10/11/2015 – Complex Types of data

The generalization of spatial and multimedia data can be presented by spatial data, image data and music data. The difference between music data and image is the time series of the music. The generalising concept of object data is the object identifier, class composition hierarchies and the construction and mining of object cubes.

80% of the data is spatial data.

10/11/2015 (evening)

Other type of data is time-series, which is recorded at regular intervals, consists of sequence of values or events changing with time. Applications are for example, financial, biomedical and meteorological. The categories of time series are long-term or trend movement, cyclic movements and seasonal movements, Irregular or random movements. Methods for estimate trends are freehand method (interpolation); the least-square method and the moving-average method. The least-square method is less sensitive and more robust to deal with outliers than the moving-average. The discovery of trend in time-series is due by using seasonal variations, season index is a set of number showing the relative values of a variable during the month of the year. Deseasonalised data are adjusted for seasonal variations. A cycle can be discovered by the estimation of cycle variations and the periodicity; Estimation of irregular variations by adjusting data for trend, seasonal and cyclic variations; Prediction using systematic analysis of the trend. The similarity search in time-series analysis consists in find a slightly different data sequence in a query; Whole matching and subsequence matching operate two categories of similar queries to match patterns. Typical applications are financial and

market, scientific databases and medical diagnosis. As usual, data transformation are necessary for treating data before using, some applications in time series data of signal filtering are DFT, DWT in order to treat the noise and compare data. Subsequence pattern mining analyses frequency and patterns to extract information. Before applying the association rules there is the need of define transactions.

12/11/2015

Text databases are large collections of documents from different sources where traditional information retrieval techniques are inadequate for the increasingly vast amounts of text data. Some problems of DB such as concurrency control and recovery are not presented in information retrieval as some problems of IR such keywords search and the notion of relevance are not presented in DB. The information retrieval problem consists in locating relevant documents based on user input, such as keywords or example documents. Basic measures for text retrieval are precision and recall. Keyword-based retrieval uses expressions of keywords to search for documents that are correlated to the query search. The similarity-based retrieval is based on the nearness of the keywords or relative frequency of the keyword. Basic techniques are stop list, word stem, a term frequency table and similarity measures such as cosine distance. Latent semantic Index has the basic idea that similar documents have similar word frequencies. A term frequency table is built to register the occurrences of terms. Other text retrieval indexing techniques are inverted index and signature file. Types of text data mining are keyword-based association analysis, automatic document classification, similarity detection, link analysis and hypertext analysis. Keyword-based analysis seeks to gather sets of keywords or terms that occur in a frequent way

together and then find the association or correlation relationships among them. The process consists of Pre-process, association mining algorithms and term level association mining. Automatic document classification makes use of a training set generate by humans to be used in a classification and later application model. Document clustering seeks to automatically group related documents based on their contents, no training set is required and the taxonomy is generated at runtime using hierarchical clustering techniques.

17/11/2015

Mining the world-wide web

Data mining is a bigger context that englobes text mining. The web is huge repository of text data, there are many algorithms for text mining. Document clustering is a process based on a user goal and implies the pre-processing of data and clustering. The process of applying data mining in the web is the most extreme challenge of data mining due to the large amount of data and complex and heterogeneous structures. This last factor causes the impossibility of data integration and transformation for storage and a data warehouse. The information growth of the web is growing at exponential rate from the end of the 90's. 99% of the web information is useless to 99% of web users. To find high-quality web pages on a specified topic, one of the options of web search engines. Web search is divided into two steps the first one is crawl web pages; second one is mining the data. The crawling is domain restrained to provide better quality results and reduce the dimension of search. Index based are built based on keywords-based indices. Many documents that are highly relevant can be left behind in case they do not have the searched

keywords. Web context, web structures and web access patterns are the tree main categories of search data. Limited customisation, coverage and limited query interface are some of the problems of web searching engines. Regular expressions can be used to more specific and complete searches. Page rank is clever way to prioritize web pages. A web page that is accessed often it means it is important for determined keywords. The structure of web links are important to infer the notion of authority. If a web page has a large number of links, it shows the high value of a specific page. The hyperlink pointing to a web page is endorsement of that page. A hub is a set of web pages that provides collections on links to authorities. HITS is an algorithm that explore the interactions between hubs and authority pages.

19/11/2015

Automatic classification of web documents is made with assignment of class labels to each document from a set of predefined topic categories, based on a set of sample documents. Some method used to perform the automatic classification are keyword based classification methods and statistical models. Multi layered web information base divides the web on a hierarchical way. The layer 0 is the web itself; Layer 1 the web page descriptor layer; Layer 2 and up, various web directory services built on top of layer 1. XML can be used to web mining, its structure may be used to extract descriptors of elements. XML can help solve heterogeneity for vertical applications, but the freedom to define tags can make horizontal applications on the Web more heterogeneous. Some benefits of multi-layer meta web are Multi-dimensional Web info summary analysis;

Approximate and intelligent query answering; Web high-level query answering (WebSQL, WebML) and Web content and structure mining. The benefits of constructed such a meta-web may justify the cost of tools development, standardisation and restructuring. Some application are target potential customers for electronic commerce; Enhance the quality and delivery of Internet information services to the end user and improve web server system performance. Techniques for web usage mining are construct multidimensional view on the Weblog; Perform data mining on Weblog records and conduct studies to analyse system performance, improve system design by Web caching, Web page pre-fetching, and Web page swapping.

24/11/2015

Review of questions for exam.

1. Explain the main three steps of DM processing?

Data pre-processing.

Data analysis to extract patterns.

Knowledge discovery to summarize patterns - Evaluation.

2. Describe the tasks that need to be performed at each step of the process.

Pre-processing: Data cleaning, data transformation, data reduction.

Data analysis: Association rules, clustering, build the model.

3. What is spatio-temporal dataset?

Is a dataset where components of space and time are fundamental to represent the data. A data warehouse is formed based on a historic of data (temporal) and can be related to a determine space.

4. What is sequential pattern mining?

Sequential pattern mining try to discover frequency of patterns related to time or other sequences.