Hypothesis and Evidence

Hui Chen

CUNY Graduate Center

Spring 2019

Acknowledgement

- Justin Zobel. 2015. *Writing for Computer Science* (3rd ed.). Springer Publishing Company, Incorporated.

How does research begin?

- Develop specific questions that the research aims to answer
- Require an understanding of how something works, or interacts, or behaves (an informal model)
- Establish a hypothesis, a statement of belief about how the object being studied behaves

Hypothesis

- Question: Is it possible to make better use of the cache on a CPU to reduce computational costs (New algorithms? New systems?)
- Hypothesis: a tree-based structure with poor memory locality will be slower in practice than an array-based structure with high locality, despite the additional computational cost.
- Research question: can a particular sorting algorithm be improved by replacing the tree structure with the array structure?
- Phenomenon: as the number of items to be sorted is increased, the tree-based method should increasingly show a high rate of cache misses compared to the array-based method.
- Evidence: the numbers of cache misses for several sets of items to be sorted. Alternatively, external evidence might be used, such as changes in execution time as the volume of data changes.

Hypothesis must be testable

- Assuming a computer scientist is investigating two data structures for searching, P-lists and Q-lists
- Example hypothesis: Good or bad?
  - Q-lists are superior to P-lists

Hypothesis must be testable

- Example hypothesis: Good or bad?
  - As an in-memory search structure for large data sets, Q-lists are faster and more compact than P-lists.

Hypothesis must be testable

- Example hypothesis: Good or bad?
  - As an in-memory search structure for large data sets, Q-lists are faster and more compact than P-lists. We assume there is a skew access pattern, that is, that the majority of accesses will be to a small proportion of the data

Hypothesis should be capable of falsification

- Example hypothesis: Good or bad?
  - Q-list performance is comparable to P-list performance
  - Our proposed query language is relatively easy to learn
  - Our search engine can find interesting Web pages in response to queries

Forms of Evidence

- Proof
- Model
- Simulation
- Experiment

Proof

- A formal argument that a hypothesis is correct
- Examples

Proof: Example in Mayers, 1999

- Gene Myers. 1999. A fast bit-vector algorithm for approximate string matching based on dynamic programming. *J. ACM* 46, 3 (May 1999), 395-415. DOI=http://dx.doi.org/10.1145/316542.316550

LEMMA 2. $Xh_j(i) = \exists k \leq i, Peq[t_j](k)\ and\ \forall x \in [k, i-1], Pv_{j-1}(x)$.[3]

PROOF. Observe from formulas (4b) that for all k, $Mh_j(k)$ is true iff $Pv_{j-1}(k)$ and $Xh_j(k)$ are true. Combining this with Eq. (8), it follows that $Mh_j(k) \equiv ((Pv_{j-1}(k)\ and\ Peq[t_j](k))\ or\ ((Pv_{j-1}(k)\ and\ Mh_j(k-1))$. Repeatedly applying this we obtain the desired statement by induction:

$$
\begin{aligned}
Xh_j(i) &= Peq[t_j](i) \quad or \quad Mh_j(i-1) \\
&= Peq[t_j](i) \quad or \quad (Pv_{j-1}(i-1)\ and\ Mh_j(i-2)) \\
&\qquad\qquad\quad\ \ or \quad (Pv_{j-1}(i-1)\ and\ Mh_j(i-2)) \\
&= Peq[t_j](i) \quad or \quad Pv_{j-1}(i-1)\ and\ Peq[t_j](i-1)) \\
&\qquad\qquad\quad\ \ or \quad (Pv_{j-1}(i-1)\ and\ Pv_{j-1}(i-2)\ and\ Peq[t_j](i-2)) \\
&\qquad\qquad\quad\ \ or \quad (Pv_{j-1}(i-1)\ and\ Pv_{j-1}(i-2)\ and\ Mh_j(i-3)) \\
&= \ \ldots \\
&= \exists k \leq i, Peq[t_j](k)\ and\ \forall x \in [k, i-1], Pv_{j-1}(x) \qquad (\text{as } Mh_j(0) = 0).
\end{aligned}
$$

□

Proof: Example in Jansson, et al., 2016

- Jesper Jansson, Chuanqi Shen, and Wing-Kin Sung. 2016. Improved Algorithms for Constructing Consensus Trees. *J. ACM* 63, 3, Article 28 (June 2016), 24 pages. DOI: http://dx.doi.org/10.1145/2925985

We now analyze the worst-case running time of Algorithm `Maj_Rule_Cons_Tree`.

THEOREM 3.3. *Algorithm* `Maj_Rule_Cons_Tree` *constructs the majority rule consensus tree of* $\mathcal{S}$ *in* $O(nk)$ *time.*

PROOF. We first show that in Phase 1, every iteration of the main loop in Step 3 takes $O(n)$ time. To perform Step 3.1 in $O(n)$ time, run Day's algorithm (see Section 2.1) with $T_{ref} = T_j$ and then check each $\Lambda(T[v])$ to see if it occurs in T_j. By Theorem 2.1, this requires $O(n)$ time for preprocessing, and each of the $O(n)$ nodes in $V(T)$ can be checked in $O(1)$ time.[4] The *delete* operations take $O(n)$ time in total since every node's parent is changed at most once (the nodes are handled in top-down order, so if some node is deleted then the new parent of its children cannot be deleted in the same iteration). Next, Step 3.2 can be implemented in $O(n)$ time by letting $P :=$ `One-Way_Compatible`(T_j, T) and $Q :=$ `Merge_Trees`(P, T), updating the structure of T to make T isomorphic to the obtained Q, and setting the counters of all new nodes to 1. This works because according to Theorem 2.8, P is a tree consisting of the clusters occurring in T_j that are compatible with the set of current candidates, and by Theorem 2.6, Q is the result of inserting each such cluster into T, if it did not already occur in T. There are $O(k)$ iterations in the main loop, so Phase 1 takes $O(nk)$ time.

In Phase 2, Step 5.1 is executed in $O(n)$ time by applying Day's algorithm like in Step 3.1. Thus, the loop in Step 5 takes $O(nk)$ time. Step 6 can be carried out in $O(n)$ time by treating the nodes in top-down order as above. In total, Phase 2 also takes $O(nk)$ time. □

Model

- A mathematical description of the hypothesis or some component of the hypothesis

Model: Example in Blei, et al., 2003

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.[1]

LDA assumes the following generative process for each document $\mathbf{w}$ in a corpus D:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

-- In Blei, et al., 2003

Model

"Essentially, all models are wrong, but some are useful."

--- In Box, George E. P.; Norman R. Draper (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley. ISBN 0471810339.

Simulation

- An implementation or partial implementation of a simplified form of the hypothesis, in which the difficulties of a full implementation are sidestepped by omission or approximation

Simulation: Example in Lin, et al., 2003

- Yi-Bing Lin, Wei-Ru Lai and Jen-Jee Chen, "Effects of cache mechanism on wireless data access," in *IEEE Transactions on Wireless Communications*, vol. 2, no. 6, pp. 1247-1258, Nov. 2003.
  doi: 10.1109/TWC.2003.819019,
  Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1244802&isnumber=27887

“We utilize discrete event simulation to model wireless data access. In the simulation two types of events are defined … The flowchart of the CB simulation is shown in Fig. 7, and is described as follows.

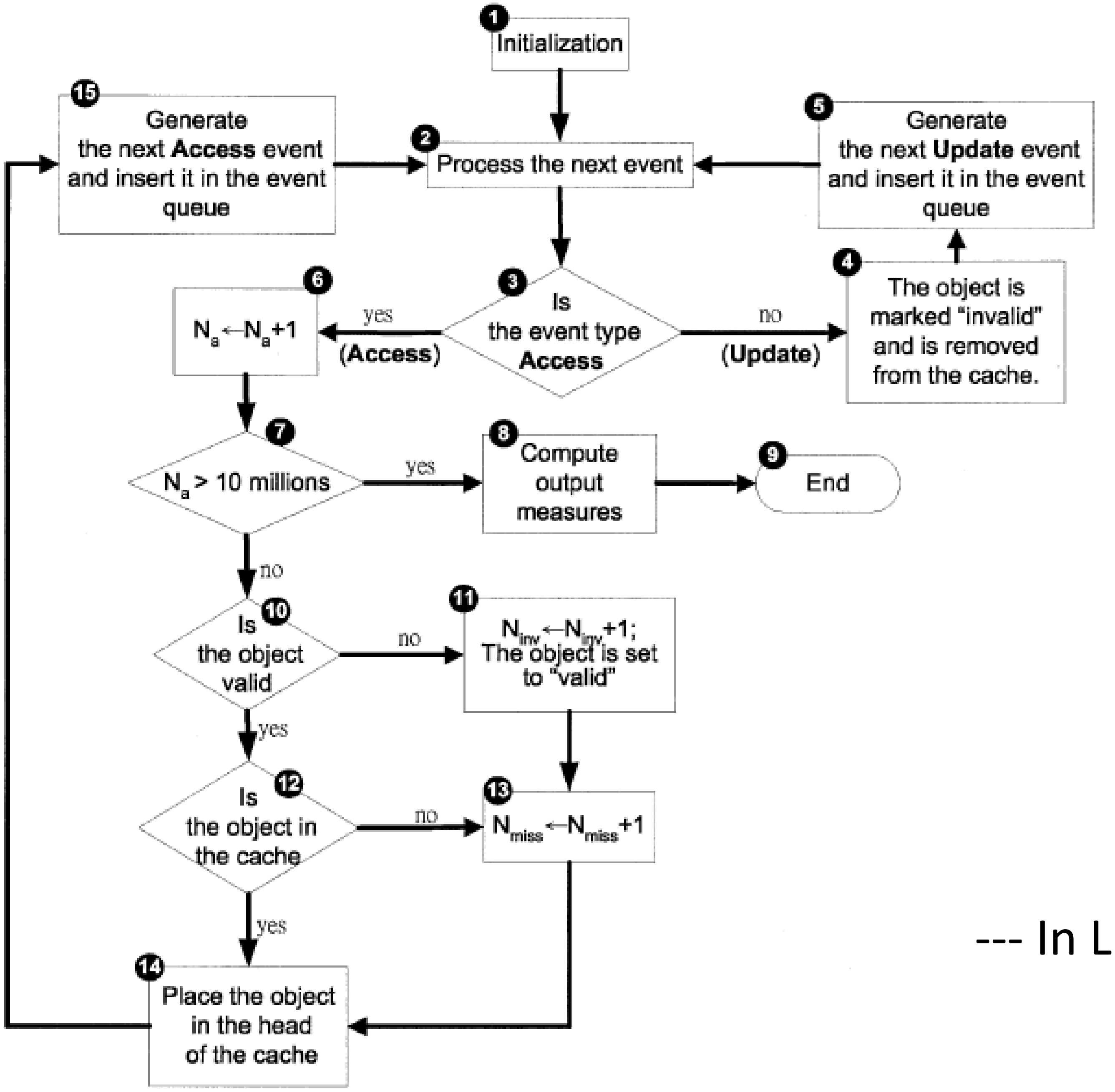
Step 1. Initially, all data objects are marked valid. The first Update and Access events are generated. The timestamps of these events are computed based on the interarrival time distributions previously mentioned. The events are inserted in the event list in the nondecreasing timestamp order.

Step 2. The event at the head of the event list is processed.

Step 3. If the event type of is Access, then Step 6 is executed. Otherwise Step 4 is executed.

……”

--in Lin, et al., 2003



--- In Lin, et al., 200

Some Remarks on Simulations

- Researchers can adjust parameters in simulations to observe behavior across a wide spectrum of inputs or characteristics and to answer *what-if* questions
- Simulations are often considered as a modeling technique.
- There is always a risk that it is unrealistic or simplistic, with properties that mean that the observed results would not occur in practice.
- Simulations need to be verified against reality.

Experiment

- A full test of the hypothesis, based on an implementation of the proposal and on real—or highly realistic—data

Example: Experiment in Guy, et al., 2016

- Ido Guy, Inbal Ronen, Elad Kravi, and Maya Barnea. 2016. Increasing Activity in Enterprise Online Communities Using Content Recommendation. *ACM Trans. Comput.-Hum. Interact.* 23, 4, Article 22 (August 2016), 28 pages. DOI: http://dx.doi.org/10.1145/2910581

4. EXPERIMENTAL SETUP

4.1. Owner Survey

Our evaluation was based on a survey of community owners that included four rounds of recommendations. Each round was only sent to the owners who had responded to the previous round. Rounds were 2 weeks apart over a period of 6 weeks in total. The owner received a personal email invitation that included a link to the community and a link to the online survey of the particular round. The email described the survey and thanked the owner for participating in previous rounds, when relevant (full wording is provided in Appendix). We opted for a multiple-round survey, since we wanted to simulate, even if roughly, a real-life situation wherein recommendations are available on a regular basis, and inspect the effect on communities over time.

……

4.2. Research Tasks and Hypotheses

Our experiments were designed to address three key research tasks, each with its own hypotheses, as follows.

RT1: Compare the effectiveness of member-based, content-based, and hybrid profiles for recommendation sharing by owners

--- In Guy, et al., 2016

Remarks on Experiment

- In some cases, the distinction between simulation and experiment can be blurry.
- In principle, an experiment only demonstrates that the hypothesis holds for the particular data that was used, while modelling and simulation can generalize the conclusion (however imperfectly) to other contexts.
- Ideally an experiment should be conducted in the light of predictions made by a model, so that it confirms some expected behavior.
- An experiment should be severe seek out tests that seem likely to fail if the hypothesis is false, and explore extremes.

Use of Evidence

- Different forms of evidence are *complementary*
- Evidence needs to be *persuasive*
- Quality of evidence is often used to evaluate one's research work

Approach to Measurement

- To collect evidence, often requires to take measurements
- *What is to be measured?* and *what measures will be used?*
- Are the measurements logically connected to the aims of the research?
  - Research aim is qualitative
  - Measurements are quantitative

“Two philosophers are arguing in a bar. The barman goes over to them and asks, “What are you arguing about?”

“We’re debating whether computer science is a science”, answers one of them.

“And what do you conclude?” asks the barman.

“We’re not sure yet,” says the other. “We can’t agree on what ‘is’ means”.

--- In Zobel, 2014

Reflection on Research

- An iterative process in which theory and hypothesis dictate a search for evidence—or “facts”— while we learn from facts and use them to develop theories (Zobel, 2014).
- Falsification, confirmation, and proof
- Good or bad science? Pseudoscience? Strong or weak research?

Regarding hypotheses and questions (Zobel, 2014)

- What phenomena or properties are being investigated? Why are they of interest?
- Has the aim of the research been articulated? What are the specific hypotheses and research questions? Are these elements convincingly connected to each other?
- To what extent is the work innovative? Is this reflected in the claims?
- What would disprove the hypothesis? Does it have any improbable consequences?
- What are the underlying assumptions? Are they sensible?
- Has the work been critically questioned? Have you satisfied yourself that it is sound science?

Regarding evidence and measurement (Zobel, 2014)

- What forms of evidence are to be used? If it is a model or a simulation, what demonstrates that the results have practical validity?
- How is the evidence to be measured? Are the chosen methods of measurement objective, appropriate, and reasonable?
- What are the qualitative aims, and what makes the quantitative measures you have chosen appropriate to those aims?
- What compromises or simplifications are inherent in your choice of measure?
- Will the outcomes be predictive?
- What is the argument that will link the evidence to the hypothesis?
- To what extent will positive results persuasively confirm the hypothesis? Will negative results disprove it?
- What are the likely weaknesses of or limitations to your approach?