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 4

Handout on: Thursday, November 27, 2003

Due on: solutions will be discussed on Friday, Dec 05, 2003

Exercise 4.1:

Discuss how to determine sessions in a query log from

a) a search engine's web server log (i.e., incoming http requests from all users of the search engine),
b) your own browser history (i.e., outgoing http requests).

Which difficulties do you need to address? How would you tackle them?

Exercise 4.2:

Discuss how you can apply the Apriori frequent-itemset mining algorithm
(see, e.g., Chapter 14, Section 14.2, of the Information Retrieval and Data Mining course
in Winter Semester 2002/03, http://www-dbs.cs.uni-sb.de/~irdm02/skripte/eng/irdm-kap14.ppt)
to find the most important query-log-induced correlations between Web pages
and/or query keywords.

Exercise 4.3:

Write more detailed pseudo-code for the incremental DBSCAN clustering algorithm.

Consider also the following generalization of the notion of clusters

(which supports arbitrarily shaped, possibly non-convex, clusters):
The neighbourhood N(p) of a point p is the set of points q such that dist(p,q) <= max_dist.
Point p is directly density-reachable from point q if p is in N(q) and |N(q)| >= min_points.
Point p is density-reachable from q if there is a finite chain q=p0, p1, ..., pn=p such that
pi is directly density-reachable from p(i-1).
Points p and q are density-connected if there is a point c such that both p and q are
density-reachable from c.
A cluster C is a non-empty subset of the data points with the following conditions:
1) if p is in C and q is density-reachable from p then q is also in C, and
2) all points in C are density-connected.

Exercise 4.4:

Assume you are using information that has been mined from query logs for automatic query expansion.
Now consider the fact that the user's thematic preferences evolve over time.
How could you take this into account in the query-log mining and/or query-expansion algorithms?