From punched cards to Google: an outline history of information retrieval

De las tarjetas perforadas a Google: una historia esquemática de la ciencia de la información

Alan GILCHRIST

Cura Consortium, alangilchrist77@gmail.com

Resumen

Se revisa la historia de la recuperación de información (IR) desde las tarjetas perforadas y la primera computadora programable (el ENIAC de 1945) hasta el actual buscador web Google y la "tecnología cognitiva" Watson de Microsoft. La revisión se basa en tres factores principales en el desarrollo de IR; (1) el enorme aumento en el poder de cómputo en los últimos 72 años, (2) la "competencia" entre el análisis estadístico del texto y el procesamiento del lenguaje natural (NLP) en la que ambos finalmente han convergido en gran medida, y (3) los cambios correspondientes en la intervención humana en el proceso de IR.

Palabras clave: Ciencia de la información. Recuperación de la información. Historia. Tendencias. Prospectiva.

Abstract

This paper reviews the history of information retrieval (IR) from punched cards and the first programmable computer (the ENIAC of 1945) to the present day Web searcher Google and Microsoft's "cognitive technology" Watson. The review is based on three major factors in the development of IR; (1) the enormous increase in computing power over the last 72 years, (2) the "competition" between statistical analysis of text and Natural Language Processing (NLP) in which the two have finally to a large extent converged, and (3) the corresponding changes in human intervention in the IR process.

Keywords: Information science. Information retrieval. History. Trends. Prospective.

1. Introduction

There have been four great revolutions in the history of human communication: spoken language, the invention of writing in Sumeria in around 2600 BCE, the Gutenberg press which introduced modern printing in Europe in the mid 15th Century; and the present age of the computer starting in earnest in 1945 with the launch of the ENIAC computer. The gap between writing and printing is thus roughly 4000 years and that between printing and the computer about 500 years. Seventy years after ENIAC we are increasingly affected by a rapidly developing revolution in information technology. Whether we think of clay tablets, papyrus scrolls, paper or tapes and disks people have stored these information carrying media in collections and libraries. So, from the earliest times people have 'retrieved' information embedded in documents, but it was not until 1950 that the term 'Information retrieval' was coined, attributed to Calvin Mooers by *the Oxford English Dictionary* with the words "The requirements of information retrieval, of finding information whose location or very existence is *a priori* unknown…"

This paper is concerned with the development of information retrieval (IR) from that time until the present with three major strands in mind: first, the extraordinary increase in computing power; second, the techniques employed in information retrieval, notably statistical analysis of text and competing techniques of Natural Language Processing (NLP); and third the radically changing activities of human intervention in the IR process due to the increasing sophistication brought about by the first two strands.

2. Genesis of the modern computer

Most histories of computers start with descriptions of the early use of punched cards at the end of the 19th and beginning of the 20th Centuries and which can be seen as precursors of the modern computer, but mention should also be made of two early pioneers in their attempts to produce calculating machines. The first is the brilliant French mathematician and philosopher, Blaise Pascal who, at the age of eighteen, devised a machine in 1642 based on cogs and gearwheels to operate addition and subtraction, thus helping his father who was a tax inspector. Some thirty years later the German polymath Gottfried Leibniz devised a similar but more sophisticated machine which was capable of multiplication and division as well as calculating square roots. Whereas these two innovations were born in the field of mathematics, punched cards were first devised in the late 18th Century and refined in 1801 by a French inventor who has given his name to the

‘Jacquard loom’, used to drive the machines employed for the weaving of complex patterns. His revolving chains of punched cards were made of wood! While the Englishman Charles Babbage was certainly aware of, and influenced by, the use made of punched cards by Jacquard he must also have been aware of the ideas of Leibniz probably because the latter had been elected a member of the Royal Society of London of which Babbage was also a member. Babbage worked on two complicated pieces of machinery, first the ‘Difference Engine’ and later the ‘Analytical Engine’, earning the title given by some as the ‘Father of Computers’. Neither of these devices were finished before his death in 1871, due to lack of funding and, it is said, his grumpy treatment of colleagues and employees, but replicas have been produced in this century and are on display in London’s Science Museum. The Analytical Engine could deal with all four arithmetic operations plus comparison and, optionally, the calculation of square roots. It incorporated an input mechanism using punched cards with output to a printer or a curve plotter, and stored the procedures for calculations by use of pegs inserted into revolving drums (a device used by Leibniz). A close acquaintance of Babbage was Ada, Countess of Lovelace (neé Ada Byron, the only legitimate daughter of the famous poet). Lovelace was a competent mathematician who understood the work of Babbage and later said that “machines could never truly think”, thus countering the extraordinary question put by an English Member of Parliament with the words “Pray, Mr Babbage, if you put into the machine wrong figures, will the right answers come out?”.

Despite these early innovations involving punched cards it was not until 1890 that the first seriously practical application of punched cards sorted by machine was invented by Herman Hollerith who worked for the U.S Census Bureau. Daunted by the time and effort expended on analysing the returns of the census of 1880, he set out to build what was subsequently called a ‘tabulating machine’, which ‘read’ the holes and their positions on the cards to arrive at accurate results of the data gathering exercise in one year rather than the eight taken for the previous census. Encouraged by this success Hollerith set up the Tabulating Machine Company, which became the Computing-Tabulating-Recording Company (C-T-R) in 1911 and which, in turn and after a series of mergers and acquisitions, later became the International Business Machines Corporation (IBM) in 1924. The stage was now set for big advances in the technology of computing and these were accelerated by a brilliant young Englishman called Alan Turing who conceived of a theoretical ‘Logical Computing Machine’. This was a philosophical leap from the engineering aspect of punched cards into the realms of mathematics and logic and represented a huge advance in the history of computing. His concept included the theoretical possibility of handling any mathematical computation from an input carried on continuous paper tape carrying data in binary code; the calculations being based on a ‘table of instructions’. In 1936 Turing continued his studies, started at Cambridge University in England, with a visit to Princeton University in America where he met and discussed his ideas with John von Neumann, who was later to be involved in most important developments some ten years later. This resulted in Turing’s seminal paper ‘On Computable Numbers’, and this fed into his work at the now historically famous Bletchley Park codebreaking centre where, during World War II and with Turing’s collaboration, a team built in 1943 the first electronic digital and programmable computer and which, called the Colossus, was used exclusively for the decryption of German signals.

3. The Modern computer

The success at Bletchley Park and a surge of theoretical and practical work in the U.S. and other countries produced significant results which heralded the start of years of concerted research and development. It is generally agreed that the first truly modern computer was the ENIAC which was unveiled in 1945, a colossal machine which, according to Wikipedia, weighed almost 35 tons and occupied a space of 167 square metres. It was equipped with 17,468 vacuum tubes to control currents, 7,200 crystal diodes, 1,500 relays, 70,000 resistors, 10,000 capacitors and 5 million soldered joints. It was kept in a carefully heat and atmosphere controlled environment, but even so, was very vulnerable; several vacuum tubes burnt out every day and had to be replaced. There is a story, perhaps apocryphal, that on one occasion moths got into the machine and short circuited it – giving rise to the now common expression “bugs in the machine”. Even with this vast array of components, the input and output was effected by punched cards and had no memory until new equipment was installed in 1953. However, its arithmetic calculations were some thousand times faster than electro-mechanical machines. With these dimensions and performance in mind one can forgive IBM President Thomas J. Watson, for saying in 1943 “I think there is a world market for about five computers”. Watson was soon proven wrong by an enormous wave of research and development in a number of countries leading to rapid and important advances in computer technology involving new start-up companies earning large amounts of money, and the establishment of the famous ‘Silicon Valley’, not to mention a number of Nobel prize winners on the way. From 1945 on, the history of the development of computers, seemingly rather paradoxical, is one where computers got smaller and smaller and at the same time more and more powerful. As early as 1946 the first commercial computer arrived, crafted by two engineers who had been responsible for the design of ENIAC, and named UNIVAC (UNIversal Automatic Computer). In fact, it was not released until 1951 when the U.S Census Bureau purchased a machine even though initially, UNIVAC employed clumsy input and output processes. In the 1950s two famous computer programming languages were released, COBOL (Common Business-Oriented Language) and IBM’s FORTRAN (FORmula TRANslation) this latter designed for numerical computation and scientific applications, and these were responsible for starting a revolution in the portability and standardization across different hardware platforms and operating systems. At the end of the 1950s there was another huge advance with the invention and introduction of integrated circuits replacing the cumbersome and unreliable vacuum tubes; this miniaturization of hundreds of tiny transistors onto a small piece of semiconductor material being popularly known as a microchip. These smaller, cheaper and much faster devices paved the way for the introduction and spread of mini- and microcomputers, and later the personal computer, thus confounding another poor prediction made in 1977 by Ken Olson, the founder of the Digital Equipment Company when he said, “There is no reason anyone would want a computer in their home”, another piece of evidence that showed how fast things were moving. In fact minis, micros and terminals spread very rapidly as they became cheaper and were used widely in workplaces, then installed in universities and later in schools so that students could be taught how to use, and even program them. The Proceedings of a Conference held in London in 1980, co-organized by the British Computer Society and the Institute of Information Scientists was published under the title *Minis, Micros and Terminals for Libraries and Information Services* (Gilchrist, 1981). In the preface it was noted that a computer specialist attending the Conference asked what was an inverted file: this was, in retrospect, an interesting question because it was the information scientists who had ‘invented’ the inverted file for searching computer held alphabetical indexes, where computer scientists used data dictionaries for their work in, mainly, data processing. Previously, information scientists had introduced the punched card known as a feature card which, unlike the edge-notched card, represented a subject and had usually 10,000 positions at which to punch the numbers given to the documents containing that subject – in practice, an inverted file. Since then, computer scientists have embraced text processing and the Institute of Information Scientists no longer exists. Refuting Olson, the computer then invaded the home with, in quick succession, the desktop, the laptop, the portable; and then in 2007 in yet another advance in miniaturization the iPhone was launched, a final combination of computer, screen and keyboard, and also the first device that combined these with the telephone (whose development is briefly reviewed below). These advances confirmed what became known as ‘Moore’s Law’ which stated that his observation that the number of transistors per square inch on a microchip had doubled every two years and was likely to continue to do so, and consequently that computer power would double at the same rate. This has proved to be the case though it is recognized that being an exponential growth can not continue indefinitely but it is an astonishing fact that with advances in nanotechnology, some transistors are smaller than a virus. In just over 70 years from the ENIAC to the iPhone it is possible to compare the two devices: the iPhone costs 17,000 times less than ENIAC; it is 40,000 times smaller; it uses 400,000 times less power; it is 120,000 times lighter…; but… it is 1,300 times more powerful.

Flops	*Computer*	*Decade*
Hundred – ten squared	ENIAC	1940s
Kilo – ten cubed	IBM 704	1950s
Mega - ten to the power ^{6}	CDC 6600	1960s
Giga – ten to the power ^{9}	Cray 2 (CERN)	1980s
Tera – ten to the power ^{12}	ASCI Red	1990s
Peta – ten to the power ^{15}	Jaguar	2000s
Exa – ten to the power ^{18}	???	

Table I. Processing power

What happens next? What happens if you pack large and powerful computers into a single connected system? The answer is ‘parallel processing’, already widely employed and known as the ‘supercomputer’, and whose applications will be briefly discussed at the end of this paper. The Table above shows the growth in computing power from the ENIAC onwards, and is measured in the technical term FLOPS standing for Floating

Point Operations per Second. This refers to a method of encoding real numbers within the limits of finite precision available in computers.

Jaguar was developed by Cray at the Oak Ridge National Laboratory in the U.S. In 2010 it was then ranked as the most powerful supercomputer in the world but in the same year was overtaken by a supercomputer developed in China. Since then more supercomputers have arrived in what looks like a race between the U.S and China. Currently the top spot is occupied by China's Sunway Taihu Light with an amazing strength of 93 petaflops and is also the most energy efficient of the top ten supercomputers.

Almost in parallel with the growth of computer technology, telegraphy arrived to join it in a spectacular way leading to today's revolution in information access and mobile computing. But, to complete a trilogy of embarrassingly bad predictions, when Edison visited London to promote his invention of the telephone he was rebuffed by Sir William Henry Preece, Chief Engineer of the Post Office with the words "…there are conditions in America which necessitate the use of instruments of this kind more than here. Here we have a superabundance of messengers, errand boys and things of that kind". Not only did extensive worldwide telephone networks rapidly develop, but by the late 1960s computer networks were developing such as ARPANET (Advanced Research Project Agency Network) which led directly to the Internet as a merger of a number of independent networks of which ARPANET was the leader. Then in the early 1990s Tim Berners-Lee working at CERN in Geneva is credited with inventing the World Wide Web, and in 1996 Serge Brin and Larry Page launched the Google search engine, these two innovations now being used by literally billions of people every day across the world. Information and Communications Technology has since had a fundamental effect on the social, political and commercial spheres and the world is still adjusting to its present impact and considering its future.

4. Retrieval techniques

4.1. Mathematical techniques

In the famous library in Assyrian Nineveh, the librarian was known as "the man of the tablets". Callimachus of the Alexandria Library devised a catalogue (on scrolls) listing the contents of several different subject-based collections, each arranged by author. Since these early endeavours man has diligently stored and arranged documents in various ways, making it easier to find specific items or documents that might be of interest to an enquirer. Classifications have been devised and catalogue cards created in the increasing endeavour to guide searchers to content 'hidden' within documents not easily detected from the titles or classification, or even multiple catalogue cards. Increasingly, books were being added to by other forms of document, notably scientific and technological reports and articles in journals. Within these documents, the scope, complexity and granularity of the topics discussed began to overwhelm librarians and a new intermediary came into being – the information scientist; initially a person (in the U.K.) officially qualified with a degree in science or technology plus a second language and five years experience in handling complex information. By the time of the arrival of the ENIAC computer in 1945, the problem was becoming worse and was brought to a head by the acquisition by the U.S. of a vast number of scientific and technological reports captured at the end of World War II, and this is when modern computers entered the field of text processing and the term 'Information retrieval' was born. However, before computers began to develop more sophisticated ways of handling text, mechanised sorting of punched cards was developed further by Calvin Mooers (mentioned earlier as the originator of the term 'information retrieval'). In the 1950s Mooers devised a complex system called Zatocoding such that a broad picture of the contents of a document could be stored on a single Hollerith card. The rationale for this was that a Hollerith card contained 960 positions and that with a four punch Zatocode it was possible to store 165 different 'subject ideas'. Mooers described Zatocoding as "the most efficient coding system presently known". It was also made available by his Company for selecting the smaller edge-notch cards using small mechanical sorters. Another primitive approach to the IR problem was initially applied by its inventor Hans Peter Luhn to the rapid indexing of the large quantity of captured reports mentioned earlier. This was the KWIC Index (Key Word In Context), a simple device that rotated the words in a title presenting them in an alphabetical order. For example, after deleting non-informative words, the title "Analysing sentences, an introduction to English syntax" would be presented as:

- Analysing sentences introduction English syntax
- English syntax analysing sentences introduction
- Introduction English syntax analysing sentences
- Sentences introduction English syntax analysing
- Syntax analysing sentences introduction English

The first realistic use of the computer for IR was to use standard Boolean logic (named after George Boole, the 19th Century English mathematician and inventor of Boolean algebra). This used the relationships AND, OR and NOT to combine search terms, usually taken from a list used in indexing the documents in the collection. Initially, complex constructs were fed into the computer which operated overnight in batch processing mode. The three simple combinations A AND B, A OR B, A NOT B were simple, but more complicated search expressions had to use brackets such as {(A OR B) AND C} or {(D AND E) NOT F} which became more complicated when all three Boolean operators were used in longer strings for a single search. Any mistake in the placing of brackets would lead to ambiguity causing false retrieval and necessitating the running of a corrected search again the next night. Online retrieval overcame this by the ability to conduct two consecutive searches, the second using the NOT function to reject unwanted documents. This approach was, of course, the simplest possible and was followed by many new improvements and alternatives. For example, in the late 1970s Stephen Robertson, a mathematician at the U.K. Cambridge University working with his computer scientist colleague Karen Sparck Jones applied probability theory using Bayesian statistics, to tackle the IR problem. This approach posited that the probability of finding a document that was relevant to the query addressed to a collection could be statistically calculated by comparing mathematical representations of the terms in the document and in the query, and even that portions of the document could be identified as being richer in relevance. These two academics later developed and applied a complete IR package under the name of Okapi BM25 (where BM stands for Best Matching). Here, the probabilistic approach was expanded to rank a set of documents based on the query terms appearing in each document without regard to such aspects as their proximity or other relationships. The system was applied successfully at London's City University where Robertson had become Professor of Information Science. In 1983 Gerard Salton of the American Cornell University and one of the most accomplished thinkers in the history of IR research, proposed the 'Extended Boolean model' which introduced the concepts of partial matching and term weights (also later used with the probabilistic approach described above) in an algebraic approach to the IR problem, and which involved more complex mathematics to create a 'vector space'. This used the algebraic approach to the problem that he had first introduced in the 1960s and continuously developed. Salton's full Vector Space Model (VSM), launched into the public domain as SMART (System for the Mechanical Analysis and Retrieval of Text) is a difficult concept to grasp for anyone not familiar with mathematical abstractions. In technical terms the model is a hypercube: an n-dimensional space in four (or geometrically more) dimensions. Vector quantities have both direction and size and can be used to describe a movement from one point to another. Within this hypercube vectors are stored and used in calculating the relationships between documents, words (as tokens) and queries with reference to the relevance of results from addressing queries to the collection of documents. Not surprisingly, the algebraic equations addressed by the computer are complex but there was more to come, specifically with Latent Semantic Indexing (LSI) introduced in the late 1980s. This approach introduced the idea that patterns between the semantics of terms, expressed as unit words or their compounds used in the texts of a collection could be correlated with their underlying abstract concepts. The mathematical technique used is called Singular Value Decomposition (according to Wikipedia "A technique used in linear algebra…a factorization of a real or complex matrix"), a description which underlines the complexity of the mathematics used in IR!

The techniques described briefly above fall into three mathematical categories: set-theoretic, probabilistic and algebraic and the examples given are the most notable in a wide field incorporating variations of these three approaches. As these techniques developed, commercial software companies added various refinements and further approaches in the hope of making the packages either more powerful or user-friendly, or both. Very many packages were developed and sold in countries world-wide, tailored to specific languages and requirements. In the U.K some of these were commercial off-shoots of software created and used in-house by large companies and institutions, such as the nuclear research laboratory at Harwell and a package called ASSASSIN created by Imperial Chemical Industries Ltd, which initially ran on IBM machines only, but was later developed to operate with the UK ICL computers. Competition was fierce and various ideas, some of which appeared odd, reached the market place. One example was a package called Excalibur., which was based on early experiences of its inventor started by watching an immobile chameleon which had the ability to identify a passing insect and to automatically launch its rolled-up tongue to catch it. The inventor saw this as pattern matching and developed a sensing machine that could distinguish between different species of felled tree trunks by the patterns visible on the cut planes. More ambitiously he moved on to applying pattern matching to IR.

4.2. Natural Language Processing

Another, initially more ambitious, approach to IR is provided by Natural Language Processing (NLP), one that has not solved the problem by itself but which has made many significant contributions to the IR problem and is likely to continue to do so. Elizabeth Liddy, not only an expert in NLP but a qualified librarian defines NLP in these words "Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications." This is a careful definition that does not claim that NLP can solve the IR problem by itself but that it can, and has, provided much significant help. The combination of the two different approaches, one involving the rigid discipline of mathematics and the application of the even more complex linguistic analysis inherent in NLP has proved increasingly fruitful in tackling the complex problems of IR, both employing the increasing power of computers. Liddy and others have described the complexity of language by defining six levels of semantic characteristics:

- Phonetic: the way in which words are pronounced.
- Morphological: referring to the smallest piece of a word that has a meaning, for example by ignoring prefixes and suffixes such as un- and -ation (though this can, by itself, lead to confusion in retrieval).
- Syntactic: the structure of a sentence, taking into account the roles played by different types of words (nouns, verbs etc.)
- Semantic: the meanings of words individually or in combination, taking into account synonyms and various combinations.
- Discourse: the different ways in which structure is used in different communications, for example in newspapers or technical articles.
- Pragmatic: the use of what is called 'synecdoche', for example the "White House" to indicate the seat of the U.S. Government (rather than a white house).

As evidence of how difficult a problem it is that faces NLP in its application to IR, it has been reported, some years ago, that Liddy and colleagues developed a prototype software IR package based on the six semantic levels above, but it was discovered that, at the time of its completion, there were too few computers in operation that had sufficient power to run the programs effectively and the plan was discontinued. However, the basic principles have been successfully applied since then, both manually by information scientists and incorporated in machine processing by computer scientists, thereby providing significant power to tackling the IR problem. Broadly speaking NLP can be used to improve relevance ranking by using such devices as automatic stemming of words, identification of proper nouns or, more cleverly identifying related terms at the concept level. Using many analytical tools NLP, it has been said, can extract additional information at all four stages of the IR process: document processing, query processing, query matching, and ranking and sorting: in short, attacking for example the problems of too many synonyms, too many meanings, inability to specify vague concepts, improving indexing consistency and avoiding variations and errors in spelling. As mentioned above, Latent Semantic Indexing has been applied within NLP to analyse the relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. If all of these actions seem (or did seem) ambitious one only has to look at the performance of Google in searching the Web.

4.3. Human intervention

Before the arrival of the modern computer the information scientist's role was carried out by librarians acting at the interface between the document collections and the readers. With the onset of the new technology, information scientists found themselves working at two interfaces; the first, working sometimes between the collections and the computers; and also at the interface between the computers and the users. In the first intervention the information scientist indexed the documents, later providing abstracts (some of which, in the scientific area were structured in a standard manner). The indexing was also formalised through the application of a standard vocabulary which device became increasingly sophisticated over time. Initially, the words used in indexing were called 'terms', though the American Mortimer Taube coined the word 'Uniterm' to indicate that it was a simple 'unit term'. This word did not catch on, but Taube did introduce another word which had a more lasting and fundamental effect – this was the word 'concept' which he used to define the basic idea underlying the word (or words); in reality, the actual meaning beneath the choice of words. Other semantic techniques, more concerned with the operations of the computer, consisted of simple NLP with such devices as stemming, phrase searching and word order, an unlikely example of this last being quoted in the early days of disambiguating between ' Blind Venetians' and 'Venetian blinds' (a type of window covering). Initially, searches came to be conducted by matching search terms with index terms and to facilitate this exercise simple vocabularies were created, many of which were limited in size (initially for use with punched cards). Tricks were used to economise on the number of words used in the vocabulary such as 'Father USE Male + Parent' (a device called Semantic Factoring), But soon, as computers became more widely used more extensive vocabularies were created and called thesauri. This term, from the Greek word 'thesauros' meaning 'treasury' was initially suggested by Helen Brownson of the U.S. National Science Foundation at a Conference in England, borrowing the word from its well-known use created by Peter Mark Roget, compiler in 1852 of the dictionary of synonyms and related terms of the English language, a work that has undergone numerous revisions ever since. The number of thesauri created then grew dramatically so that by the late 1970s the Directorate of the European Commission responsible for matters of information and technology was able to publish a directory containing over 1500 of these vocabularies used for IR.

One of the pioneer thesauri was that created by the Engineers Joint Council in the U.S., which was followed by two more produced by the Armed Services Technical Information Agency and the other by the American Institute of Chemical Engineers. In 1965 it was decided that these agencies should collaborate in combining these three thesauri into what became a huge publication created by 329 scientists working in a number of coordinated committees, whose effort was later calculated to amount to over 23 man-years. The thesaurus, called TEST (Thesaurus of Engineering and Scientific Terms) contained 17,800 descriptors with a further 5554 lead-in terms and cross-references adding up to 162,657 line entries. In addition there were a permuted index of the compound terms, a subject category index and an accompanying hierarchical index. Other thesauri, one which preceded TEST and the other which followed, broke new ground, the first in its display, the second in its structure. In 1965, Euratom (a European Commission research programme for nuclear research) published a thesaurus with graphical displays. These were initially created as sets of Euler circles and then presented to the users in the shapes of abutting polygons containing randomly placed descriptors with the top terms in each polygon underlined. These displays were augmented by others called 'arrowgaphs' in which uni-directional arrows indicated hierarchical relations and two-directional arrows indicated other relationships. This approach was further emulated in Europe and is later to be seen in what were called 'Topic maps' and later again in the more complex diagrams representing ontologies used in the semantic web (see below). A second innovation in thesaurus construction was the intellectual advance pioneered by Jean Aitchison with the Thesaurofacet, so called because it combined the features of a faceted classification and a thesaurus. In fact, the primary construct was a disciplined faceted classification from which the descriptors were extracted and alternatively displayed in thesaurus format employing normal thesaurus features. This thesaurus was published by the English Division of the American giant GE (General Electric), and contained some 16,000 descriptors augmented by a further 7,000 lead-ins, all terms meticulously connected by the faceted classification. Aitchison went on to create other dual classification/thesaurus works, principally the two-volume UNESCO Thesaurus. The thesaurus had now become the standard device for supporting computer-based systems so that, for example, the National Library of Medicine transformed its Index Medicus subject heading system into a more thesaurus-like structure to support its computerised database called MEDLARS (MEdical Literature And Analysis Retrieval System). Since those early days many more thesauri have been created including some large international and multilingual schemes such as those created by the Food and Agriculture Organization (called AGRIVOC), another by the European Commission to support parliamentary documentation (called EUROVOC) and the Art and Architecture Thesaurus produced by the Getty Foundation.

While all this was happening a revolution occurred in IR which began to diminish and in many cases removed the need for indexing. This was full text searching, an advance which relied far more on the computer to do the retrieval work supported by more advanced mathematical and NLP techniques, the origins of which have been described earlier in this paper. This wider range of semantic tricks included requests for two words within x words of each other and suggestions of synonyms derived from a dictionary incorporated in the system. The use of thesauri became search aids, not always used by what had become known as the 'end-users'. With many users of large organizations on scattered sites many were unsupported by trained information professionals and research in the private sector has shown a continuing dissatisfaction with systems of information access and provision. One small study in a U.K. government department in the 1980s was instructive; the information service had set up a template on users' screens that offered a choice of 'simple' or 'advanced' search. The first was a single box for a subject entry while the second offered a menu of different metadata that could be combined including subject but also author, publisher, etc. (The large topic of metadata is not included here). A small study showed that 95% of the users used simple search against 5% that used advanced search – a foretaste of Google.

There had been a number of national standards presenting guidelines for thesaurus construction since the late 1970s but a major new and thorough standard was published by ISO (2011,2013) in two parts with the title Thesauri and Interperability. A major advance in this work was the recognition of the importance of other vocabularies and the desirability of connecting them in shared subject areas. One of the new forms of structured vocabulary was the 'taxonomy' which, though the word means simply classification was a thesaurus-like structure with more elastic construction rules devised mainly as more user-oriented aids, principally to websites and intranets. In this, it followed in spirit the earlier 'folksonomies' created organically by end-users and preceding the now more common tagging. Meanwhile there were separate developments within the Semantic Web of which the most relevant to this paper can be seen in what is known as the 'Semantic Web Stack' showing the interrelated standards used from the underlying bases up to the user interface. At the lowest level are found the standards, Unicode for character sets and URIs (Uniform Resource Indicators). The URIs disambiguate descriptors with the same name, for example: Segovia, the city and Segovia, the musician. Further up is the standard for data interchange, the Resource Description Framework (RDF). This dictates the way in which resources are described using 'triples', composed of subject-predicate-object, for example: "Prado is situated in Madrid" Above the RDF level is found 'Taxonomies' used here to indicate any standardized vocabulary, and 'Ontologies', (another word borrowed from another era and use). This ontology is defined by Tom Gruber as "a formal specification of a shared conceptualization". As a more complex 'topic map', it can be created for describing the contents of a collection of entities, and unlike the thesaurus that proposes strict rules for relationship types (for example genus-species, part-whole, instance for hierarchies). However, the ontology can define such relationships and more, for example as in the Prado example used above in explaining the RDF. While these are powerful devices for relatively small and homogeneous collections problems grow as the collection increases in size and scope.

It might be useful here to mention the work done in previous years on retrieval performance. The influential experiments carried out in the 1960s in the U.K. at what is now Cranfield University. The Cranfield measures of retrieval effectiveness were 'Recall', the number of correct results divided by the number of results that should have been retrieved and 'Precision', the number of correct results divided by the number of all results. Cranfield compared four systems using the same corpus specially indexed for the experiment: an alphabetical subject catalogue, the Universal Decimal Classification, a faceted classification (specially constructed for the experiment) and a 'Uniterm' index. The actual results of the experiment are now a matter of history; what mattered was that the idea of evaluation using these two measures caught on and were used in many more tests. To these two measures the concept of user-judged relevance was added and relevance feedback loops were built into the IR process. More measures and measurement techniques were added and now the world leader in IR evaluation is undoubtedly TREC (Text Retrieval Conference), a series of workshops addressing different areas of IR research (called 'tracks'). TREC celebrated its 25th Anniversary in 2016 and supplies test collections and evaluation software to 93 groups working in 22 countries.

5. Epilogue

What next? First, 'quantum computing'. The brilliant physicist Richard Feynman once said, "I think I can safely say that nobody understands quantum mechanics." Nevertheless, according to the science journal Nature Google and Microsoft may release post-research models of quantum computers in 2017. If Feynman is right few people will understand the definition of quantum computing. According to Wikipedia, "Quantum computers make direct use of quantum-mechanical phenomena such as superposition and entanglement to perform operations on data". Meanwhile, in a recently published paper with the title "Combining word semantics within complex Hilbert Space for information retrieval", the introduction says" In quantum theory, states are represented by vectors defined as a complex-valued Hilbert Space…In information retrieval, pioneering work showed that the quantum formalism encompasses many state-of-the-art retrieval models and subsequent works proposed many quantum-like models." If all this sounds like futurology one should look at what Google and Microsoft are currently doing. After some years of operating massive powers of parallel computer processing, employing many of the retrieval models and NLP techniques (briefly described above, and augmented by "page-ranking", an algorithm to measure the relative importance of web pages, Google now employs another technique called "looping", which uses information about each individual's search patterns to "improve" their search results. This means that two people carrying out identical searches at exactly the same time may retrieve different results. This is intelligent manipulation of "big data", using NLP and other techniques, an approach which is said to be revolutionizing the older approach to artificial intelligence that started out trying to mimic the workings of the human brain. Microsoft is working in the same direction with its powerful system called Watson (named after IBM's first CEO). Early experiments with IBM's Deep Blue chess-playing machine which competed with Gary Kasparov, the chess champion prompted an interesting comment from him when he said, "What if, instead of human versus machine, we played as partners?". After trying that he commented "We could concentrate on strategy planning instead of spending so much time on calculations. Human creativity was even more paramount under these conditions." Subsequently, Professor Thomas Poggio of MIT has said "These recent achievements have, ironically, underscored the limitations of computer science and artificial intelligence. We do not understand how the brain gives rise to intelligence, nor do we know how to build machines that are as broadly intelligent as we are." Nevertheless, Watson has enormous power and it has been reported that it can process 500 gigabytes, the equivalent of one million books, per second. In fact, it has been reported (Otake, 2016) that Watson cross-referenced a patient's records with 20 million research papers and produced a correct diagnosis within ten minutes! While such power is being deployed successfully in such areas it is not yet replacing human decision-making. What can we expect to see by next year's Ibersid Conference? And what of Information Science and the application of information retrieval and knowledge organization in social systems? Contrary to some appearances, it has not vanished but is now even more diffuse and necessary and, though it is not fully understood or supported by much of academia and employers, is embedded in the applications of very many organizations, notably on the Web. The first winner in 1999 of the Tony Kent Strix Annual Award for advancing the art and science of information retrieval was Stephen Robertson (one time Professor of Information Science at City University and pat-time Research Associate at Microsoft's Cambridge Laboratory). In 2015 the Award was presented to Dr. Susan Dumais, Senior Researcher at Microsoft.

References

Gilchrist, Alan (Editor) (1981). Minis, Micros and Terminals for Libraries and Information Services. London. Heyden & Son Ltd., on behalf of the British Computer Society/ 1981.

International Standards Organization (2011). ISO 25964. Thesauri and Interoperability. Part 1: Thesauri for Information Retrieval. Geneva: ISO.

International Standards Organization (2013). ISO 25964. Thesauri and Interoperability. Part 2: Interoperability with other Vocabularies. Geneva: ISO.

Otake, Tomoko (2016). IBM big data used for rapid diagnosis of rare leukemia case in Japan. // The Japan Times. Aug. 11, 2016. https://www.japantimes.co.jp/news/2016/08/11/national/science-health/ibm-big-data-used-for-rapid-diagnosis-of-rare-leukemia-case-in-japan/#.Wyo-wS0rxSN

Further reading

This does not purport to be an academic study and material for its compilation was drawn from many places, including Wikipedia and other numerous websites. Consequently, it would be tedious and largely unprofitable for the reader to cite them all. There follows a short list of references that might be useful, but idiosyncratic and by no means comprehensive.

For the first section on Computers:

Isaacson, Walter (2014). The Innovators: how a Group of Hackers, Geniuses and Geeks created the Digital Revolution. London: Simon and Schuster.

For IR techniques:

Liddy, E.D. (2001-2). A breadth of NLP applications. // ELSNEWS, 10.4, Winter2001-2 (Newsletter of the European Network in Human Language Technologies).

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schutze (2008). Introduction to Information Retrieval. Cambridge. Cambridge: Cambridge University Press.

For human intervention:

Frické, Martin (2012). Logic and the Organization of Information. New York: Springer.

Vickery, Brian C. and Alina Vickery (2004). Information Science in Theory and Practice. Munich: K.G. Saur.

Enviado: 2017-03-13. Segunda versión: 2018-01-12.
Aceptado: 2018-01-12.