Qualitative Research and Computer Analysis: New Challenges and Opportunities

Allan H.K. Yuen

Database Approach and QDA Methods

Why or how QDA methods are different from the database approach as means of managing and exploring unstructured data

Implicit Properties of a Database

- A database represents some aspect of the real world, called the Universe of Discourse (UoD). Changes to the UoD are reflected in the database.
- A database is a logically coherent collection of data with some inherent meaning. A random assortment of data cannot correctly be referred to as a database.
- A database is designed, built, and populated with data for a specific purpose. It has an intended group of users and some preconceived applications in which these users are interested. (Elmasri & Navathe, 2000)

Characteristics of Database Approach

- Self-describing nature of a database system – system catalogue, metadata
- Insulation between programs and data, and data abstraction
- Support of Multiple views of data

- Sharing of data and multi-user transaction processing

Scientific and Statistical DBMS Application

- Successful application in science data analysis (Fayyad, 1997)
- The information in these databases consists of unstructured data, natural-language texts of documents; number-oriented databases primarily contain information such as statistics, tables, financial data, and raw scientific and technical data … etc
- Natural science fits in a mathematical or computational representation system
- The need for more powerful and flexible data models to support scientific database application, in particular social sciences

Data in QDA

- An 'assemblage' of data (Lee, 1993) – multi-stranded, derives from multiple sources, has multiple forms: transcripts, field notes, documents etc.
- The emergent character of QDA also tends to encourage 'data promiscuity' (Fielding & Lee, 1998) – data of different kinds are collected 'just in case'
- Unstructured, context-specific and recalcitrant – not imposing an a priori analytic structure on the data
- Tend to accumulate large volumes of data

Analyzing Qualitative Data

- Mysteries and tedious, always gives surprises
- In order to code something, the research has to understand what the respondent (or document) is saying
- Understanding and knowledge – complex cognitive process

- Process of making sense of the data
- Scaffolding of meanings

Issues of Incompatibility

Symbolic Computation

DBMS is not only a tool helping researchers in managing data, it tempts researchers to try to tie up knowledge as codified knowledge which stored in a database. This rigidity and formality regarding knowledge leads to the stultification of creativity (Allee, 1997) and limits the data analysis and exploration.

Data Representation

How you define data determines how you manage it. A database lends itself to a mathematical representation of the UoD. How a DBMS store and manipulate data affects not only matter of automation or efficiency, but it affects the process of making sense about the qualitative data. The representation of rich qualitative data in structured database systems is questionable.

Cognitive Process

Shannon and Weaver (1949) define information as data that "reduce uncertainty" – data that extend the realm of understanding, lead the mind to new awareness of the world. Once the data are acknowledged by the senses, the mind becomes involved and begins the cognitive driven segment (Debons et al., 1988). How can a database system support such the cognitive process of the researchers?

Goals and Values

The development of database technology has been dictated by the traditional symbolic-oriented computational theory. The focus is on well-defined problems. However, QDA methods contain

computational elements that cannot be described entirely in computational terms.

Data Mining and Web Search

What do data mining and web search tools offer to qualitative research?

Nature of Data Mining (DM)

1. DM refers to the mining or discovery of new information in terms of patterns or rules from vast amount of data. To date, it is not well-integrated with DBMS (Elmasri & Navathe, 2000)
2. DM is a methodology for discovering hidden patterns in data (Pawlak, 1999)
3. DM is a discovery-oriented process, designed to turn up relationships that you might not have suspected, as opposed to requiring that you specify a query (Carr, 1998)
4. DM as a part of the knowledge discovery process (knowledge discovery in database, KDD)
5. Goals of DM and KD: prediction, identification, classification, & optimization

Types of knowledge discovered during DM

1. Association rules (e.g. when a female retail shopper buys a handbag, she is likely to buy shoes)
2. Classification hierarchies (e.g. a population may be divided into five ranges of credit worthiness based on a history of previous credit transaction)
3. Sequential patterns: a sequence of actions or events is sought

4. Patterns within time series: similarities can be detected within positions of the time series
5. Categorization and segmentation: a given population of events or items can be partitioned or segmented into sets of "similar" elements

Process of DM

DM is a continuous, iterative process. A summary of the major stages of a data mining process is:

1. **Goal definition**: This involves defining the goal or objective for the data mining project. This should be a business goal or objective which normally relates to a business event such as arrears in mortgage repayment, customer attrition, energy consumption in a process, etc.
2. **Data selection**: This is the process of identifying the data needed for the DM project and the sources of this data.
3. **Data preparation**: This involves cleansing the data, joining/merging data sources and the derivation of new columns (fields) in the data through aggregation, calculations or text manipulation of existing data fields. The end result is normally a flat table ready for the application of the data mining itself (i.e. the discovery algorithms to generate patterns). Such a table is normally split into two data sets; one set for pattern discovery and one set for pattern verification.
4. **Data exploration**: This involves the exploration of the prepared data to get a better feel prior to pattern discovery and also to validate the results of the data preparation.
5. **Pattern Discovery**: This is the stage of applying the pattern discovery algorithm to generate patterns. The process of pattern discovery is most effective when applied as an exploration process assisted by the discovery algorithm.
6. **Pattern deployment**: This stage involves the application of the discovered patterns to solve the business goal of the DM

project. This can take many forms:

- Patterns presentation: The description of the patterns (or the graphical tree display) and their associated data statistics are included in a document or presentation. This requires the DM tool to generate text reports and WMF (Windows Meta File) representations of the graphical decision tree.
- Business intelligence: The discovered patterns are used as queries against a database to derive business intelligence reports.
- Data Scoring & Labelling: The discovered patterns are used to score and/or label each data record in the database with the propensity and the label of the pattern it belongs to.
- Decision Support Systems: The discovered patterns are used to make components of a decision support system.
- Alarm monitoring: The discovered patterns are used as 'norms' for a business process. Monitoring these patterns will enable deviations from normal conditions to be detected at the earliest possible time.

1. **Pattern Validity monitoring**: As a business process changes over time, the validity of patterns discovered from historic data will deteriorate. It is therefore important to detect these changes at the earliest possible time by monitoring patterns with new data. Significant changes to the patterns will point to the need to discover new patterns from more recent data.

Technology

1. It involves the use of software, sound methodology and human creativity to achieve new insight through the exploration of data to uncover patterns, relationships and dependencies.
2. Data mining solutions are based on the implementation, through programming, of interfaces to generally available and privately developed algorithms which enable the efficient exploration and organization of data.

3. Technique including: decision tree, case-based reasoning, statistics, neural nets, optimization algorithm … etc, see (Elmasri & Navathe, 2000)

Knowledge and skills required

1. A successful data mining project needs developers with the following knowledge and skills:
    - Deep knowledge of the data and its history.
    - Insight into the specific business area.
    - Proficiency in the use of the data mining tool.

1. The above skills may be combined in one person or may require more than one person. However, even in the largest of organizations there is a relatively small number of such specialists/teams with the above skills.

Principles of Web Search Engine (WSE)

1. Use statistics and probability to predict whether a document and a query are similar and estimate how similar (Feldman, 1998)
2. Each engine has its own algorithms for computing document relevancy
3. All the engines assume that the more times a term appears in a document, more likely that the document cover that subject
4. Terms which appear more frequently in a document than they do in the database as a whole indicate that the term in question is a major topic of that specific document
5. Relevant ranking works off weights assigned to each term. Higher rankings go to documents that have co-occurrence of terms.
6. Some kind of automatic stemming and truncation
7. They are not designed to work well on document records that

do not contain substantial text.

What is the state of WSEs today?

1. WSEs morphing into portals
2. Most WSEs contain elements of DM
3. Searching experience and searching behavior (Rappoport, 1999)
4. WSEs need to build in flexibility and options to serve the needs of a wide variety of searchers – "best" results for specific searches
5. Automatic clustering and categorization
6. Summarization attempts to reduce document text to its most relevant content based on the task and user requirements
7. Visualization of information
8. Searching multimedia, e.g. audio search using speech recognition www.compaq.com/speechbot (Jones, 2000)
9. Meta-search engines e.g. (www.metacrawler.com)
10. Intelligent agents
11. Cross-language information retrieval
12. Knowledge management
13. Data mining or text mining
14. The Search Standards Project – as a cooperative effort between the major services and industry observers to help establish standards that will benefit users, without hurting competition
15. W3C - the Resource Description Framework (RDF), as the proposed mechanism, is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. RDF encourages the view of "metadata being data" by using XML (the eXtensible Markup Language) as its encoding syntax.

Comparing Experience from DM and QR Project

Advanced Scout: Data Mining and Knowledge Discovery in NBA Data (Bhandara, et al., 1997)

- An IBM project - to discover interesting patterns in game data
- Data collection (who took a shot, type of shot, outcome, …….)
- DM using Attribute Focusing (AF) – occurrence and probability
- Define an "even" (E) as a string of attributes and "interestingness" measure using I(E)
- Result of DM – "when X was point-guard, Y missed 0%(0) of his jump field-goal-attempts and made 100%(4) of his jump field-goal-attempts"
- Provide interactive analysis to gain additional contextual information
- However, the best opportunity to interpret a pattern is via video tape of the actual events

Studying ICT-Supported Pedagogies in Hong Kong (Law, Yuen, Ki, Li, Lee, & Chow, 2000)

- Multiple case study
- Data collection (video, interviews, & documents)
- Coding scheme (challenge, instruct, set, evaluate, ……)
- Some patterns of instruction are quite unique, it is hard to generalize in certain mathematical representation
- The development of a "video analyzer" for data display

Discussion

1. Commercial DM or KDD is in its infancy (Kim, 1999)
2. Both DM and QDA aim at discovering hidden patterns in data. DM emphasizes the automated application of algorithms to detect patterns in data (Bhandari et al., 1997) from vast amount of data, whereas QDA is a kind of semi-automated processing and the data set sometimes is unique

3. Goals and outcomes are different. DM is domain-specific (e.g. "Advanced Scout" is data mining and knowledge discovery in NBA data) and the outcomes are rules or formal patterns to facilitate decision making.
4. QDA projects take place in a "rich" and "ill-structured" domain. An ill-structured problem lacks a fixed solution algorithm and maybe even clear goal achievement criteria. Most ill-structured problems of interest to QDA are highly context sensitive. DM data are more structured or sometimes semi-structured and less context sensitive.
5. Engineering-driven development, User research
6. DM tools are continually evolving with latest algorithms and including non-standard data
7. Mathematical tools that fit QDA – fuzzy set, rough set
8. Visualization and multimedia data in QDA, agent technology
9. Marriage of DM and Web searching - information on the web, data collection from the web
10. Ranking web document and "core" category
11. Metadata and standards – "metadata being data", building professional community for qualitative researcher (to join the rich text revolution)

Web-based Research and QDA

What should QDA software do to support web-based research? What methodological and technical problems are posed for web-based QDA by the web itself?

Application of Web-based Research – Marketing Research

1. Web-based surveys to be a cost-effective method for collecting data (Pealer, 2000).

2. Marketing researchers increasingly are using the WWW as research tool. Despite the deficiencies of online research, Web-based research is helping advertisers plan Web ad campaigns. Online searchers must be cognizant of the fact that while market research surveys conducted on the Web can provide valuable data, results may be skewed due to the sample population (Maddox, 1997).
3. A growing number of market studies rely on Web-based marketing research to gain critical data upon which some of the most fundamental business decisions are based. The increasing use of Web-based marketing research can be attributed to several factors. These include the elimination of the need for an interviewer which, in turn, eliminates interviewer bias and provides the researcher with greater control over data quality. Web-based surveying is also cheaper and generates information faster than traditional methods (McCullough, 1998a).
4. Customer feedback will be more readily available when important business decisions need to be made. To ensure the success of a Web-based survey, researchers should state clearly the questions they want answered, identify the population segment they want to reach, formulate a questionnaire and post it on the Web, boost traffic to the questionnaire, and conduct analysis of data collected (McCullough, 1998b).

Application of Web-based Research – Tax Research

1. With Web-based tax research solidly established, Internet-enabled tax preparation is just now getting off the ground (McCausland, 1999).
2. Raising the stakes in the Internet tax publishing wars, CCH Inc. has launched a Web-based, browser-driven tax library for accounting professionals. CCH said its Internet Tax Research Network offers a full range of federal, state, estate, gift,

business entity, payroll and pension tax information. Users can also conduct complex tax research by searching topics, such as corporate or sales taxes, across any or all states (http://tax.cch.com). (Faulkner & Gray, Inc., 1997).

3. The Internet will make it easier to get to the most up-to-date tax information, and may even make it more affordable for the user who can't justify big buck subscription costs. But the largest benefit, and the one that will probably have the most impact on the profession, is that by putting these products on the Web, vendors will be making them not only more accessible, but easier to use (Needleman, 1997).

Web-based Research Tools

1. "WebZinger" (www.webzinger.com) is designed to assist Web-based research projects by providing a simple query interface and a comprehensive result-reporting tool. (Kirkpatrick, 2000)
2. "Web Buddy" offers an extensive feature set with page, branch, and Web-site grabbing capabilities in real time and on a scheduled basis. Other Web-friendly features include the ability to convert HTML to Word documents and to import Netscape bookmarks (Ozer, 1998).
3. "HotPage" indexing and quick-reference tool for Web pages offers some features when used as an offline browser or as a bookmark manager, it may be suitable for those who wish to include pictures of Web sites in printed materials. (Bailes, 1996).
4. The biomedical engineering community recognizes the need to organize physiological data in a comprehensive, user-friendly, and accessible format, for example for Escherischia coli (http://ecocyc.pangeasystems.com/ecocyc/) at the University of Virginia (Ley & Brewer, 1999)
5. "Global Access" is a Web-based research tool, providing integrated access to company information from multiple

sources (IEE, 1999)

6. Taking an iterative design approach, a user-centered field study investigated requirements to support remote scholarship. By observing scholars' use of a library special collection, a requirements analysis study established a model of scholarly inquiry and specified the initial design requirements for a World Wide Web based research library. Upon implementing a prototype providing intelligent information retrieval capabilities to access the full text and page image representations of 28 books from the special collection, its use was evaluated by observing scholars conducting research remotely. Supporting the evaluation study, a Web-based field laboratory administered survey instruments and instruction to participants. Promotional activities attracted an international group of scholars to use the prototype and participate in its evaluation. The results provide insight into Web-based inquiry and requirements for remotely supporting scholarship (Downs & Friedman, 1999).

Standards and Security of Web-based Research

1. LURHQ set up a secure intranet that the firm's 300 employees could access from anywhere and the Internet while protecting confidential internal data, which used security products from Network Associates and RSA Data Security (Rogers, 1999)
2. The CHIC-Pilot project is a TERENA coordinated effort investigating the feasibility of setting up a large-scale. It brings together existing distributed indexing service for searching Web-based research information search services throughout Europe, integrating them into a coherent architecture. (Valkenburg, et al., 1998).

Summary of Issues

1. QDA query and database query – query in NUD*IST (www.qsr.com.au)
2. Reporting and visual representation – visualization contributes to the conceptualization of qualitative data in QDA, ways of visual representation in QDA, such as concept map, event-state networks, causal network, chronological event chart, flow diagram, tables … etc. can be used in reporting or data display (www.scolari.co.uk)
3. Numerical data to open-ended questionnaire – data collection and analysis, SPSS TextSmart 1.0 (www.spss.com/textsmart/overview.htm)
4. Information retrieval techniques
5. Searching, data mining and web mining

Multimedia / Hypermedia Data

1. Automated video analysis (Evans, 2000) - data modeling for dealing with complex objects
2. Documents constitute a specialized area and deserve special consideration (Elmasri & Navathe, 2000)
3. Storage of multimedia database presents problems of representation, compression, mapping to device hierarchies, archiving, and buffering during I/O operation. DBMS will be required to deal with synchronization and compression/decompression, and will be coupled with indexing problems, which are still in the research domain.
4. Multimedia database opens up issues in queries and retrieval, such as efficient query formulation, query execution, and optimization. (www.qbic.almaden.ibm.com)
5. One of the reasons for low precision in text information retrieval systems is that words have multiple meanings. One possible way to resolve ambiguity is to use an on-line dictionary; another is to compare the context in which the two words occur.

Final Remarks

1. ICT implementation - automation, rationalization, re-engineering, paradigm shift
2. ICT transforms the process of qualitative data analysis, but is not the only agent to the transformation
3. The role of data in qualitative methodology