Great Ideas in Computing

University of Toronto CSC196
Winter/Spring 2019

Week 5: October 12-16 (2020)

Announcements

Announcements

- Assignment 2 has been posted. There may be an additional one or two questions. Although it is not due until October 28, I suggest that you should begin working on the assignment as soon as possible as questions on the assignment will be helpful for the quiz.
- The quiz will take place Monday, October 19 at 11:10 AM. We will start at 11:10 and end at 12:10 so that you will have a full hour for the quiz. I will make an annlouncement as to whether we will be running the quiz on Quercus or Marcus.

Agenda

Agenda for the week

- On Wednesday, our meeting was led by Professor de Lara discussing virtualization. Professor de Lara provided a very information packed discussion and in some sense an overview of operating systems (OS). It is interesting a topic (e.g., virtualization, limited memory) can be active for a while (say in the 70s), becomes somewhat inactive and then comes back (say in the late 90s) as a major topic due to changes in technology and applications.

  As discussed, the main goals of an OS is managing the memory, the CPU, and the I/O. Virtualization is used to provide the effect of separate OS's for many processes running simultanneously. In particular, virtualization provides *isolation* of the processes.

  I consider even the basic concept of an OS (even running one process at a time) as a great idea.
- On Friday, we begin our new topic, namely search engines. I will also answer questions you may have concerning the upcoming quiz.

New topic: search engines

As I have mentioned, I want to talk about search engines as a great idea in the sense of being a "killer application" and also leading to interesting computational issues that have energized the field of computing.

In doing so I am mostly talking about search engines as they exist today in terms of searching web documents. That is, search engines that take queries (in the form of key words or phrases) and produce a ranked list of documents.

I am mostly going to talk about search engines independent of the importance (and necessity) of having large pools of fast machines, high speed communication and massive storage.

That is, I am mostly going to talk about search engines in terms of their functionality and the basic computational ideas that make them work (so) well. This is another example of a great idea where greatness depended on new technology and where greatness may also be an inhibitor for thinking about how to move beyond the *current norm of key word based search*.

A little search engine history

Search engines are part of the topic of "information retrieval" once the domain of library science. Computerized information retrieval has been an application idea since the start of modern computing.

On the web page there is a link to a prophetic July, 1945 Atlantic article "As We May Think" by Vannevar Bush where he envisions something quite close in many respects to the modern web and hyperlinked documents.

The article begins with the following: "Consider a future device . . . in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

That is, some kind of semi-automated information retrieval has been thought about for over 70 years.

Some quotes for the Vannevar Bush article

There are a lot of anachronisms (in terms of what the technology will be, gender roles) in this article but more important there are many insightful ideas about the future of accessing information. Here are some quotes from that article.

"Much needs to occur, however, between the collection of data and observations, the extraction of parallel material from the existing record, and the final insertion of new material into the general body of the common record. For mature thought there is no mechanical substitute. But creative thought and essentially repetitive thought are very different things. For the latter there are, and may be, powerful mechanical aids."

More quotes from Bush's article in the Atlantic

"Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. ... The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain."

"Man cannot hope fully to duplicate this mental process artificially, but he certainly ought to be able to learn from it. In minor ways he may even improve; e.g., for his records have relative permanency. The first idea, however, to be drawn from the analogy concerns selection. **Selection by association, rather than indexing**, may yet be mechanized."

"Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified." Think now of hyperlinks.

"Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, memex will do."

The debate as to the nature of information retrieval

In the 1960's and 70', there was a “debate” (albeit not widely discussed outside of those interested in information retrieval) between those who felt that information retrieval (IR) (i.e. finding documents to satisfy an “information need”) was a subfield of AI (and more specifically natural language understanding) verses those who thought it could be best realized by more well established combinatorial, algebraic and statistical ideas.

That is, one constituency felt that we needed to be able to "understand" what a document was saying (and what people were requesting) so as to find relevant documents.

The other constituency felt that the claims of many AI researchers were not at all feasible and that again a more statistical/algebraic/combinatorial approach (devoid of any real "intelligence") would produe better results.

The debate continued

I had a course (1967) in IR from Gerald Salton, who (according to Wikipedia) was "perhaps the leading computer scientist working in the field of information retrieval during his time". His group at Cornell developed the SMART Information Retrieval System".

I am not a great historian but I believe the vector space model (which we will discuss) was his idea. Salton was a proponent of the statistical/algebraic/combinatorial approach. I think that he always felt that AI was over-hyped.

So who the debate?

The debate continued

I had a course (1967) in IR from Gerald Salton, who (according to Wikipedia) was "perhaps the leading computer scientist working in the field of information retrieval during his time". His group at Cornell developed the SMART Information Retrieval System".

I am not a great historian but I believe the vector space model (which we will discuss) was his idea. Salton was a proponent of the statistical/algebraic/combinatorial approach. I think that he always felt that AI was over-hyped.

So who the debate?

As of today, it is clear that the approach of the constituency represented by Salton has turned out to be the basis for the way we currently do search in the internet.

The debate continued

I had a course (1967) in IR from Gerald Salton, who (according to Wikipedia) was "perhaps the leading computer scientist working in the field of information retrieval during his time". His group at Cornell developed the SMART Information Retrieval System".

I am not a great historian but I believe the vector space model (which we will discuss) was his idea. Salton was a proponent of the statistical/algebraic/combinatorial approach. I think that he always felt that AI was over-hyped.

So who the debate?

As of today, it is clear that the approach of the constituency represented by Salton has turned out to be the basis for the way we currently do search in the internet. However, search engines today do incorporate ML into their retrieval algorithms.

Key word search

At a very very general level, we can think of current search as the following process:

1. A user converts an "information need" into a query (i.e. a set of key words)
2. The search engine is then an algorithm for the mapping: query × {collection of documents} → <ranked list of "relevant documents">.
3. Upon receiving highly ranked documents, the user may choose to refine the query.
4. This process continues until the user is either satisfied or gives up. How often do you have to refine your queries?

As we discuss the ideas behind key word search in search engines, it should be noted that there are many specific ideas and engine specfic details that go into making a search engine successful (in terms of the quality, speed, and coverage) and these ideas and details are kept reasonably confidential. Why?

End of Friday, October 16 class

We ended at slide 10 on Friday. We will continue Wenesday, October 21 repeating slide 10.

I hope that this discussion will make it clear why I think the "debate" was won (for the time being) by the those advocating the combinatorial, algebraic, statistical approach for wweb search.