Human Memory and Computer Caches

Abstract

The abstract should be one paragraph, indented 1/8 inch on both sides, in 9 point font with single spacing. The heading **Abstract** should be 10 point, bold, centered, with one line space below it. This one-paragraph abstract section is required only for standard spoken papers and standard posters (i.e., those presentations that will be represented by six page papers in the Proceedings).

Keywords: memory, caching, information retrieval

General Formatting Instructions

For standard spoken papers and standard posters, the entire contribution (including figures, references, everything) can be no longer than six pages. For abstract posters, the entire contribution can be no longer than one page. For symposia, the entire contribution can be no longer than two pages.

The text of the paper should be formatted in two columns with an overall width of 7 inches (17.8 cm) and length of 9.25 inches (23.5 cm), with 0.25 inches between the columns. Leave two line spaces between the last author listed and the text of the paper. The left margin should be 0.75 inches and the top margin should be 1 inch. **The right and bottom margins will depend on whether you use U.S. letter or A4 paper, so you must be sure to measure the width of the printed text.** Use 10 point Times Roman with 12 point vertical spacing, unless otherwise specified.

The title should be in 14 point, bold, and centered. The title should be formatted with initial caps (the first letter of content words capitalized and the rest lower case). Each author's name should appear on a separate line, 11 point bold, and centered, with the author's email address in parentheses. Under each author's name list the author's affiliation and postal address in ordinary 10 point type.

Indent the first line of each paragraph by 1/8 inch (except for the first paragraph of a new section). Do not add extra vertical space between paragraphs.

Problem Formulation

Suppose all the data a person or agent has seen and used is the set D. We will have a finite amount of memory $C \subset D$ items $c_0, c_1, ..., c_k$. We also have H, an access history of each item. This can be described as a set of sets that each item was accessed at, ie $t_0, t_1, ..., t_n$ for each $c_i \in C$. Whenever an item needs to be accessed, it must be retrieved from memory. If the item is in C, we can retrieve it for a small cost of $cost_c$. If the item is not in cache, we must retrieve it from D, for $cost_d$, where $cost_d > cost_c$.

Our goal then, is to minimize our cost the retrieving items we need from memory. This means that given C and H, we want to minimize our cost of retrieving sequences of items we need. To do this, we must manage C in order to maximize retrevals from C, and minimize retrevals from D, where the cost to retrieve is much higher.

For computer science purposes, C represents our cache, a small but very fast block of memory in a computer. D represents disk space, which has far more storage, but is orders of magnitude slower.

The Algorithms

Least Recently Used (LRU) Perhaps the most simple of caching, LRU evicts the item that was used the least recently. Because the only information needed to implement LRU is order of uses, it can easly be implemented with just a list. LRU's strength function can be described by

$$S_{LRU}(i) = \frac{1}{t_{current} - t_n}$$

Quite simply, the strength of a memory is proportional to how much time has passed since its last access.

We also used a random replacement algorithm to approximate LRU. For this algorithm, a random item is selected for eviction.

Least Frequently Used (LFU) Another straightforward approach is to evict the item that has been used least. While often a good approach, it is less feasible to implement, as LFU requires an ubounded amount of additional memory to store counts of item uses. LFU's strength function is

$$S_{LFU}(i) = n$$

since an item's strength grows the more times it is used. LFU can be very effective when items are used often but not necessarily in temporal proximity. A major downside to LFU is that the cache can become littered with items that were once extremely popular, but might never be used again.

LRU-2 LRU-2 is a specific instance of the LRU-K algorithm. LRU normally doesn't do a good job of accounting frequency, so LRU-K is a way to approximate LFU without all of the memory requirements LFU suffers from. LRU-K's strength function is

$$S_{LRU-2}(i) =$$

2Q 2Q is an algorithm that tries to find a balance between accounting for recency and frequency by splitting the cache into two queues. The first queue is managed as an LRU queue. If a hit occurs in this queue, the item is promoted to the second queue, which is managed as a LFU queue. The LFU queue has a predefined maximum size, so items will be evicted from the LFU queue if it is larger than the predefined size. The strength of an item in the cache will depend on how large the LFU queue is. If it is larger than the set limit, the strength of an item will be $S_{LFU}(i)$ if the item is in the LFU queue, and infinite otherwise (it wont be evicted). If the LFU queue is smaller than this threshold, an item's strength is $S_{LRU}(i)$ if it is in the LRU queue, and infinite otherwise.

One thing that 2Q also does is keep track of the items that were evicted from the LFU queue, by storing pointers to their address in memory. If one of these evicted items should be used again, the item can be brought back into memory much faster, making the miss far less costly. Because storing a pointer takes a very small amount of space, it can reduce the cost of the LFU misses with a small addition to the memory requirement.

An issue with 2Q is the fact that where LFU queue size has a strict set maximum size, and this maximum size may be hard or impossible to estimate effectively ahead of time.

Adaptive Replacement Cache (ARC) ARC is very similar to 2Q but makes generally improving complication. The maximum size of the LFU queue can adapt to the data incoming. ARC keeps track of items that have been evicted from each queue. Upon a cache miss of an item that was recently evicted from one of the queues, ARC will make that queue larger, as it got rid of an item that it should have kept. ARC's strength function is identical to that of 2Q.

LRFU LRFU subsumes both LRU and LFU. Each item has a combined recency-frequency count. Intuitively, an item strength continually climbs the more it is used, but that strength decays with time. The exact function and parameter that compute this strength can vary. As suggested in Lee et. al. (2001), we calculated the strength of an item as:

$$S_{LRFU}(i) = \sum_{i} \frac{1}{2}^{\lambda(t_{now} - t_i)}$$

λ is a tunable parameter. For our purposes, .001 worked well.

The Anderson Model Anderson Schooler (1991) designed a model of human memory to explain phenomena they witnessed in human data. There were 3 observed phenomena they wished to explain:

* Recency - The more recently an item was used, the more likely it is to be used again.

* Practice - The more times an item has been used, the more likely it is to be used again.

* Spacing - The pattern of prior uses of an item is also a predictor of furutre usage probability. This pattern can be described by the amount of time passing in between uses, and how much time passed since this interval occured.

The reasoning behind spacing may not be intuitive, but it desribes the idea that acesses of an item may follow a pattern. For the library example, books about taxes are probably accessed far more frequently during tax season. Once tax season is over, these books can probably be brought back to reserves despite their frequency and recency.

$$S_{And}(i) = A \sum_{i} s(t_i) s(t_i) = t_i^{-d_i} d_i = max[d_1, b(t_i - t_{i-1})^{-} d_1]$$

Third-Level Headings Third-level headings should be 10 point, initial caps, bold, and flush left. Leave one line space above the heading, but no space after the heading.

Formalities, Footnotes, and Floats

Use standard APA citation format. Citations within the text should include the author's last name and year. If the authors' names are included in the sentence, place only the year in parentheses, as in ? (?), but otherwise place the entire reference in parentheses with the authors and year separated by a comma (?, ?). List multiple references alphabetically and separate them by semicolons (?, ?, ?). Use the et al. construction only after listing all the authors to a publication in an earlier reference and for citations with four or more authors.

Footnotes

Indicate footnotes with a number[1] in the text. Place the footnotes in 9 point type at the bottom of the page on which they appear. Precede the footnote with a horizontal rule.[2]

Tables

Number tables consecutively; place the table number and title (in 10 point) above the table with one line space above the caption and one line space below it, as in Table 1. You may float tables to the top or bottom of a column, set wide tables across both columns.

Table 1: Sample table title.

Error type	Example
Take smaller	63 - 44 = 21
Always borrow	96 - 42 = 34
0 - N = N	70 - 47 = 37
0 - N = 0	70 - 47 = 30

Figures

All artwork must be very dark for purposes of reproduction and should not be hand drawn. Number figures sequentially, placing the figure number and caption, in 10 point, after the figure with one line space above the caption and one line space below it, as in Figure 1. If necessary, leave extra white space at the bottom of the page to avoid splitting the figure and figure caption. You may float figures to the top or bottom of a column, or set wide figures across both columns.

CoGNiTiVe ScIeNcE

Figure 1: This is a figure.

[1] Sample of the first footnote.

[2] Sample of the second footnote.

Acknowledgments

Place acknowledgments (including funding information) in a section at the end of the paper.

References Instructions

Follow the APA Publication Manual for citation format, both within the text and in the reference list, with the following exceptions: (a) do not cite the page numbers of any book, including chapters in edited volumes; (b) use the same format for unpublished references as for published ones. Alphabetize references by the surnames of the authors, with single author entries preceding multiple author entries. Order references by the same authors by the year of publication, with the earliest first.

Use a first level section heading for the reference list. Use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 1/8 inch. Below are example references for a conference paper, book chapter, journal article, technical report, dissertation, book, and edited volume, respectively.