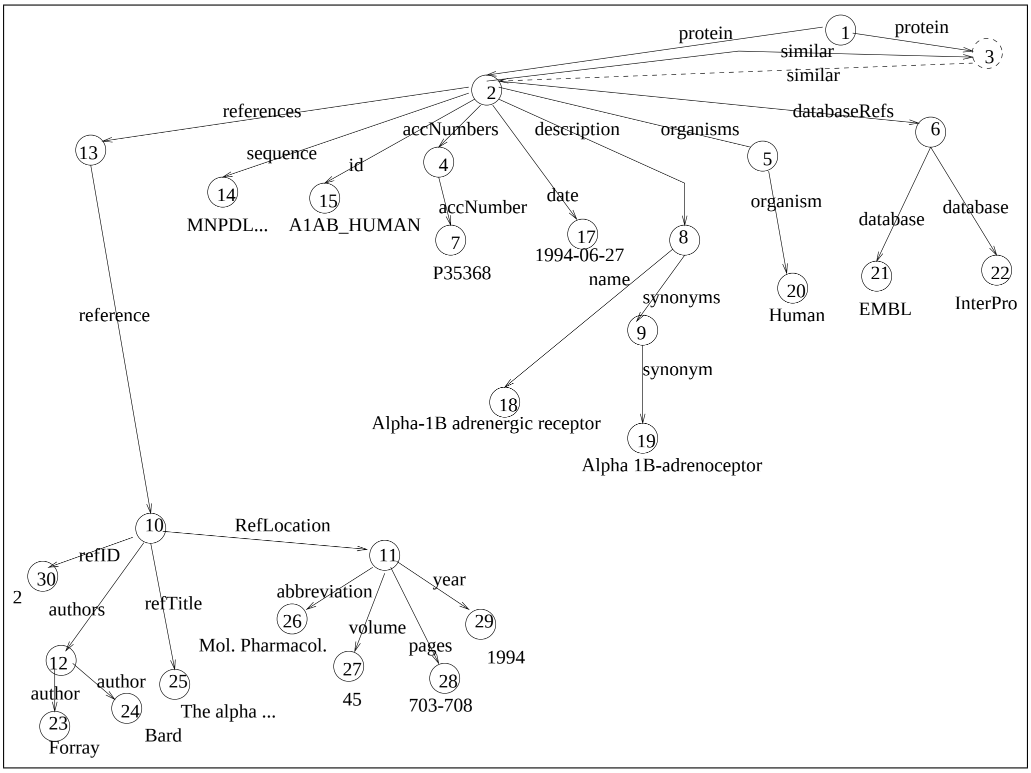

Figure 1: SwissProt db -1

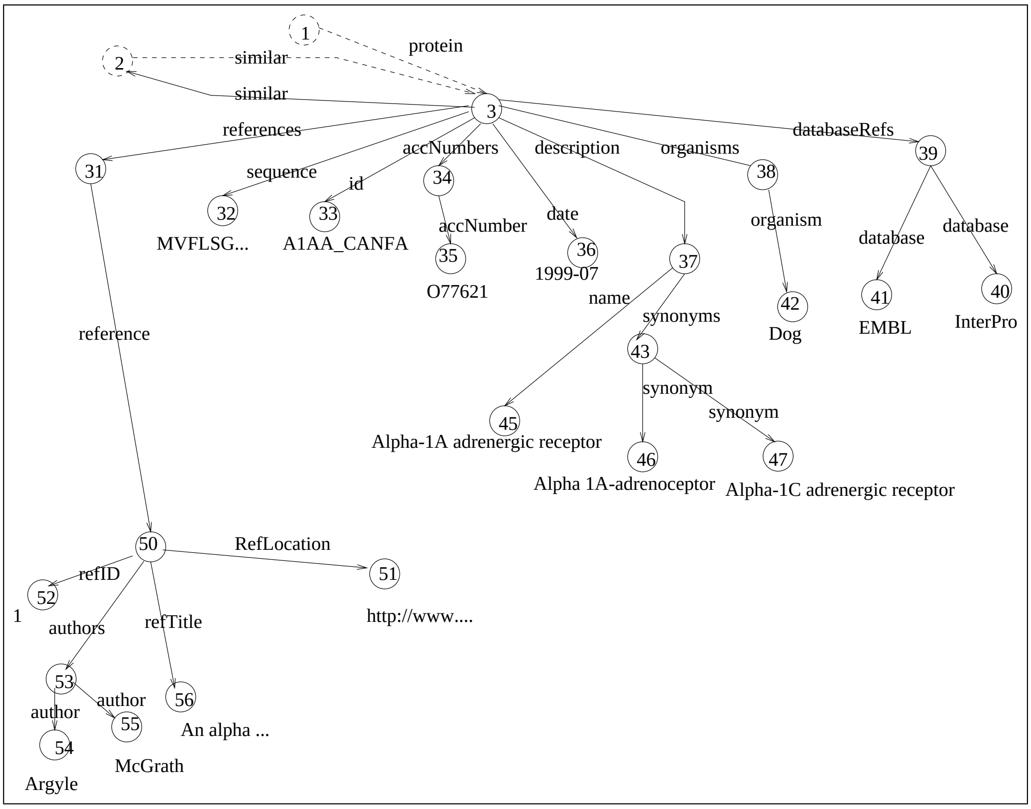

Figure 2: SwissProt db -2