University of Mannheim

Prof. Dr. Christian Bizer

IE670 Web Data Integration

Example Exam Questions

Solutions are found on page 3.

Example Question 1 (5 Points):

For which type of strings do edit-based string similarity measures deliver good results (1.5 points)? For which type do token-based similarity measures (assume token = word) deliver good results(1.5 points)? How does the Monge-Elkan string similarity measure try to combine the advantages of both classes of measurements (2 points)?

Example Question 2 (5 Points):

Which problems does the standard blocking approach have with respect to recall and bucket size (3 points)? How does the Sorted Neighborhood Method try to overcome these problems (2 points)?

Example Question 3 (5 Points):

Given the XML file shown below, specify XPath expressions for:

(1) retrieving the total population counts of all settlements (as a list of separate numbers, not the sum of all counts) (2 points)

(2) retrieving the names of those settlements for which the population timestamp (*asof* attribute) is 2013 or later. (3 points)

```
<?xml version="1.0" encoding="UTF-8"?>
<settlements>
      <city>
            <name>Mannheim</name>
            <country>Germany</country>
            <population asof="2013">
                  <total>294627</total>
```

```
            <density>2000</density>
        </population>
    </city>
    <city>
        <name>Akaigawa</name>
        <country>Japan</country>
        <population asof="2013">
            <total>1264</total>
            <density>4.5</density>
        </population>
    </city>
    <city>
        <name>Windsor</name>
        <country>UK</country>
        <population asof="2012">
            <total>26885</total>
        </population>
    </city>
    <city>
        <name>Berlin</name>
        <country>Germany</country>
        <population asof="2014">
            <total>3517424</total>
            <density>3900</density>
        </population>
    </city>
</settlements>
```

Answers:

Example Question 1:

For which type of strings do edit-based string similarity measures deliver good results (1.5 points)? For which type do token-based similarity measures (assume token = word) deliver good results(1.5 points)? How does the Monge-Elkan string similarity measure try to combine the advantages of both classes of measurements (2 points)?

Answer:

- **Edit-based** similarity measures work well on strings containing typos (e.g. the and teh).
- **Token-based** similarity measures work well on strings containing different orders of words (e.g. name and surname).
- **Monge-Elkan** makes use of edit-based similarity measures to compare the tokens of a string.

Example Question 2:

Which problems does the standard blocking approach have with respect to recall and bucket size (3 points)? How does the Sorted Neighborhood Method try to overcome these problems (2 points)?

Answer:

- **Recall**: Matching records might end up in different buckets and are thus not compared, which leads to a lower recall, as this match will not be found by the identity resolution.
- **Bucket Size**: The size of the buckets can vary drastically.
- **SNM**: Sorts all records by the blocking key and compares all records within a specific window size, and therefor has less problems with the two mentioned problems.

Example Question 3:

Given the XML file shown above, specify XPath expressions for:

(1) retrieving the total population counts of all settlements (as a list of separate numbers, not the sum of all counts) (2 points)

(2) retrieving the names of those settlements for which the population timestamp (*asof* attribute) is 2013 or later. (3 points)

Answer:

(1) //total/text()

(2) //city/population[@asof>=2013]/../name/text()