Search

(Forthcoming, The Johns Hopkins Encyclopedia of Digital Textuality)

Yuk Hui

When computer scientist Donald Knuth started his 5th chapter on Searching in the third volume of *The Art of Computing Language* (1998), he began with the sentence “the chapter might have been given the more pretentious title ‘Storage and Retrieval of Information’; on the other hand, it might simply have been called ‘Table Look-Up’” (392). In between these two terms – one that is professional and geeky and another that sounds more layman and quotidian –one can find a history of search and the evolution of search technologies from the indexations of pictograms and logograms to modern information retrieval systems characterized by automation and various algorithms such as binary search, Fibonacci search, etc. (see ALGORITHM).

Indexation precedes all searches. To search is to look for something within a larger set of items by identifying one or multiple specified features; this ranges in our daily life from choosing an apple in the grocery store to looking for a result on Google. These actions require an index that allows us to find the most relevant result. This index is often referred as a “key” in information retrieval. Searching, sorting and indexing are activities that can hardly be discussed separately. Conceptually speaking, indexing establishes relations between different entities, and sorting in this sense organizes these relations. To understand searching it is necessary to understand the evolution of relations.

One can find the earliest indexing and sorting technology in the Sumerian culture towards the end of the fourth millennium. The Sumerians used pictograms inscribed on clay tablets to index their stocks in order to set up an ‘accounting system’ that recorded debts, archives for future search (Goody 1977: 75). We have to bear in mind that strictly speaking there is no semantic meaning in pictogram writing; what exists are *relations* created by *visual differences*. The set of visual differences present in pictograms is also the first metadata system that describes the relations between the entities, if we understand “metadata” according to its conventional definition: “data about data”. Search techniques evolve according to the emergence of new relations specific to the technical natures of the writings (pictographic writing, logographic writing, analogue writing and digital writings). For example, with the introduction of morphemes and phonetics, the varieties of relations between objects and words escalates and the description of objects becomes more explicit since more and more semantic meanings can be inscribed as metadata.

We have to recognize that relations, materialized and formalized in different technical forms, are at the core of information retrieval. Searching allows us to follow these relations and thus arrive at a certain end according to specific instructions. An information system that consists of a large amount of data demands a fine order of granularity of relations as well as algorithms that allow it to organize these relations effectively across different tables. The relational database introduced by the mathematician and information scientist Edgar F. Codd in the 1960s was the first to allow searching across multiple tables. Today we know that most of information systems are built on top of relational databases, which are based on two main concepts: ‘(1) data independence from hardware and storage implementation and (2) automatic navigation, or a high-level, nonprocedural language for accessing data.

Instead of processing one record at a time, a programmer could use the language to specify single operations that would be performed across the entire data set'(CSTB 1999:162). The attributes of a table specify the properties and relations of the data stored in it, at the same time they also generate a "second layer" of relations among themselves through comparisons, e.g. difference, sameness, etc.

The emergence of the World Wide Web in the late 80s posed a new challenge to search, since HTML files are not databases, but more like light-weighted SGML(Standard Generalized Markup Language) annotations. Thus we saw the emergence of search engines that archive and index contents and algorithms that display 'relevant' result to users. With the development of the web and the rapid growth of user-generated data, the question of indexation and annotation became more and more crucial as the digital objects could easily get lost in the universe of data and become invisible. Early technologies such as HTML (prior to HTML5) are not expressive enough to make explicit relations among various contents. Moreover, relational databases only gave permission to authorized users and it was difficult to query across different databases without multiple authentications.

At the beginning of the twenty-first century, we saw several significant solutions to these questions. One thing these solutions have in common is that they go beyond searching for specified features of digital objects (as one might search for a specific color red in an apple) to social searching. This new movement in search fostered three parallel movements. The first one is the development of Google's indexation technique known as PageRank. The PageRank algorithm recursively calculates an index of a web page according to the incoming links that are associated with it; the search engine then can determine where the page belongs in a search result. Google also record the online activities of its users, in the name of personalization, in order to determine the most relevant results for specific users. Google's attempt to find a compromise between "relevance of contents" through Page Rank and the "relevance to users" by personalization has become an important method for producing and organizing relations.

The second of these three movements is the semantic web proposed by Tim Berners-Lee (2000). The semantic web proposes to annotate web pages according to compulsory vocabularies, known as web ontologies and can be understood as a semantics shared between machines and human users. This virtual "sharing" of semantics implies certain sense of "social". The basic technical idea is to use RDF (Resource Definition Framework) tuples (ordered lists of elements) to describe web page contents in detail(by comparison with HTML). The semantic web aims to create a web of information that keeps all data in RDF stores, so everyone can openly search them according to semantic relations.

The third contemporary search solution is social tagging. In contrast to the semantic web, tagging doesn't use compulsory vocabularies, but encourages users to annotate digital objects with their own text. Searching through tagging provides an experience of serendipity based on the tags contributed by other users. It allows discoveries which are not formally related to search results. For example, by searching for a painting of shoes by Van Gogh, one can come to discover the writing of Heidegger on the origin of the work of art.

The connection between the social and the machinic has two significant implications. The first is the organization of knowledge on a global scale. Knowledge—ranging from simple facts, to videos on how to make bombs, academic seminars, book contents, etc.—can all be searched online. Search engines are the biggest players in the contemporary organization of knowledge, since they determine which comes first and which has to disappear while relying very much on the black boxes they created in the name of algorithms, personalization, etc. This has important political implications, in terms of privacy and information manipulation.

The second implication is the organization of cognition. Search tools provided by computational devices – including computers, mobile phones, tablets, etc. – become more and more important for our attempts to acquire knowledge, to the extent that we find it increasingly difficult to learn without them. This issue has been explored as a philosophy of externalities by such thinkers as Bernard Stiegler (tertiary retention), Andy Clark and David Chalmers (extended mind), John Haugeland (embedded mind) and Fred Dretske (externalism). Anthropologists such as André Leroi-Gourhan and Jack Goody have demonstrated the transformations of cognitive functions through writing, which create new circuits and become supports of the brain. Contemporary technologies such as search further complicate the manner in which humans think. The organization of cognition and organization of knowledge continue to converge; this is a subject that deserves exploration in the context of search.

References and Further Reading:

Berners-Lee, Tim. 2000. *Weaving the Web: the Origins and Future of the World Wide Web*. Orion Business.

Brin, Sergey and Page Larry, 1998. "The Anatomy of a Large-scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems* 30: 107-117

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58:10–23.
Codd, E.F. 1970. "A Relational Model of Data for Large Shared Data Banks" *Communications of the ACM* 13(6): 377-387.

Computer Science and Telecommunications Board (CSTB). 1999. Funding a Revolution: Government Support for Computing Research. Washington, D.C.: National Academy Press

Goody, Jack. 1977. The Domestication of the Savage Mind.: Cambridge University Press

Knuth, Donald Ervin. 1998. *The Art of Computer Programming*,vol. 3. Redwood City, CA: Addison Wesley Longman

Stiegler, Bernard. 2009. *Technics and Time*, vol.e 2: *Disorientation*. Stanford: Stanford University Press.