Max-Planck-Institut für Informatik

Databases and Information Systems Group (AG 5)

Prof. Dr.-Ing. G. Weikum Dipl.-Inform. Hanglin Pan
Dipl.-Inform. Sergej Sizov
Dipl.-Inform. Stefan Siersdorfer

Selected Topics in Web Information Retrieval and Mining (WS 03/04)

Assignment 3

Handout on: Thursday, November 20, 2003

Due on: solutions will be discussed on Thursday, Nov 27, 2003

Exercise 3.1:

Consider "reshaping" a vector-space-norm-based similarity function between documents by changing weights according to relevance feedback.
How can you do this if your original similarity function is the cosine measure?
Discuss the consequences of your approach (*e.g., which aspects of the formal vector space model may need to be adjusted, which complications or additional costs in a system implementation may arise*).

Exercise 3.2:

Consider a linearly weighted multi-criteria distance function with two or three different criteria (*e.g., **relevance** and **authority** as two criteria, or **relevance**, **authority**, and **recency** as three criteria*).
Develop closed-form solutions for the optimal weights derived by linear regression from user feedback.

Exercise 3.3:

Reconsider the separation-based approach to query expansion.
How many single-feature query candidates would you add to a given query, based on the feedback?
Can you think of alternative criteria for identifying good candidates?
When would you remove a feature from the given query?

Exercise 3.4:

How can you carry over the document-oriented notion of ranked retrieval relevance feedback to structured data? Consider, for example, a user searching an electronics product catalog. The catalog may contain structured data on notebooks with the following (relational database) schema:

Notebooks (Vendor varchar, Model varchar, Processortype varchar, Clockrate float, Memory integer, DiskCapacity integer, DiskSpeed float, Price integer, CDwriter boolean, DVDwriter boolean, ...)

The user may ask queries of the kind:

Processortype=Centrino and ***Clockrate>=1.4 [GHz]*** and ***Memory >=1 [GB]*** and ***DiskCapacity>=60 [GB]*** and ***Price<1000 [Euro]*** and ***DVDwriter=true***

If such queries were posed in SQL and evaluated according to the usual Boolean retrieval model, the user would often obtain empty results or would be flooded with too many matches. So the user would rather prefer a ranked retrieval model where the system returns a ranked list of approximate matches.

How would you define a similarity function between records and the similarity between queries and records?
What kind of user relevance feedback would you like to have supported by the system?
How should the feedback influence the query?
Try to cast your answers into a formal model.

Exercise 3.5:

Now to carry over the notion of ranked retrieval with relevance feedback to semi-structured hierarchical data like XML documents (essentially following the considerations of Exercise 3.1).
Consider, for example, a collection of music CD descriptions where a typical XML document may look as follows:

```
<?xml version="1.0"?>
<cds>

    <cd>
      <title>Time Out</title>
      <artist>Dave Brubeck Quartet</artist>
      <recording date="June-August 1959" place="NYC"/>
      <catalogno label="Sony/CBS" number="Legacy CK 40585" format="CD"/>
      <category>Jazz</category>
      <price>9.99 USD</price>
      <personnel>
            <player name="Dave Brubeck" instrument="piano"/>
            <player name="Paul Desmond" instrument="alto sax"/>
            <player name="Eugene Wright" instrument="bass"/>
            <player name="Joe Morello" instrument="drums"/>
      </personnel>
      <tracks>
            <track title="Blue Rondo a la Turk" credit="Brubeck" timing="6m42s"/>
            <track title="Strange Meadow Lark" credit="Brubeck" timing="7m20s"/>
            <track title="Take Five" credit="Desmond" timing="5m24s"/>
            <track title="Three To Get Ready" credit="Brubeck" timing="5m21s"/>
            <track title="Kathy's Waltz" credit="Brubeck" timing="4m48s"/>
            <track title="'Everybody's Jumpin'" credit="Brubeck" timing="4m22s"/>
            <track title="Pick Up Sticks" credit="Brubeck" timing="4m16s"/>
      </tracks>
      <notes>Possibly the DBQ's most famous album, this contains
          <trackref link="#3">Take Five</trackref>, the most famous jazz track
          of that period.  These experiments in different time signatures are
          what Dave Brubeck is most remembered for.  Recorded Jun-Aug 1959 in
          NYC. See also the sequel,<albumref link="cbs-timefurthout">Time Further Out</albumref>.
```

```
    </notes>
    <review>
        <reviewer name="anonymous" date="December 14, 2000" comments="Excellent!!!"/>
        <reviewer name="Renato Buchert" date="July 26, 2002" comments="Essential for Jazz Fans..."/>
        <reviewer name="Ask Manny" date="August 28, 2003" comments="Time out, timeless..."/>
    </review>
  </cd>

  <cd>
    <title>Fascinoma</title>
    <artist>Jon Hassell</artist>
    <recording date="June 1999"/>
    <catalogno label="Water Lily Acoustics" number="WLACS70CD" format="CD"/>
    <category>Jazz</category>
    <price>16.99 USD</price>
    <personnel>
        <player name="Jon Hassell" instrument="trumpet"/>
    </personnel>
    <tracks>
        <track title="Nature Boy"/>
        <track title="Datura"/>
        <track title="Caravanesque"/>
        <track title="Wide Sky"/>
        <track title="Mevlana Duke"/>
        <track title="Secretely Happy"/>
        <track title="Poinciana"/>
        <track title="Sensuendo"/>
        <track title="Suite de Caravan"/>
        <track title="Estate (Summer)"/>
    </tracks>
    <notes>This is a rare chance to hear the "fourth world" experimenter's trumpet virtually
        devoid of its accustomed heavily-processed sound. This album is also Hassell's
        first stab at playing cover versions, a partially nostalgic exercise in revisiting
        moments from his childhood, tuning in to unfamiliar sounds over the airwaves.
    </notes>
    <review>
        <reviewer name="anonymous" date="January 8, 2000" comments="Perfect"/>
        <reviewer name="costisd" date="August 10, 2001"
                comments="Misty like a winter evening. Gradually, evening images give
                their place to night dreams. It is incredible that such things can happen on disc..."/>
    </review>
  </cd>
  ...
</cds>
```

Suppose the user is looking for inexpensive Jazz CDs with lots of brass instruments (trumpet, horn, trombone, tuba) that have recently received good reviews (e.g., containing words like "excellent", "very good", "love this ...", etc.). The system would return approximate matches in some ranking order.
What kind of relevance feedback would you like to be supported by the system?
How should the system change queries based on the user's feedback?

Exercise 3.6:

Prove the optimality of the solution for adjusting a query and a quadratic-form feature-similarity matrix, based on relevance feedback. (*This is a difficult exercise; the solution makes uses of Lagrange multipliers. You may consider studying the paper by VLDB 1998 paper by Ishikawa et al.*)