The big data between your ears: Human inspired heuristics for forgetting in databases

Conference Paper · July 2015

DOI: 10.1109/ICMEW.2015.7169754

2 authors:

Gisela Susanne Bahr
Harris Corporation
48 PUBLICATIONS **143** CITATIONS
SEE PROFILE

Stephen Wood
Florida Institute of Technology
31 PUBLICATIONS **352** CITATIONS
SEE PROFILE

THE BIG DATA BETWEEN YOUR EARS: HUMAN INSPIRED HEURISTICS FOR FORGETTING IN DATABASES

Gisela Susanne Bahr & Stephen Wood, Florida Institute of Technology

ABSTRACT

Inspired by explanations of human forgetting, the primary research question that motivates this research is whether databases can be manipulated to resemble human long term memory in performance. We hypothesize that selective degradation of a database using different heuristics proposed as explanations of human forgetting, produce different recall results. Furthermore, we explore how these results resemble human performance by mapping performance to five memory phenomena that typify human forgetting. For this research we implemented the artificial long term memory of a fictitious character, Mr. Alfred Polly, using the Ardemia modelling architecture, which is built on a relational database management system. The experimental investigation of three versions of Mr. Polly's selectively degraded artificial memory revealed that forgetting heuristics produce query results that resemble human forgetting. The implication of this study is that big data algorithms inspired by "human forgetting inspired heuristics" can be a tool for shrinking and managing large data using attributes or metadata.

Index Terms— Artificial Memory, Ardemia, Forgetting, Databases, Human Inspired Heuristics, TimeGlue

1. INTRODUCTION

The primary question that motivates this research is whether explanations of human forgetting can inspire databases manipulations that give rise to performance patterns that resemble human long term memory (H-LTM). The relationship between H-LTM and big data is easy to see: conservative estimates of the capacity of H-LTM capacity range in the order of petabytes [1] based on 1 billion neurons. The actual number of neurons in the human brain is close to 100 billion, hence this figure required a scaling factor of 100 that is unlikely linear [2]. Human data resembles big data not only by volume but also by variety: H-LTM is supported by a multiplicity of systems dealing with different data types, such as semantic (factual), episodic (personal and emotional), modality (sensory), skills and motor based data [3]. In addition, human memories can be classified into implicit (non-declarative) and explicit (declarative) data types, of which the former can be verbalized and articulated but not the latter.

How the brain manages its big data and multi-media content over time has been investigated by cognitive psychologists who study retrieval and forgetting mechanism as well as specific retrieval phenomena. The mechanisms correspond to forgetting processes and the retrieval phenomena are memories that represent the product of forgetting. These two research topics form the basis and inspiration for the current study that investigates forgetting processes using human inspired heuristics and their effect on retrieval and memory phenomena (patterns) in the context of a database, with the overarching goal on meaningful big data reduction. To date, this is the only study that is based on a relational artificial memory that is designed based on the organization and associativity of H-LTM [4]. We present a review of the relevant research on H-LTM in section 2, which is followed by the experimental section in section 3. Section 3 includes a brief description of the simulation architecture and environment (Ardemia[1]), heuristics (manipulation) and metrics designs, analyses and results. The last section, section 4, presents the discussion of the results and implications and future directions, i.e., how heuristics can be used to make databases forget as if they were human.

2. CHANGE OVER TIME: FORGETTING MECHANISMS & PHENOMENA

Cognitive psychologists have proposed a number of explanations why people forget. These ideas fall generally into two classes of forgetting mechanisms [5]: the first is based on compromised data health that results from data decay and link loss (lost associations). In this first category, forgetting occurs because memories (data) change over time as a result of updates, additions, modification, and deletions of data. The second class of explanations of forgetting focuses on data volume and data access. These approaches explain forgetting as the result of interference from too much available data or the failure to use appropriate retrieval cues to get to the data. Rather than focus on the data changes over time, such explanations are concerned with the investigation of search and retrieval strategies (algorithms) and their effect on available memories (data). In the current study, we focus on modelling human inspired

[1] The name "Ardemia" is pronounced ('Ar *–Dee* – Me – Ah') and the phonetic spelling of the acronym RDMA, which stands for *Relational Data Memory Architecture [4]*.

forgetting that affects the system as a whole over time and involves data modifications and deletions.

In addition to studying the generic underlying mechanism for change over time, human forgetting can be studied by product or output, i.e., recall patterns; in fact, abundant anecdotal and empirical evidence indicates that humans display specific recall phenomena in everyday memory [7]: examples are *Source Forgetting, People Forgetting, Schema Knowledge, Temporal Knowledge, and Dream Amnesia*.

- Source forgetting is the inability to remember how, when, where or from whom information was learnt;
- People forgetting is the failure to recall a person or a person's attributes;
- Pattern and schema knowledge is the recall of typical occurrences or event;
- Recall of the temporal knowledge is the ability to remember whether an event happened before, during or after another event the ability to report which events happen more often than others, and
- Sleep and dream amnesia is the inability to remember sleep duration and stages, including REM (dream).

Within these recall phenomena there exist general quantitative differences that are based on anecdotal and experimental observations. For instance, humans exhibit enhanced schema knowledge compared to detail knowledge such as to source memory [7]. Last but not least, dream forgetting is a general form of amnesia that affects the majority of humans and results in overall depressed levels of recall compared to all other phenomena [8].

In summary, there are two classes of forgetting mechanism and a set of recall phenomena that are associated with different kinds of forgetting and typify human forgetting during recall.

3. HUMAN INSPIRED FORGETTING RESEARCH QUESTIONS AND STUDY DESIGN

The primary research question is whether databases can be manipulated, inspired by explanations of forgetting, to resemble H-LTM in performance. It is self-evident that degrading a "perfect recall" database is going to affect its output. This approach resembles the memory impairment or forgetting. We expect that using different heuristics for making databases forget leads to different recall results, hence selectively impair recall. What is not clear is how these impairments map to the everyday recall phenomena, i.e., are any of the forgetting mechanisms investigated here more likely to result in a human-like recall performance? In summary, we hypothesize that by selectively degrading databases with forgetting mechanisms, different recall results will emerge for each mechanism and can be observed in the query results. Furthermore, we explore how these results resemble human performance by mapping the memory phenomena and typical human performance. Simply, we initially show that the results vary based on the forgetting mechanism and then how they compare to human memory performance.

Table 1. Every day memory query in natural language

Q1: Where did you meet the Jets ?
Q2: How did you feel when he first met Sam?
Q3: Where did you last meet Sam?
Q4: How did you feel when you first met the Jets?
Q5: Which Jet did you meet first? Lance or Mike?
Q6: Do you remember George? How you happen to know him, where you met and how you felt at the time?
Q7: Do you remember Neil, how do you happen to know him, where you met him and how he felt when you met him?
Q8: How many times per week do you usually pick up your laundry from Mr.Pipps?
Q9: Where did you usually see Ruby?
Q10: Sorry to be indiscreet but exactly did you do?
Q11: Where did you actually see her?
Q12: Mr. Polly, What item do you sell the most?
Q13: Mr. Polly, What do you sell the least?
Q14: Mr. Polly, Do you less gloves or less mufflers?
Q15: Mr. Polly, In the springtime do you sell more hats or handkerchiefs?
Q16: Have you ever met the Archbishop of Canterbury, what is his name?
Q17: What you dream about at night?
Q18: What have you been dreaming about recently?

The design of the study was based on explanations of the forgetting mechanism and observable recall phenomena. Specifically, forgetting explanations inspired the design of the data manipulation with heuristics (independent variable), and the recall phenomena served as observable and dependent variables (metrics). More specifically, in the current investigation forgetting was treated as a function of

the data changing over time. Hence, data driven explanations were the focus of the current simulation experiments. (Interference and inappropriate cues are topics of future investigations.) We investigated two cases: decay manifested in missing data, and data implicification evident in changing data tags (binary attribute in Ardemia) from explicit to implicit. (Implicification is a made up term consisting of implicit and the Latin facere, i.e., to make implicit.) While the change from explicit to implicit memory is not necessarily an intuitive notion, there appears to be a forgetting mechanism that changes explicit experiences to implicit knowledge. As mentioned in the introduction, explicit memories can be articulated while implicit memories seem to be forgotten and cannot be articulated; nevertheless, they are recallable and support behavior and cognition. The details of the heuristics design and the dependent measures are described in the following paragraph of this section.

3.1. Simulation Architecture and Environment

For the simulation an artificial memory was constructed in a relational data base system, whose schema represented the various memory subsystems as relations and the associativity between the subsystems (relations), using a temporal variable called TimeGlue. TimeGlue was operationalized as date & time stamp. TimeGlue supports an associative and as well as inference systems [4].

The architecture and environment for the artificial memory is called Ardemia and the current instance of an artificial memory was populated with six months of memories of a fictitious character Mr. Alfred Polly. The primary source of the data population was the novel "The history of Mr. Polly" and secondary sources were historic events, daily activities and probable affect based on Mr. Polly's profile and personality. Mr. Polly's artificial memory was implemented in MSSQL 2008R. This choice was arbitrary. Other solution approaches, such as Oracle 12c or open sources alternative such as MySQL can be used to implement the artificial memory.

3.2. Designing the manipulation (Heuristics)

The goal was to design the manipulations in order to test how to make a database more human and less perfect. (A perfect memory system was operationalized as a database with perfect recall, i.e., complete and correct query output.) The instrumental concept was forgetting, which was operationalized as degrading perfect memories. Since cognitive psychologists continue to entertain multiple ideas why humans forget, it seemed reasonable to selectively degrade data to test our hypotheses. The degradations created two conditions: in the implicification condition, explicit memories were shifted to implicit storage, and in the decay condition data were deleted. These two manipulations were a small subset of the possible combinations of degradation designs but established a starting point for future research

3.2.1. Heuristics

In the decay condition, data were selectively preserved and removed using the following heuristic: preserved memories included memories that were accompanied by intense affect and those that were highly distinctive, i.e., unusual. All other data were removed including their TimeGlue.

In the implicification condition, data were modified (implicified) using the following heuristic: all explicit memories without consciousness during encoding (while sleeping[2]), without distinctiveness during encoding (while feeling bored, while engaging in routine activities) were implicified.

Each manipulation required its own database. Therefore a database was created for each condition and populated with the original data which remained preserved (un-manipulated) in the Control Condition. The code and the effect on the data conditions can be seen in in Table 2.

3.3. Experimental Design & Dependent Variables/Metrics

Three conditions (two with different types of degradation and one control) were evaluated based on their performance on 18 queries, which served as probes into an artificial memory, by tapping into every day memory and the five recall phenomena. (The queries in natural language can be seen in Table 1.) From the queries we constructed two sets of dependent measures: query accuracies and phenomena recall accuracies. (See the right most column of Table 4 for the mapping of queries to the phenomenon.) All measures were based comparisons of the degradation conditions to the control condition and a comparison of the degradation conditions to each other. The comparisons between manipulation and control were prefixed with the term *absolute*, e.g., absolute query accuracy. The comparisons between the conditions were prefixed with the term *relative*, e.g., relative query accuracy. Overall the design of the dependent measure was motivated by the degree of memory accuracy in human recall [9] and the degree of overlap between the causes of forgetting modelled (relative metrics).

More specifically, the measures were designed as follows: the result of each query was a relation with zero or more tuples. Hence it is easy to construct a more or less complex metrics involving the number of tuples recalled, the size of the tuples recalled or the content of the memory.

---

[2] One might argue that no encoding takes place during sleeping, but to investigate this question is not the objective of our study. The periods of sleep were included in the database because (a) they involved dreaming which may be memorable and (b) humans have autonoetic awareness of having slept.

Table 2. Code for Database Degradation by Condition and Effects on Rows by Condition

Condition1: Control	No changes to database	No rows affected.
Condition 2: Implicified Data	/*Implifying work memory while feeling bored, while engaging in routine activities*/ UPDATE EPISODIC_M SET RECALLABLE = 'implicit' WHERE HowMuch = 'low' AND (WHAT = 'sells' OR WHAT = 'talks to' OR WHAT = 'serves' OR WHAT = 'cleans' OR WHAT = 'studies' OR WHAT = 'checks' OR WHAT = 'decorates' OR WHAT = 'counts' OR WHAT = 'folds' OR WHAT = 'sweeps') ; /*Sleep and Dream Memory*/ UPDATE EPISODIC_M SET RECALLABLE = 'implicit' WHERE HowMuch <>'high' AND (WHAT LIKE '%sleep%' OR WHAT LIKE '%dream%'); /*Memory Bias for Age Peers in Gangs*/ SET RECALLABLE = 'implicit' WHERE AGE <> 20 AND (CIRCLE = 'sharks' OR CIRCLE = 'jets');	This preserved 3427 rows of explicit episodic memory and creates 940 implicit memories. The ten most common (non-distinctive) work activities were degraded. This preserves 46 explicit sleep or dream related memories of 1283 original and implicifies 1236. This preserves 61 of 78 explicit memories and implicifies 17 memories
Condition 3: Missing Data	use MrPollyLTM_MissingData; DELETE FROM EPISODIC_M WHERE HowMuch = 'low'; (This requires the disabling of any foreign key constraints)	This reduces the number of rows in the Episodic Memory from 4367 to 703.

We chose semantic content of the tuples because it is the only measure that supports the evaluation of the accuracy of recall comparing it to a control condition; this was our absolute accuracy metric. (Comparisons between the degradations yielded the relative absolute accuracy metric.) Consequently, the conditions were not compared based on their time performance (as in traditional benchmarking) but based on how much the memories (tuples) retrieved resembled each other and across conditions including the control database.

Another dimension of the queries is their mapping by everyday memory phenomena. Therefore, the raw data of the query outputs can be factored by memory phenomena. This created five a priori factors, namely Source Forgetting, Forgetting, Schema knowledge, Temporal Knowledge, and Dream Amnesia. This set of dependent measures was called phenomena recall accuracies. (The comparisons between degradations were prefixed with the term relative, such as relative source forgetting.)

3.4. Analyses Plan

When comparing three conditions in an experimental study, standard approaches to the analyses engage correlational techniques and mean comparisons following analysis of variance techniques like ANOVAs or in the case of multivariate measures, MANOVAs. Such techniques make assumptions about the dependent variables. For example, the data should be normally distributed and the dependent variable should be at least on an interval scale. Our experimental data are relations and do not meet these assumptions. Therefore, analyses based on comparisons of means or correlations are inappropriate. Instead set comparisons were conducted on the memories retrieved consisting of zero or more tuples. To our knowledge, a significance test for mathematical set comparisons does not exist. Therefore comparisons were performed based on the percentage of the content intersection (percentage overlap) between each condition and the control and between the conditions using both types of query accuracies and phenomena accuracies. Qualitative comparisons to known human performance patterns were examined to contextualize the results. Given the nature of these results they were presented in the results section.

3.5. Experimental Results

The results are reported for two sets of dependent measures absolute accuracies and phenomena recall accuracies: For the absolute accuracies the percentage of overlap between query outputs is shown by condition in Table 3. The

rightmost column indicates the everyday memory phenomena involved.

The results indicate that the implicification condition performed with 100% absolute accuracy for 15 of 18 queries and the missing data condition performed with 100% absolute accuracy for 8 of 18 queries. The combined relative accuracy of implicified memory and missing data memory was 10 of 18 queries. The relative absolute accuracy for queries 17 and 18 was based on semantic comparisons of sets whose tuples (raw data) were either retrieved in a different order or whose tuples were a proper subset of the other condition.

Table 3. Absolute and Relative Accuracies by Query

	Absolute Accuracies		Relative Accuracies
	Control ∩ Implicified	Control ∩ Missing Data	Implicified ∩ Missing Data
Q1	100%	100%	100%
Q2	100%	0%	0%
Q3	100%	0%	0%
Q4	100%	100%	100%
Q5	100%	100%	100%
Q6	100%	100%	100%
Q7	0%	100%	100%
Q8	100%	0%	0%
Q9	100%	100%	100%
Q10	100%	100%	100%
Q11	100%	100%	100%
Q12	100%	0%	0%
Q13	100%	0%	0%
Q14	100%	0%	0%
Q15	100%	0%	0%
Q16	100%	0%	0%
Q17	50%	50%	100% (different order)
Q18	0%	0%	100% Implicified (subset Missing Data)

For the phenomena accuracies the performance of each "forgetful" condition compared to the control and each other listed by memory phenomenon in Table 4. The pattern of performance is summarized in Table 4. For ease of data perception and interpretation see Figure 1. It appears that phenomena accuracy for source memory and people memory were least impaired in the implicification condition and most impaired in the missing data condition (83% vs 67% and 90% vs. 70%, respectively). There was some overlap between the conditions as seen in the relative source and people accuracy (67% and 70%, respectively). A similar pattern was observed for the temporal knowledge phenomenon accuracy (83% vs 50%) with 56% relative temporal knowledge accuracy.

Table 4. Absolute and Relative Phenomena Accuracies

Everyday Memory	Absolute Phenomena recall accuracies (PRA)		Relative PRA	Query mapping
	Implicified (avg. % recalled)	Missing Data (avg. % recalled)	Implicified vs Missing (avg. % recalled)	
Source Forgetting	83%	67%	67%	1, 3, 4, 6, 7, 16
People Forgetting	90%	70%	70%	1, 2, 3, 4, 5, 6, 7, 9, 11, 16
Temporal Knowledge	83%	50%	56%	8, 9, 10, 11, 12, 13, 14, 15, 17
Schema knowledge	94%	39%	44%	2, 3, 4, 5, 6, 7, 8, 15, 17
Dream Amnesia	25.0%	25.0%	100.0% (75%)	17,18

4. DISCUSSION

The results of the absolute accuracies indicated that implicifed and missing data memories differ from the control group memories. Moreover the results of the relative metrics indicated that simulations of different degradations produced different recall patterns. Hence both, the absolute and relative accuracies by query indicate that selective impairments can be modelled using artificial memories, like Mr.Polly's in Ardemia. Furthermore, the results of the degradations may provoke ideas into explanations of forgetting by contextualizing the performance of the queries on the five memory phenomena.

Both degraded long term memories showed the greatest impairment for dream and sleep related memories. This result was realistic because memory for sleep related experiences is generally not available for recall: many of our dreams and the experience of sleeping are forgotten. Overall, implicified memories showed some impairment on the remaining phenomena but ranged in the 83th to 94th recall percentiles. Recall was greatest for schema knowledge, then for remembering people, which was followed by a tie for remembering temporal knowledge and sources. It could be argued that this pattern is somewhat realistic by optimizing the recall of schema knowledge,

considered the hall mark of human cognition, and by minimizing remembering temporal memory and source memory; Source forgetting is a well-established phenomenon and based on our results it could be argued that implicification is the mechanism for source forgetting that preserves schema knowledge. Nevertheless, the recall results from implicified memory are overall elevated and source memory, although least memorable, was recalled in 83% of cases. The results for the Missing data memory or decayed memory ranged from 39th to 70th percentile and indicated a different performance pattern. Recall was greatest for people, followed by sources, then temporal knowledge and lastly, ranking lowest with 39%, schema knowledge. While the overall depressed recall percentage seemed to resemble low recall expectations associated with Ebbinghaus' forgetting curve [10], the pattern seemed less realistic because schema memory, considered critical to human cognition, was recalled less than in 40% of cases. While the pattern of recall in a system compromised by decay did not resemble human recall, the absolute values of source forgetting greater for decayed memories than for implicified memories and hence more realistic.

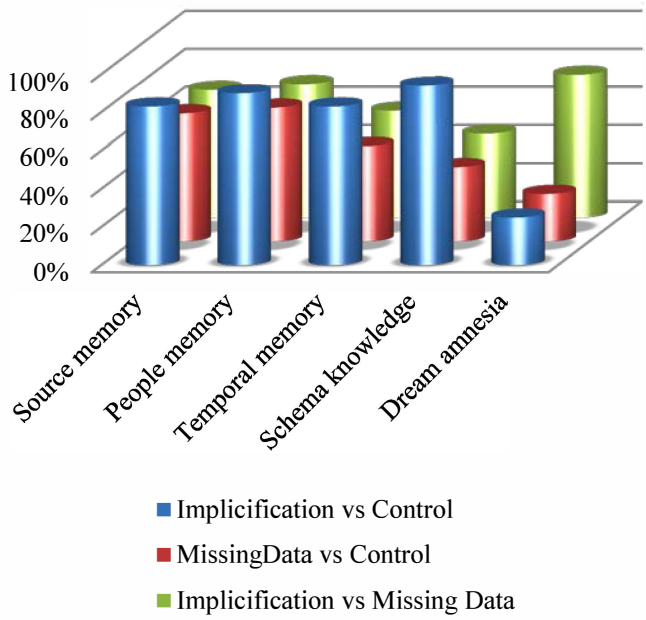

Figure 1. Phenomena Recall Accuracies by Condition

4.1. Conclusions

It is obvious simulations of artificial memory have virtues and shortcomings. The important findings are that forgetting can be selectively modelled using an artificial memory that are implemented in a relational database system. The decision for the current study to design the forgetting algorithms using attributes, was data and implementation driven. Data other than structured data, such as semi-structured or unstructured data can be degraded based on annotations and metadata. The key observation is that human-like forgetting can be accomplished in a database that the resulting, unique performance patterns can be analyzed. Shrinking oversized data or identifying which data can be transferred to tertiary storage are a few of the applications of human inspired heuristic forgetting. The method of evaluating database corruption using a set of benchmark queries, in this study to investigate every day memory, may be of interest for tampering detection as well. In the current study, the manipulations or degradation were designed a priori. In reality, the causes of data corruption are not necessarily known but it may be inferred from performance patterns. It follows that benchmark queries may be developed that reveal specific types of database corruptions thereby creating a diagnostic for maintaining and verifying data health. In general, the implication of this study is that big data algorithms inspired by human forgetting heuristics can be a tool for shrinking and managing large data using attributes or metadata.

5. REFERENCES

[1] P. Reber. (2010, April, 19). What Is the Memory Capacity of the Human Brain? [Online]. Available: http://www.scientificamerican.com/article.cfm?id=what-is-the-memory-capacity

[2] R. W. Williams and K. Herrup, "The control of neuron number," Annu. Review Neurosci., vol. 11, pp. 423–53, 1988. doi:10.1146/annurev.ne.11.030188.002231

[3] E. Tulving 1985, How many types of memory are there? American Psychologist, 40, 385-398.

[4] Bahr, G.S. (2014). Of Human Memory and Databases: Ardemia The Relational Data Model and Management System as a Set Theory Based Modeling Architecture for Human Long Term Memory. Retrieved from Proquest Dissertations & Theses. (Publication number AAC 3648513).

[5] R. A. Bjork, "Interference and forgetting," in Encyclopedia of learning and memory, 2nd ed., J. H. Byrne, Ed. New York: Macmillan, 2003, pp. 268-273.

[6] Neisser, U., & Hyman, I. E. (2000). Memory observed: Remembering in natural contexts. New York: Worth Publishers. ISBN 978-0716733195

[7] D.L.Schacter, J.L. Harbluk, and D.R. McLachlen (1984). Retrieval without recollection: an experimental analysis of source amnesia. Journal of Verbal Learning and Verbal Behaviour, 23(5): 593-611.

[8] C. Koch (2004). The Quest for Consciousness: A Neurobiological Approach. Englewood, CO: Roberts & Co.

[9] A. Koriat, M. Goldsmith, and A. Pansky, "Towards a psychology of memory accuracy," Annu. Review Psychology, vol. 51, pp. 481–537, 2000.

[10] H. Ebbinghaus, On memory (1885), H. A. Rutger and C. E. Bussenius, Translators. New York: Teachers' College, 1913. Paperback edition, New York: Dover, 1964. http://psychclassics.yorku.ca/Ebbinghaus/index.htm