Building Knowledge around Complex Objects using Inforbright Data Warehousing Technology

Julia Ann Johnson
Department of Mathematics and Computer Science
Laurentian University, Ontario, Canada jjohnson@cs.laurentian.ca

Genevieve Marie Johnson
Centre for Psychology, *Athabasca University*
Alberta, Canada gjohnson@athabascau.ca

Abstract

There are considerable challenges in analysing and reporting on word-based data. Infobright data warehousing technology was used to build knowledge around qualitative data that are subject to human interpretation. Infobright was chosen as a system for implementing the data set because its rough set based intelligence appears to be extensible with moderate effort to implement the data warehousing requirements for automatic interpretation of word based data. An example of social sciences research data was used for illustration.

Keywords: *rough set, data warehousing, Infobright, semantic similarity, qualitative data, qualitative data analysis, knowledge grid, WordNet.*

1. Introduction

Infobright (IB) is a database server that uses a rough set based data compression method to efficiently process queries on very large databases. The use of information that is already provided by IB or can be easily provided with minor modification is advanced for analysis of qualitative or word-based data emerging from research in the social sciences.

IB is reviewed in Section 2. A representational word-based data set is reviewed in Section 3. Implementation of that data set by building knowledge around the complex objects abstracted from it is discussed in Section 4. Methodology for implementing the proposed extensions while streamlining with existing IB methodology is presented in Section 5. Conclusions are provided in Section 6.

2. Infobright

A column-oriented database management system stores content by column rather than by row. A column-oriented approach has advantages for data warehouses and library catalogues where aggregates are computed over large numbers of similar items [1][2]

IB [3][4][5] uses a column-oriented database architecture. Column-oriented, as opposed to row-oriented, databases lend themselves to data compression techniques. Since all values in a column are of the same type, the values of each column may be split into separately

compressed value chunks. Information about the column type and the patterns occurring within the value chunks characterize the column. The amount of information stored to describe the value chunks is smaller than that required to represent the row chunks of comparable size.

IB uses what is referred to as the database knowledge grid that equates to metadata of conventional databases but organized with knowledge nodes of compact information about the value chunks. The historical counterpart in conventional databases is data blocks corresponding to large portions of the database. Much of query processing can be accomplished using the IB knowledge grid without need or with significantly limited need to fetch compressed data.

3. Social Sciences Research Data: Qualitative and Quantitative

The social sciences comprise academic disciplines concerned with the study of the social life of human groups and individuals including anthropology, communication studies, cultural studies, demography, economics, education, political science, psychology, social work, and sociology. Inevitably, research methods in the social sciences include collecting data on human subjects. Research tools for collecting social sciences data include direct observation in various contexts, questionnaires and surveys completed by individuals, and the administration of standardized instruments such as IQ tests [6]. Two types of data emerge from such collection strategies, -- qualitative and quantitative. Qualitative data are expressed in words, texts, narratives, pictures, and/or observations; quantitative data reflect numerical representation of phenomena such as performance test scores, physiological measurements, and numerical ratings [7] as illustrated in Figure 1.

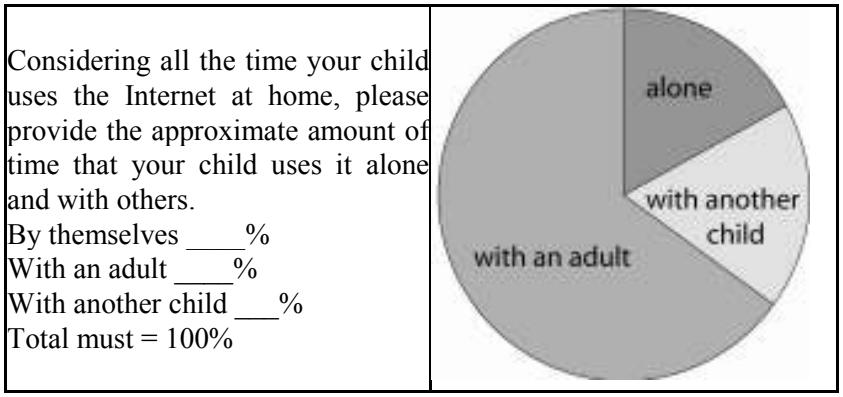

Figure 1. Quantitative data from social sciences research

Increasingly, the use of both qualitative and quantitative research methods is promoted as necessary to valid investigation of complex human behaviour. The theoretical assumption underlying a mixed methods approach is that "the world can be represented through both numbers and words and that numbers and words should be given equal status" in data collection and analysis in the social sciences [8].

Quantitative and qualitative data necessarily require differing methods of analysis. Quantitative number-based data are analyzed with statistical software, most commonly SPSS (Statistical Package for the Social Sciences). SPSS was first released in 1968 [9] and continues to be a particularly popular software in social sciences quantitative or statistical data analysis

[10]. Qualitative word-based data, in contrast, are coded or classified by researchers in order to organize a large amount of words into manageable chunks or meaningful attributes. Such subjective or human treatment of data introduces bias and error as different individuals interpret text differently [11]. An example of social sciences research data may clarify some of the challenges in qualitative or word-based social sciences data analysis and interpretation.

3.1. Word-based social sciences research data: A survey of those who self-injure

Non-suicidal self-injury (SI) is defined as direct, deliberate destruction of one's own body tissue without suicidal intent [12]. Polk and Liss [13] found that 20% of college students reported having self-injured at some point in their lives. Many self-injurers find support in virtual communities that typically include a website with e-message boards. Important information for psychologists and e-health practitioners includes description of individuals who participate in virtual communities for those who SI (e.g., the nature of self-harm, reasons for participating in virtual communities, and perception of the effect of such on level of SI).

Sixty-seven individuals who participate in virtual communities responded to 12 questions posted on two e-message boards for those who SI. Respondents ranged in age from 16 to 60 years (mean age = 26.5 years; SD = 9.93 years). Almost 15% (n = 10) of the sample indicated that they were male, one respondent indicated transgender, and 83.6% noted that they were female. Participants ranged in duration of self-injurious behavior from one month to 35 years; average duration was 9 years and 5 months (SD = 7.8 years). Website administrators, themselves recovering from SI, were suspicious of researchers and would not allow questions to be posted unless approved. Approved questions, ultimately, were all open-ended and thus only answerable with text or written words. For purposes of illustration, Table 1 provides some of the posted questions and the responses of one participant. Note that in qualitative data, grammar, spelling, and punctuation are not corrected.

Table 1. Sample posted questions and responses (abbreviated)

Posted Questions	Sample Response
Why and how do you self-injure?	… I discovered it after a bad day when I was 5 and had broken a glass and accidently cut myself with the glass…
How long have you self-injured?	I started around 5 or 6 and I am 34 now.
Have you ever tried, or are you now trying, to stop self-injuring?	I have never really made a major effort to quit but there have been periods where I just stopped for periods …
If you are attempting to stop self-injuring … what methods …	One of the methods that I used to use was to drink which nearly always backfired. Now days I try to write …
Do you have an eating disorder, or abuse drugs or alcohol?	I do not have an eating disorder but as a form of control when I try to stop cutting I have had disordered eating …
Do you feel you have control …	For the most part yes
When start using these boards?	I joined … a little over 3 years ago.
How often do you visit this message board?	I have been visiting as a regular member …

Why do you visit this website?	it gives me an emotional outlet, allows me the chance to say what I will in a nonjudgemental manner …

As can be seen, responses to the posted questions contain many words and a few numbers. Relative to words, numbers are easy to organize and interpret. For example and as illustrated in Figure 2, responds to the question of frequency of visiting online communities for those who SI was organized and presented with precision. The majority of posted questions, however, elicited word-based data that may be organized and interpreted differently across researchers. As contrasted in Tables 2 and 3, a seemingly straightforward question of method of SI resulted in conflicting word-based data organization and corresponding interpretation. Two equivalently trained research assistants independently coded the word-based responses to the posted questions. In both cases, coding took many hours of focused effort. Divergent interpretation of data was apparent between the two coders and, in both cases, description of methods of SI were limited, despite considerable effort and expenditure of personnel resources. In general, the greater the number of words in a response, the greater the human time required to organize or code the data and the greater the influence of subjective human interpretation. There would be little need for human coding if responses to questions could be evaluated based on semantic comparison of the text fields of different respondents to the same question.

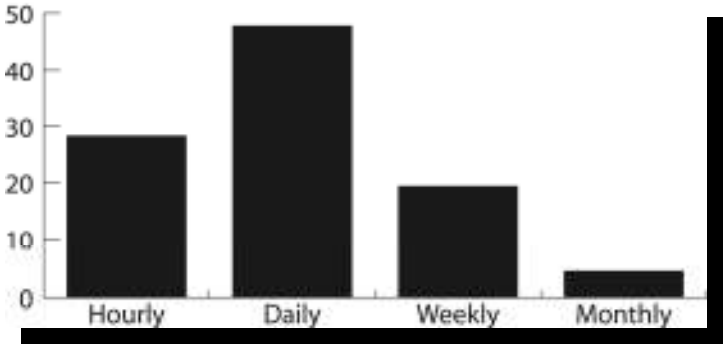

Figure 2. Coding simple word-based data: How often do you visit this site?

Table 2. Coder A: How do you self-injure?

Method	Sample Response Phrase	% Indicating
cut	cut words into my skin	92.5
burn	burn using an iron	38.8
hit/whip	hit myself to bruise	20.9
scratch	scratching (fingernails, paperclip)	13.4
overdose	painkillers, diet pills, and laxatives	6.0
other	over exercise	6.0
bite	bitten myself	3.0
strangulate	hit, choke, scratch, and cut myself	3.0

break bones	many ways including bone breaking	3.0
disordered eating	starvation, bingeing, purging	3.0

Table 3. Coder B: How do you self-injure?

Method	Exclusively	Non-exclusively	Rarely
cut	41.5%	53.8%	1.5%
burn	3.1%	33.8%	4.6%
hit	1.5%	6.2%	4.6%
scratch	1.5%	12.3%	1.5%
choke		1.5%	1.5%
bang head		4.6%	
bite		4.6%	
drug use		6.2%	1.5%
exercise		1.5%	
pick scabs		4.6%	1.5%
cause bruising		9.2%	1.5%
break bones		6.2%	1.5%
salt in wounds		1.5%	

3.2. Software for analysis of social sciences word-based research data

To some extent, the coding of qualitative social sciences data has been improved with the development of software. Weitzman and Miles [14] applied code and retrieve functionality to qualitative text data. Current code-based software includes content analysis tools, word frequencies, word indexing with key word in context retrieval, and text based searching tools [15]. Such software, however, has not been readily adopted and the fundamental assumption that machines can make meaning of word-based data has been questioned [16] in the social sciences. Computer scientists are generally more favorable toward the idea of machines extracting knowledge from data, but all would agree that the existing software for analyzing qualitative data is inadequate for making meaning of text. The available software amounts to data management systems requiring the user to reformulate the data by putting his/her own interpretation into it in a preprocessing step before it can be input into the system. To avoid the subjectivity so introduced, a view based on building knowledge around complex objects was investigated.

3.3. Previous semantic similarity comparison metrics

Various algorithms for semantic similarity have been developed by researchers. Chien and Immorlica [17] investigate the idea of discovering semantically similar queries of search engines. Their technique rests on the similarity in behaviour of the queries over time. They developed a method of finding temporally correlated input queries to serve as a means of

quantifying the relatedness between queries. Their work may be relevant for relatedness of questions, but logically related qualitative responses may not be temporally correlated as are the requests of a web server. However, this work may be applicable to social sciences data for measuring semantic similarity of responses collected on an ongoing basis.

In the investigation at hand, the focus has been on analysis of responses that have been collected from a web site posted for a short period of time. It was assumed that each question on the posted questionnaire could be formulated as a precise query. Relaxation of this assumption for future work is expected because the same techniques used for analyzing narrative responses may be applied to narrative questions.

Algorithms for the various aspects of measuring similarity are available that rely on use of the WordNet lexical database [18]. WordNet gives specific meanings of words and establishes connections between parts of speech. A freely downloadable interface is available that accepts words and gives a measure of their similarity. Various WordNet based semantic relatedness algorithms are available online, for example, one as a Perl module and another as Java pseudo code.

WordNet has been used to detect spelling errors that go unnoticed by a regular spelling checker [19]. A difficulty to be overcome with the WordNet route is that grammatical correctness and spelling accuracy to any small extent cannot be assumed for social sciences qualitative data. For example, respondents, especially in the internet, tend to use phrases such as “good4U”. Therefore, the matter of checking for errors and grammatical correctness using WordNet needs to be revisited when dealing with qualitative data.

The capability of Infobright (IB) to store and efficiently retrieve large amounts of data makes it possible to store massive amounts of WordNet information together with application data organized and integrated within the relational data schema. In this way, we build knowledge around complex objects (sentences and collections of sentences) based on knowledge about atomic objects (words) and links among some of them. Such an approach has an analogue in the rough mereology (RM) [20] that ties together the rough and fuzzy paradigms for modelling vague, inconsistent, imprecise, and incomplete information. RM is based on a compositional view of object construction consistent with the proposed compositional semantics approach for organizing knowledge about complex objects.

3.4. Why Infobright?

Why use IB in favor of other revolutionary high performance data warehousers that employ Field Programmable Gate Array (FPGA) technology? For example, Kickfire [21][22] uses some of the same principles as IB (columns, compression, special indexes, MySQL interface), but implements them using FPGAs which offer a platform for the implementation of processing engines. The Kickfire strategy is to “turn software into silicon” by building an appliance that contains a MySQL chip. The appliance plugs into a standard Linux host server where the MySQL database resides and where its storage engine and the Kickfire utilities are also run. MySQL appliances are marketed under different names (eg., Netezza TwinFin [23], DB^X from XTREMEDATA [24]) and they provide a high degree of interoperability.

FPGA devises are re-programmable in-house [26] speeding up SQL operations concerned with large data movement and time-consuming functions, such as joins, sorts, groupby, orderby and aggregations. IB designers have also emphasized fast joins but implement them in software. Of seeming relevance to the study at hand is the FPGA coprocessor board [25] used to accelerate the processing of queries on a relational database that contains texts and images.

However, we find that text databases are quite distinct from qualitative data sets.

Returning to Kickfire for further illustration, given the power of the MySQL Chip, the host server is modest in capacity with just two CPUs and 16GB of memory. Contrast this with the IB server that for our relatively simple educational application has two processors and 8GB, one GB for each concurrent user. So why use IB? The answer is that FPGAs provide an implementation strategy. The IB software itself could be implemented in hardware using an FPGA. It is not inconceivable that, like the MySQL software has been implemented as a MySQL chip, the IB software may ultimately be implemented on a chip. Now is the time to strategically align ones research directions with paradigms positioned to take advantage of the next generation analytic appliances [27] that resolve the von Neumann bottleneck that limits data transfer rate between the CPU and main memory.

4. Infobright implementation

IB uses the concept of a rough set to determine, for a given query, the stored data packs that are irrelevant (disjoint with the answer set), relevant (fully inside the set) and suspect (overlapping with the answer). Only the suspect data packs need to be decompressed because for those it is necessary to determine exactly what parts of their data satisfy the query at hand.

The use of such data compression techniques is needed for efficiently storing and retrieving information in large stores of word-based qualitative data. Compression is even more important for ongoing collection of qualitative data. But, information generated as a result of decompression may also find use in evaluation of the meaning of responses to queries.

IB returns statistics about the time required to evaluate a query. There is likely a correlation between the amount of time required to evaluate a query and the amount of exact computation required for the query. Exact computation occurs when data packs must be decompressed. A way to test the hypothesis that information generated as a result of decompression has semantic utility is to see if there is a correlation between query speed (of the queries that capture the meanings of the responses) and semantic relatedness of those responses. Extensions to IB for better analysis of such data are proposed in Section 5.

4.1. Database design

In qualitative data, column values that tell a story are of interest. Column names are likewise one or more short sentences, usually questions. In the selected design each column name corresponds with a question, and there is exactly one record for each subject. A column extension corresponds to the answers by all subjects to the question given by that column.

Experience was gained on suitable column names when Excel tables were used to record coded data. The tables were designed with abbreviated column names that expanded to the full text of the question. In Microsoft Excel, columns themselves may be stretched having the effect of revealing previously hidden parts of long column names. In similar manner when using IB, it would be helpful to be able to point to a column entry and have it expand to the full text comprising the column value. Otherwise, a meaningful tabular display of qualitative data to fit conventional display media may not be possible.

4.2. Database definition and loading

Infobright Enterprise Edition was obtained on the basis of a special academic promotional offer. It was installed on an 8-core, 8 gig ram server running the Debian operating system (a

flavor of Linux). IB's MySQL pluggable storage engine architecture allowed the database server to be accessed using an SQL client running on Windows.

As illustrated in Figure 3, a database schema was defined and implemented on the IB server with 12 columns (questions). The database was populated with 67 rows corresponding to answers given by as many subjects. Abbreviation of column names was required to fit the MySQL constraints on names. Take for example Q6 that reads "If you are now attempting to stop self-injuring, or have previously tried to do so, what methods do you use when you feel the urge to self-injure? Where did you learn these methods?" The text of such questions is too long for a name.

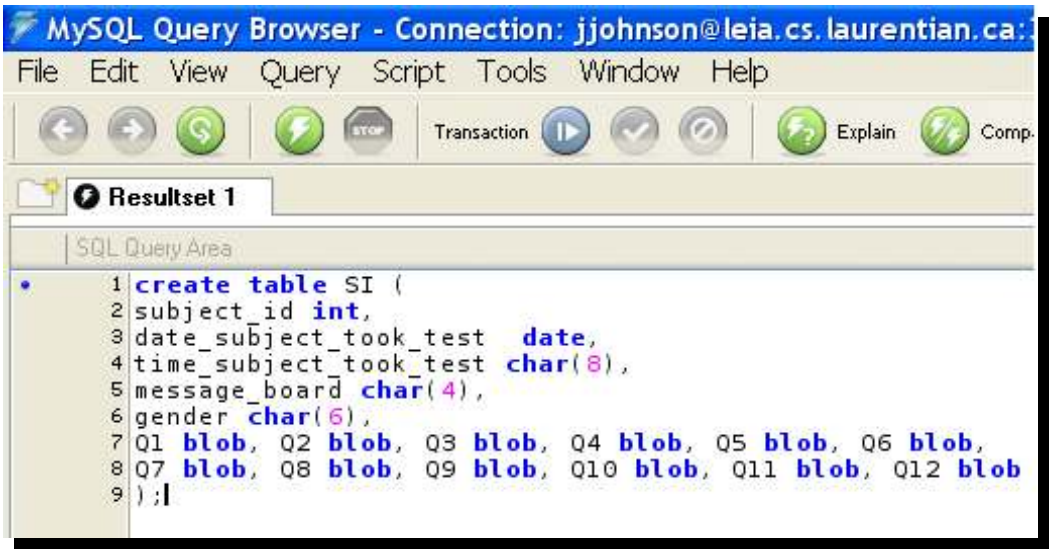

Figure 3. Schema definition using MySQL connected to Infobright server

An immense downsizing of the original MySQL database capabilities is revealed in the How-to-Work-with-Data-Types blog [36] on the IB site. ENUM, UNSIGNED INTEGER, DECIMAL with a precision larger than 18, and BLOB are not supported even though table creations that use them may successfully compile. Quoting from the How-To Blog:

> A BLOB is a binary large object. Its size is limited only by memory and disk limitations. In place, it is recommend to use the VARBINARY data type. The largest VARBINARY that can be stored in Infobright is theoretically 64K in size, but due to maximum row length limitations in MySQL, this can be reduced significantly depending on the size and number of other columns. To store an object that is a very large size, for example, a PDF document or JPEG picture, it is recommended to break the object down to manageable size for example, VARBINARY (8196) and store the object in multiple rows. The size of VARBINARY is also limited by available memory. During a load the compression process would take a huge amount of memory if, for example, VARBINARY (64000) was used, as memory has to contain 64K rows at once to do the compression.

The command in its simplest form for loading data into a table follows**:**
LOAD DATA local INFILE 'c:/tilde_data.txt'
INTO TABLE SI
FIELDS
TERMINATED BY '~';

c:/tilde_data.txt is the name of the file on the local PC on which the data were located. The tilde **'~'** as column entry delimiter was arbitrarily selected. Options were available for handling delimiters that also appear within column entries. If tildes were expected to occur in responses given by subjects, those options would have been employed.

The data required preprocessing to permit use of the IB LOAD command that was specifically designed for large quantities of data. The MySQL INSERT statement was also available but would require a script to be written for inserting the rows. The preliminary field values and answers given to the first few questions for subject 001 in preprocessed form follow:

~001~ Nov. 13~ ~9:37pm~ ~bus ~
~ no answer From question 4, the answer can be supposed to be '34' ~
~ Female ~
~ I started cutting completely by accident. I discovered it after a bad day when I was 5 and had broken a glass and accidently cut myself with the glass. I noticed that it helped release the intense emotions and made me feel more able to breathe. There were times when I would burn but that was an entirely different sensation than cutting and was rare.~
~ I around 5 or 6 and I am 34 now.~

Complementary to the CREATE TABLE command illustrated in Figure 2, the corresponding schema is presented in Figure 4. Although subject_id uniquely identified subjects, we had no facility for specifying a key constraint in IB.

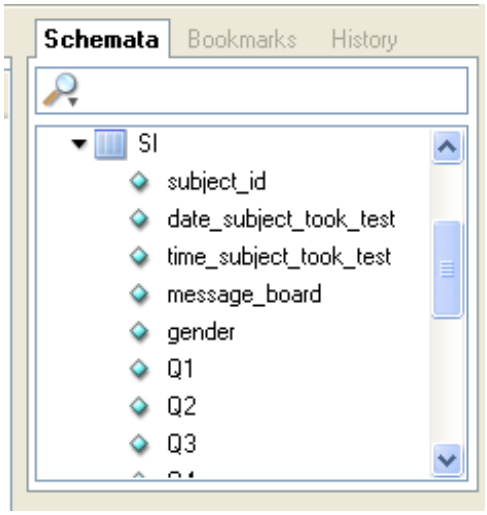

Figure 4. Fields of self injury table to collect preliminary information and answers to questions

4.3. Database query

Specific requests on the data were formulated by guessing about likely relationships, for example:

1. Do older individuals use different methods of SI than younger individuals?
2. What is different between individuals who say their SI has increased since visiting the boards and those who say their SI has decreased?
3. What differentiates those who SI with eating disorders (or substance abuse issues) and those who SI without associated disorders?
4. Do males SI for the same reasons as females SI?

Let *reason* denote the "why" part of Q1 (Why and how do you self-injure?) A MySQL query to partially answer request 4 looks like the following:

```
SELECT M.subject, F.subject FROM SI M, SI F
WHERE M.reason = F.reason AND M.gender ='M' and F.gender ='F';
```

As of now, there is no native support for weighing the semantic equivalence of two sentences either through proprietary measures or by any SQL specific extensions (Quoted from the IB Help desk). Therefore, the above query returns an empty (null) answer even if the semantic content of the two operands is identical (M.reason = F.reason). There are opportunities to leverage IB, Inc. to help with the problem by sharing in the IB Forums.

5. Streamlining the application with existing Infobright methodology

Several possibilities for introducing semantic similarity were studied: 1) expansion of the meaning of relational operators in queries 2) native support in IB for weighing the semantic equivalence of two sentences, 3) augmentation of the database in a preprocessing step to include columns that contain coded information that would permit a KDD approach to computing semantic similarity.

Option 1 is a matter best taken up with the MySQL community rather than the IB designers. For example, a suggestion regarding the HAVING and GROUP BY clauses in queries follows: Introduction of a column oriented feature extraction and classification process that clusters texts within a given column by placing those requiring a high degree of exact computation (semantically similar ones) all in one cluster. Options 2 and 3 are discussed individually in the remainder of this section.

5.1. Native support for semantic equivalence

Regarding option 2 above, proposed extensions to IB follow: 1.) The provision of a means of associating with the results of relational expressions (>, =, IN, etc.) derived from narrative responses, an indication of whether the relationship is irrelevant, suspect or relevant. In other words, to check relational expressions, find out how much decompression was required. 2.) In the case of an exact computation (suspect data packs), a measure of the amount of overlap. Such a measure differs from early measures of certainty and coverage associated with predictive rules generated by rough set based inductive learning algorithms. There the metrics were based on rows while here they are based on columns 3.) Expansion of the statistics about query speed to provide a more accurate indicator of the degree of overlap.

Additionally, we would like to see a facility provided whereby an IB database may be overlaid by a MySQL schema for specification and enforcement of key constraints and referential integrity constraints. This would make IB more applicable for managing conventional databases such as existing lexicons required for describing complex objects.

Implementation of the proposed extensions is expected to take advantage of the organization of an Infobright database as a rough level and an exact level. The rough level known as the knowledge grid with its "small, efficient, calculable units of information about data packs" reduces the need for decompression [28]. It is anticipated that the knowledge grid could also be used to provide semantic information about whether a given text and one selected from a column are identical in meaning, not similar in meaning at all, or overlapping in meaning. An indicator of the degree of overlap is already partially provided by the statistics returned from IB upon each query evaluation. The relative amount of exact computation required to interpret a narrative response can be inferred from the query speed.

5.2. KDD approach to computing semantic similarity

A strategy based on possibility 3) was discussed in the IB forums. The idea proposed was to handle semantic similarity at the data preprocessing level, comparable to the ETL (Extract Transform Load) process required to convert an existing database to an IB one. Similarity was to be expressed by means of additional columns in the data table created prior to loading the data. The idea was that, given appropriate additional columns, the semantic similarity should be at least roughly expressible by means of SQL conditions. Algorithms for semantic similarity may then be at least partially adapted to a MySQL query interface. Population of such appropriate columns would continue to emerge from subjective inputs.

Such an approach based in part on information coding is fraught with the same problems that were outlined in Section 3.1 concerning subjective hand coding of the meaning of qualitative data. Two coders will invariable provide two different interpretations of the same data. The additional columns would introduce subjectivity and considerable expenditure of personnel resources. However, some of the problems inherent in human coding may be overcome by adding to the relational data schema, columns for automatic recording of information about words, about how they are composed into sentences, and about how sentences are organized to form a story. Each column of the augmented relational schema corresponds to a single data pack. Joins and self-joins can be done efficiently on large data tables and these will be the primary operations required to query the qualitative data.

5.3. Future work

Compositional semantics applied to short stories provided by human subjects would see the interpretation of sentences and collections of sentences as deriving from knowledge about atomic objects (words) and links among them. We foresee that huge amounts of metadata will be required to automatically interpret even a small size set of qualitative data. There are 25,394 words appearing in the responses to questions across all questions and across all subjects of the SI data set. It is possible to deal with much larger data sets that this within IB even if these words were to be augmented with huge amounts of metadata.

A re-direction of this research has been necessitated by IB's inability to handle large data objects. Rather its focus is on handling large amounts of data. This apparent deficiency of IB has turned out to point us to a more appropriate direction, one distinct from text analysis systems [29]. SQL queries will be formulated on the qualitative in conjunction with the

metadata needed to understand the qualitative data. An approach to be investigated is that of automatically populating the columns of the database schema by WordNet [18] information about words, their synonyms and the relationships between parts of speech.

Automatic interpretation of qualitative data requires implementation of a natural language back end capable of translating stories to precise meanings. Previous research in the area of natural language (English) interfaces has provided researchers with an understanding of the issues involved in translating a natural language request to an SQL query with corresponding meaning. However, we wish to use non-conventional methods unconstrained by matters of data storage and retrieval efficiency. By storing WordNet information within the relational schema, it may be possible to dispense with some of the insufficient algorithms for converting English requests to a syntax tree and that syntax tree to a formal expression with the aid of additional world and contextual knowledge.

Instead, syntactic and semantic information (output from WordNet) as well as world knowledge and contextual information can be organized using a database schema capable of expressing the association of such elements with knowledge discovered from the data or specified a priori by means of database structure. The previous approach has been to have a general lexicon and a domain dependent one, but perhaps both could be entered as part of the data preparation phase. Most data warehousing systems have sufficient capability to load existent databases in various formats. An outline of methodology for future research follows:

Step 1: Available semantic relatedness algorithms may be brought to bear on the problem of automatically interpreting qualitative data. The relationships and knowledge about words available from WordNet [18] are likely insufficient if used by themselves but augmentation with other paradigms may improve the outcome.

A paradigm under consideration is that of pair wise comparison (PC) [30][31][32] and locally available PC based Concluder system for its ability to find semantic similarities among words given known similarities between pairs of words. It may be possible to provide objective pair wise similarities between words using WordNet to achieve multi-way similarities from PC/Composer that would be useful for groupings of semantically similar words and phrases. Initially, the primitive pair wise similarities will be based on subjective evaluations but must be corroborated with the relationships between pairs of words provided by more objective sources. To that end, we are also exploring algorithms for measuring similarity that are based on knowledge sources outside the WordNet lexical database.

Step 2: We are exploring the Rough Mereology (RM) paradigm as a means for enhancing the quality of relatedness measures developed in step 1. Validation will be achieved by cross-checking of degree of relatedness measures provided by the rough mereological inclusion function with those provided by both subjective and objective functions input to PC/WordNet approach. Also in this step, we will combine the RM and WordNet paradigms with the expectation of measuring degree and plausibility of the semantic relatedness outcomes among groups of words.

Step 3: Introduction to Infobright Enterprise Edition (IEE): The IB Data Warehousing engine is a column-oriented analytic data warehouse built on open-source MySQL data management software. IB was designed specifically for large volume data warehousing applications with up to 30TB of data providing sufficient capability for the maximum 8GB database developed in the previous weeks. The text part of the social sciences data set has been implemented in a conventional manner and the objective from here is to design and implement a meta database capable of expressing information about words and phrases in the narrative data.

Step 4: Viewing IB Compression Ratio Statistics: IB server provides a highly-compressed database system optimized for analytic-type queries. Specific statistics on table and column compression are provided. It remains to investigate the use of Compression Ratio Statistics not only for achieving goals of physical storage, but also, for providing information at the database conceptual level.
Step 5: Application of the IB Knowledge Grid: It is anticipated that the knowledge grid could also be used to provide semantic information about whether a given response and one selected from a database column are identical in meaning, not similar in meaning at all, or overlapping in meaning. This is an area for future research.

6. Summary and conclusions

The challenges of analyzing, interpreting, and reporting social sciences data are formidable. First, data are expressed in both words and numbers, depending on the research strategy employed. Secondly, although software for dealing with number-based data is well-developed and fully implemented in the social sciences, software for dealing with words is sorely lacking. Existing software has reduced time-demands but continues to reflect individual interpretations. A new paradigm in social sciences data collection and analysis is required.

A social sciences research project was examined that concerned people who self-injure (SI) for other than suicidal reasons. Qualitative data emerged about virtual communities for SI involving electronic bulletin boards where messages and replies are posted and available to everyone with access to the board. Since first appearing in 2001, the popularity of Internet-based SI peer-to-peer support groups has increased dramatically [33]. Whitlock, Powers, and Eckenrode [34] reported over 500 active self-injury-focused virtual communities. Tierney [35] noted that virtual communities encourage active involvement in personal wellness but also "reinforce dysfunctional or unhealthy practices and isolate individuals from society" (p. 182). Such phenomena are of great interest to social scientists.

Extensions to data warehousing functionality have been proposed to allow IB to be better used in social sciences research. IB was investigated for its ability to reduce subjectivity of interpretation and alleviate the effort required for preparing qualitative data for analysis. Methods such as compression and selective decompression for high performance data warehousing are needed for social sciences research data due to the complexity and data concentration of both column entries and column names. Proposed minimal extensions include:

1. Support for textual column values that tell a story
2. The ability to distinguish texts that require no decompression from those that require some.
3. Support for the query language to cluster texts based on the degree of decompression required to materialize them.
4. Provision of additional parameters regarding the amount of decompression required.
5. Expandable column names to reveal the full text of questions and expandable column entries to reveal the full text of answers given by respondents.

The information requested in items 1-4 above, reduces to the simple requirement that IB provide information about whether responses are the same, not the same, or overlapping and, if overlapping, provide a measure of the degree of overlap. With this we will have the ingredients to define the kind of semantic similarity that will aid analysis of social sciences qualitative research data. The information requested in items 1 through 4 is already available at the physical (storage) level, but we see it as also having a purpose at the conceptual level, specifically, for measuring the degree of overlap in the meaning of answers given by different subjects to corresponding questions.

Infobright appears to be limited for implementing a text analysis expert system. Therefore, the research problem has been redirected away from an original text analysis approach [37] and more toward a knowledge base approach to take better advantage of the IB analytic tool. In addition to this research project which was directed toward organizing knowledge around narrative stories as complex objects, IB is being used at Laurentian University is to teach database and knowledge discovery concepts to computer science students. We are in the process of developing the infrastructure for the undergraduate database course to augment the presentation of MySQL with hands on use of a data warehousing system (ICE Infobright Corporate Edition). To cover procedures and triggers, MySQL demands use of the INNODB storage organization. However, INNODB is not available in IB. Possible solutions have been suggested on the web based IB blogs. Data integrity achieved by means of constraints defined by database designers and enforced by the database management system must be covered at the undergraduate level. Hence, the possibility of overlaying a database (conveniently loaded using the simple IB LOAD command) with a schema of a richer structure and meaning was proposed. Infobright Enterprise Edition (IEE) running on our 8-core server with 8 gigs of RAM supports a maximum of 8 concurrent users sufficient for expected enrolments in a graduate course. Among the topics to be explored in the knowledge base course is the problem of how to achieve the minimal extensions to IB that were outlined in this paper. Additionally, each step of the KDD methodology outlined in Section 5.3 has been assigned to an individual graduate student.

Acknowledgement: We wish to thank Dominik Ślęzak, chief research scientist at Infobright, Inc., for his valuable input to this project. We would also like to acknowledge the funding provided by Natural Sciences and Engineering Research Council of Canada (NSERC) that permitted purchase of Infobright data warehousing/database software.

7. References

1. C-Store: A column-oriented DBMS, Stonebraker et al., Proceedings of the 31st VLDB Conference, Trondheim, Norway (2005)
2. Ślęzak et al., Brighthouse: an analytic data warehouse for ad-hoc queries, Proceedings of the 34th VLDB Conference, Auckland, New Zealand (2008).
3. Infobright open source data warehousing: Working smarter, not harder. Infobright IEE Technical Brief, March (2009).
4. Ślęzak, D., Wróblewski, J., Eastwood, V., and Synak, P. Bright-house: An Analytic Data Warehouse for Ad-hoc Queries. PVLDB 1(2) (2008) 1337-1345.
5. Ślęzak, D., Wróblewski, J., Eastwood, V., and Synak, P. Rough Sets in Data Warehousing. RSCTC 2008 (2008) 505-507.
6. Agnew, N.M., and Pyke, S.W. The science game: An introduction to research in the behavioral and social sciences (7th ed.). Oxford University Press (2007).
7. Liamputtong, P. Qualitative research methods (3rd ed.). Oxford University Press (2009).

8. Yoshikawa, H., Weisner, T.S., Kalil, A., and Way, N. Mixing qualitative and quantitative research in developmental science: Uses and methodological choice. Developmental Psychology, 44 (2008) 344-354.
9. Nie, N., Brent, D.H., and Hull, C.H. Statistical package for social sciences. New York: McGraw-Hill (1970).
10. Argyrous, G. Statistics for research: With a guide to SPSS (2nd ed.) SAGE UK, London (2005).
11. Auerbach, C.F. and Silverstein, B.L. Qualitative data: An introduction to coding and analysis. New York University Press, New York (2003)
12. Hilt, L.M., Cha, C.B., and Nolen-Hoeksema, S. Nonsuicidal self-injury in young adolescent girls: Moderators of the distress-function relationship. Journal of Consulting and Clinical Psychology, 76 (2008) 63-71.
13. Polk, E. and Liss, M. Psychological characteristics of self-injurious behavior. Personality and Individual Differences 43 (2007) 567-577.
14. Weitzman, E., and Miles, M. Computer programs for qualitative data analysis. Thousand Oaks: Sage (1995).
15. Lewins, A., and Silver, C.: Using software in qualitative research: A step-by-step guide. Sage, London (2007).
16. MacMillan, K. and Koening, T. The wow factor: Preconceptions and expectations for data analysis software in qualitative research. Social Sciences Computer Review, 22, (2004) 179-186.
17. Chien, S. and Immorlica, N.: Semantic similarity between search engine queries using temporal correlation. International World Wide Web Conference. WWW 2005, Chiba, Japan (2005).
18. Budanitsky, A. and Hirst, G. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics, 32(1) (2006) 13-47.
19. Hirst, G. and Budanitsky, A. Correcting real-word spelling errors by restoring lexical cohesion, Natural Language Engineering, 11(1), March (2005) 87-111.
20. Polkowski, L. Rough Mereology as a Link between Rough and Fuzzy Set Theories. A Survey. Transactions on Rough Sets (2004) 253-277.
21. Kickfire, Inc. A Revolution in High-Performance SQL Processing, A Kickfire Technology White Paper (2008).
22. Kickfire, Inc. Kickfire data warehouse appliance, Kickfire data sheet (2010).
23. Netezza TwinFin data warehouse and analytic appliance, Netezza TwinFin Data Sheet (2010)
24. DBX: Empower your data, XTREMEDATA Product Brief (2009).
25. Jean, J.S.N., Dong, G., Zhang, H., Guo, X., and Zhang, B., Query Processing with An FPGA Coprocessor Board(available by Microsoft Academic Search).
26. Leung, K.T., Ercegovac, M., and Muntz, R.R. Exploiting Reconfigurable FPGA for Parallel Query Processing in Computation Intensive Data Mining Applications (downloadable from Citeseer).
27. Russom, P. Next generation data warehouse platforms, TDWI Best Practices Report: fourth quarter (2009).
28. Ślęzak, D. and Marcin, K. Intelligent Data Granulation on Load: Improving Infobright's Knowledge Grid, Springer Berlin, Heidelberg, 2009.
29. Bird, S., Klein, E., and Loper, E. Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit, O'Reilly Media (2009).
30. Adamic P., Babiy V., Janicki R., Kakiashvili T., Koczkodaj W.W., and Tadeusiewicz, R. Pairwise Comparisons and the Visual Perception of Equal Area Polygons, Perceptual and motor skills 108(1) (2009) 37-42.
31. Bozoki, S., and Rapcsak, T. On Saaty's and Koczkodaj's inconsistencies of pairwise comparison matrices. Journal of Global Optimization, 42(2) Springer, (2007) 37-56.
32. Koczkodaj, W.W., Robidoux, N., and Tadeusiewicz, R. Classifying visual objects with the method of pairwise comparisons, Machine Graphics & Vision, 18(2) (2009) 143-155.
33. Adler, P. and Adler, P. Self-injurers as loners: The social organisation of solitary deviance. Deviant Behaviour, 26, (2005) 345-378.
34. Whitlock, J.L., Powers, J., and Eckenrode, J. The Virtual cutting edge: The Internet and adolescent self-injury. Developmental Psychology, 42, (2006) 407-417.
35. Tierney, S. The dangers and draws of online communication: Pro-anorexia websites and their implications for users, practitioners, and researchers. Eating Disorders, 14 (2006).181-190.
36. Infobright open source data warehousing: How to – Data Type Differences. Infobright Blogs. (2009).
37. Johnson, J.A., and Johnson, G.M. InfoBright for analyzing social sciences data. In D. Ślęzak, T. Kim, Y. Zhang, J. Ma, & K. Chung (Eds.), Communications in Computer and Information Science, 64, Berlin: Springer (2009) 90-98.

Julia Johnson is an associate professor of Computer Science in the Department of Mathematics and Computer Science at Laurentian University. She graduated with a Ph.D in Computer Science from the University of British Columbia in Vancouver, B.C., Canada in 1989. Her research includes the resolution of semantic ambiguity in natural language utterances and transportable architectures for natural language interfaces to database systems, software architectures to support concurrency controls in database systems, and applications of rough sets.	
Genevieve Johnson is an Educational Psychologist whose program of research includes theoretical, empirical, and practical understanding of Internet technologies and human learning and cognition. Increasingly complex cultural tools (e.g., the Internet) require, reflect, and facilitate increasingly complex cognitive processes (and vice versa). Genevieve was awarded a Ph.D. in 1990 by the University of Alberta (Canada) and, in 2007, a Graduate Diploma in Distance Education Technology by Athabasca University (Canada). For detailed information on her research, visit http://members.shaw.ca/gen.johnson/