Kumar Rajesh
201614060112

Data base Techniques

Home Work Three

Kumar Rajesh
11/11/2016

HOMEWORK THREE

Q:1 Describe benefits and drawbacks of a source-driven architecture for gathering of data at a data warehouse, as compared to a destination-driven architecture?

Answer: In a destination-driven architecture for gathering data, data transfers from the data sources to the data-warehouse are based on demand from the warehouse, whereas in a source-driven architecture, the transfers are initiated by each source.

The benefits of a source-driven architecture are

1. Data can be propagated to the destination as soon as it becomes available. For a destination-driven architecture to collect data as soon as it is available, the warehouse would have to probe the sources frequently, leading to a high overhead.
2. The source does not have to keep historical information. As soon as data is updated, the source can send an update message to the destination and forget the history of the updates. In contrast, in a destination-driven architecture, each source has to maintain a history of data which have not yet been collected by the data warehouse. Thus storage requirements at the source are lower for a source-driven architecture. On the other hand, a destination-driven architecture has the following advantages.
3. In a source-driven architecture, the source has to be active and must handle error conditions such as not being able to contact the warehouse for some time. It is easier to implement passive sources, and a single active warehouse. In a destination-driven architecure, each source is required to provide only a basic functionality of executing queries.
4. The warehouse has more control on when to carry out data gathering activities, and when to process user queries; it is not a good

Q:2 Suppose half of all the transactions in a clothes shop purchase jeans, and one third of all transactions in the shop purchase T-shirts. Suppose also that half of the transactions that purchase jeans also purchase T-shirts. Write down all the (nontrivial) association rules you can deduce from the above information giving support and confidence of each rule.

ANS:

Intuitively, clustering refers to the problem of finding clusters of points in the given data. The problem ofclusteringcan be formalized from distance metrics in several ways. One way is to phrase it as the problem of grouping points into k sets (for a given k) so that the average distance of points from the centroid of their assigned cluster is minimized.

Another way is to group points so that the average distance between every pair of points in each cluster is minimized. There are other definitions too; see the bibliographical notes for details. But the intuition behind all these definitions is to group similar points together in a single set.

Another type of clustering appears in classification systems in biology. (Such classification systems do not attempt topre`dictclasses; rather they attempt to cluster related items together.) For instance, leopards and humans are clustered under the class mammalia, while crocodiles and snakes are clustered under rep-tilia. Both mammalia and reptilia come under the common class chordata. The clustering of mammalia has further subclusters, such as carnivora and primates.

We thus havehierarchical clustering. Given characteristics of different species, biologists have created a complex hierarchical clustering scheme grouping related species together at different levels of the hierarchy.

Q:3 . The organization of parts, chapters, sections and subsections in a book is related to clustering. Explain why and to what form of clustering.

ANS:

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Q:4 . Suggest how predictive mining techniques can be used by a sports team, using your favorite sport as an example.

Example 1.

Table 1. Statistics for Teams A and B for the prediction of the winner based on (i) WPs, (ii) RPI, (iii) Elo rating, and (iv) Pythagorean wins.

Team	Wins	Losses	OR	OOR	Elo	PF	PA
A	5	1	0.556	0.611	1850	112	57
B	0	6	0.500	0.512	1200	41	153

(i) By considering only the winning percentages (WPs),

(ii) Team A is superior to Team B because the WPs of Teams A and B are 5/(5+1) ≈ 0.833 and 0/(0+6) = 0, respectively. Thus, Team A would be the predicted winner. (ii) When considering ratings percentage index (RPI), Team A is again superior to Team B because RPIs

of Teams A and B are (0.25*0.833) + (0.5*0.556) + (0.25*0.611) = 0.639 and (0.25*0) + (0.5*0.500) + (0.25*0.512) = 0.378, respectively.

(iii) (iii) Moreover, Team A with the Elo rating of 1850 is more superior to Team B with the Elo rating of 1200, i.e., a 650-point difference. (iv) Similar comments apply when using the Pythagorean wins. Although there is a large difference between the points Team A has scored (PF for A = 112) and the number of points that Team B has scored (PF for B = 41), but there is a small difference in the number of points either team has surrendered (PA for A is 57 and PA for B is 153). As Pythagorean wins for Teams A and B are1122.37/(1122.37+572.37) ≈ 0.832 and 412.37/(412.37+1532.37) ≈ 0.042 respectively,

(iv) Team A is gained the predicted winner (with higher Pythagorean wins than Team B). To summarize, for all four statistics, Team A is the predicted winner over Team B.

Q:5 . Suggest how predictive mining techniques can be used by a sports team, using your favorite sport as an example.

This algorithm calculates a reference count for each document identifier. A reference count of i for a document identifier d means that at least i of the keywords in S occur in the document identified by d. The algorithm maintains a list of records, each having two fields – a document identifier, and the reference count for this identifier. This list is maintained sorted on the document identifier field.

```
initialize the list L to the empty list;
for (each keyword c in S) do
begin
    D := the list of documents identifiers corresponding to c;
    for (each document identifier d in D) do
      if (a record R with document identifier as d is on list L) then
            R.reference_count := R.reference_count + 1;
      else begin
            make a new record R;
            R.document_id := d;
            R.reference_count := 1;
            add R to L;
      end;
end;
for (each record R in L) do
    if (R.reference_count >= k) then
      output R;
```