MARISTELLA AGOSTI

La storia dell'arte e della scienza dei motori di ricerca in Italia

(Memoria presentata in modalità telematica)

Estratto

Atti e Memorie dell'Accademia Galileiana di Scienze, Lettere ed Arti
già dei Ricovrati e Patavina
Volume CCCCXXI (2019-2020)
Parte II: Memorie della Classe di Scienze Matematiche, Fisiche e Naturali

PADOVA
PRESSO LA SEDE DELL'ACCADEMIA

ACCADEMIA GALILEIANA DI SCIENZE LETTERE ED ARTI
IN PADOVA
35139 Padova - Via Accademia, 7 - Tel. 049.655249 - Fax 049.8752696
e-mail: galileiana@libero.it - www.accademiagalileiana.it

MARISTELLA AGOSTI, s.e.

La storia dell'arte e della scienza dei motori di ricerca in Italia

(Memoria presentata in modalità telematica)[1]

Introduzione

Alla fine degli anni '50 del secolo scorso, si avviano gli studi dei metodi per la rappresentazione, la gestione e il recupero automatico dell'informazione che si inizia a rappresentare in formato digitale. Da questi studi nasce il settore del reperimento dell'informazione moderno – in inglese *Modern Information Retrieval*, in breve IR – che si sviluppa in continuità con il reperimento dell'informazione tradizionale, basato su metodi esclusivamente manuali per la preparazione di vari strumenti di indicizzazione e classificazione finalizzati a favorire il recupero manuale dei documenti custoditi nelle biblioteche e negli archivi. Fra questi strumenti molto comuni erano i cataloghi per autori e titoli, per materia e per soggetto, di solito realizzati con schede catalografiche cartacee. L'IR si arricchisce di metodi che contribuiscono alla rappresentazione e gestione dell'informazione in modo automatico, quindi metodi che iniziano ad utilizzare le caratteristiche della rappresentazione digitale e della gestione automatica dell'informazione mediante calcolatori elettronici – *computer* – che proprio in quegli anni cominciano ad essere resi disponibili in diversi contesti.

Il reperimento dell'informazione moderno nasce parallelamente negli Stati Uniti d'America e in Europa e l'Italia non fa eccezione.

([1]) La memoria doveva essere presentata nel corso dell'adunanza pubblica ordinaria del 14 marzo 2020, che è stata cancellata in conformità alle disposizioni riguardanti l'emergenza COVID-19. Una presentazione video della memoria è stata inserita nel mese di giugno 2020 nella sezione 'I Soci scrivono' del sito Web della Accademia Galileiana e successivamente è stata archiviata nel canale YouTube dell'Accademia ed è visionabile all'indirizzo https://www.youtube.com/watch?v=6CrTcUClvKs.

Infatti, quando nel 1961 fu fondata l'Associazione italiana per l'informatica e il calcolo automatico (AICA),[2] le attività della neonata associazione vengono organizzate anche con gruppi di lavoro dedicati alle aree già allora attive dell'informatica. Uno di questi gruppi è il Gruppo di Lavoro sull'Information Retrieval (GLIR), fondato nel 1962, che riuniva gli studiosi con background accademico e industriale che avevano iniziato ad affrontare vari aspetti dell'IR, che si basa su metodi propri dell'informatica, ma anche della biblioteconomia, della linguistica, della matematica, del calcolo delle probabilità e della statistica. Per questo motivo, l'IR è intrinsecamente interdisciplinare e il suo studio e avanzamento richiede la collaborazione di esperti con competenze diversificate e, quindi, per il suo sviluppo a studiosi di informatica si affiancano spesso studiosi di altre discipline.

Sono trascorsi quasi sessant'anni da allora e l'IR è cresciuto in modo coerente con le diverse evoluzioni, e talvolta 'rivoluzioni', che si sono verificate nell'informatica. Soprattutto a partire dalla seconda metà degli anni '90, con la diffusione dell'uso del Web, il recupero dell'informazione ha raggiunto uno dei picchi più alti di notorietà essendo il campo che ha dato vita agli strumenti di reperimento dell'informazione denominati 'motori di ricerca' e che oggi sono fra gli strumenti informatici più diffusamente utilizzati. A partire da quegli anni l'attenzione per l'IR è cresciuta così come l'interesse ai risultati scientifici e applicativi che man mano venivano raggiunti nel settore. Proprio perché i motori di ricerca sono strumenti informatici che vengono diffusamente e quotidianamente utilizzati per ottenere risposte alle più disparate esigenze informative, questa memoria illustra su quali principi opera un motore di ricerca e quali metodi è stato necessario utilizzare, se già disponibili, o ideare e sviluppare per rendere i motori di ricerca operativi e via via sempre più efficaci e performanti. Sarà necessario anche dare uno sguardo alla struttura – la 'anatomia' – di un motore di ricerca, perché la conoscenza delle componenti principali con le quali è costituito consente una sua utilizzazione con maggiore senso critico e più mirata allo scopo. Le caratteristiche dei motori di ricerca vengono qui illustrate ad un livello funzionale e non teorico o tecnico, perché sono le funzioni che rendono disponibili quelle caratteristiche che interessano i loro fruitori.

([2]) Indirizzo Web dell'Associazione italiana per l'informatica e il calcolo automatico (AICA): https://www.aicanet.it/

Che cosa è un motore di ricerca

Dal punto di vista informatico un motore di ricerca è un'applicazione software, è quindi un insieme di programmi che sono scritti in un linguaggio di programmazione idoneo a istruire un calcolatore elettronico a svolgere operazioni di ricerca e recupero di informazioni e/o documenti digitali disponibili nel Web. Le operazioni di ricerca vengono innescate dall'utente che fornisce al motore di ricerca una frase che esprime la sua esigenza informativa e che il motore di ricerca interpreta e elabora andando poi ad estrarre dagli archivi di informazioni e documenti che gestisce, quelli che valuta più pertinenti alla frase di ricerca che l'utente gli ha fornito. Infatti per usare un motore di ricerca un utente traduce una sua esigenza informativa in una frase di ricerca – *query* – e la scrive all'interno di uno spazio appositamente preparato nella pagina Web dello specifico motore di ricerca che ha deciso di utilizzare.

Ad esempio, se l'utente ha deciso di utilizzare il motore di ricerca Google[3] e desidera ottenere informazioni sulla Accademia Galileiana, la frase di ricerca che potrebbe scrivere è 'Accademia Galileiana in Padova', come indicato in figura 1, e la risposta che otterrebbe è quella indicata in figura 2.

Visto che si è utilizzato il termine 'applicazione Web', prima di procedere con la illustrazione delle caratteristiche funzionali dei motori di ricerca, e proprio perché i motori di ricerca sono applicazioni che operano all'interno del Web, occorre vedere un po' più da vicino che cosa è effettivamente il Web.

Web o World Wide Web

Il Web, forma abbreviata di *World Wide Web* (anche *WWW* o *W3*), può essere considerato un sistema informativo in cui sono disponibili numerosi documenti digitali di diverso tipo e dove ogni documento è

([3]) Google è una multinazionale, fondata nel 1998 negli Stati Uniti d'America, che oggi fornisce diversi servizi e prodotti informatici e telematici, ma che inizialmente era nota solo per il motore di ricerca denominato Google come l'azienda. Negli anni l'azienda si è molto diversificata, ma il nome con il quale il motore di ricerca viene 'raggiunto' dagli utenti per essere utilizzato è rimasto lo stesso. L'indirizzo Web dell'azienda è: https://www.google.com/

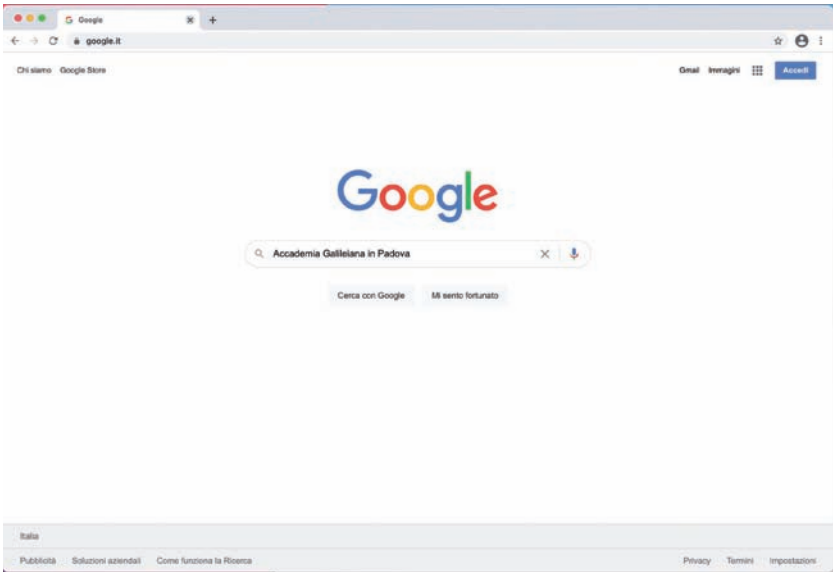

Fig. 1 - Pagina di accesso – o *home page* – del motore di ricerca Google.

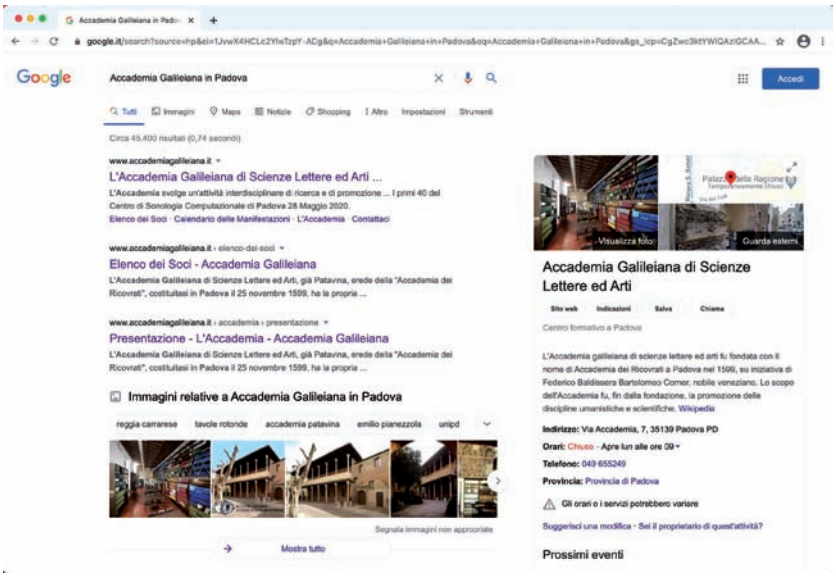

Fig. 2 - Pagina Web con le risposte alla domanda 'Accademia Galileiana in Padova'.

identificato attraverso un indirizzo Web che è unico a livello mondiale; questo indirizzo è denominato *Uniform Resource Locator*, in forma abbreviata URL.[4] Grazie a questo indirizzo unico ogni documento Web può essere identificato univocamente. Poi ogni documento Web può essere collegato ad altri documenti Web, perché contengono informazioni che arricchiscono o specificano le informazioni che il documento contiene, ma anch'essi sono tutti identificati univocamente da un indirizzo Web unico. Quindi ogni documento mantiene la sua unicità e la possibilità di essere individuato senza la possibilità che venga confuso con altri documenti.

I documenti collegati fra loro costituiscono un ipertesto che è consultabile utilizzando gli indirizzi dei documenti; infatti l'insieme dei documenti collegati fra loro costituisce il Web che è una sorta di rete, e da qui il nome Web – rete o ragnatela in inglese – che idealmente può essere percorsa grazie agli indirizzi che permettono di realizzare un collegamento ipertestuale – *link* – fra un documento e l'altro.

L'utente interessato a consultare i documenti presenti nel Web lo può fare utilizzando una applicazione software chiamata *browser Web*. Diverse aziende informatiche hanno sviluppato dei browser Web, alcuni fra quelli più utilizzati dagli utenti sono: Firefox,[5] Internet Explorer,[6] Google Chrome,[7] e Safari.[8]

I documenti Web vengono preparati da privati e da diverse organizzazioni che, per renderli disponibili, utilizzano un'applicazione software denominata *server Web*, che deve essere funzionante su un calcolatore – di tipo *server* – immerso nella rete di trasmissione dati denominata Internet, che è la rete che interconnette milioni di calcolatori di tipo server a livello mondiale. Sui server, interconnessi grazie ad Internet, funzionano diverse applicazioni software e fra queste l'applicazione server Web è una delle più diffuse e utilizzate. Per gli utenti le caratteristiche della rete Internet e dei server Web sono completamente 'trasparenti', cioè non è necessario che gli utenti conoscano le loro caratteristiche, perché essi utilizzano l'applicazione Web tramite un

(4) Il lettore interessato può trovare specifiche tecniche sulla struttura di un indirizzo Web o URL, all'indirizzo: https://url.spec.whatwg.org

(5) Mozilla Firefox, indirizzo Web dell'azienda: https://www.mozilla.org/it/

(6) Internet Explorer, informazioni all'indirizzo Web: https://support.microsoft.com/it-it/office/guida-di-internet-explorer-23360e49-9cd3-4dda-ba52-705336cc0de2?ui=it-IT&rs=it-IT&ad=IT

(7) Google Chrome, informazioni all'indirizzo Web: https://www.google.it/chrome/

(8) Safari, informazioni all'indirizzo Web: https://www.apple.com/it/safari/

browser Web e con questo possono consultare i documenti di loro interesse e navigare nella rete ipertestuale costituita dai documenti Web.

I documenti disponibili nell'ipertesto Web, chiamati pagine Web, sono molto numerosi, infatti si stima che ormai siano diversi miliardi e quindi il Web costituisce un unico punto di accesso a miliardi di documenti o pagine Web. L'ipertesto Web è costituito dall'unione di tutti gli ipertesti che i singoli privati e le singole organizzazioni hanno preparato e poi pubblicato nel Web. Ognuno di questi ipertesti prende il nome di sito Web e ha una pagina che costituisce una sorta di entrata per la sua consultazione e/o navigazione; questa pagina prende il nome di *home page*, *home* in breve. Quindi, per consultare il sito Web di una specifica organizzazione occorre conoscere l'indirizzo Web della sua home page. Ad esempio, per consultare il sito Web dell'Accademia Galileiana occorre conoscere l'indirizzo: https://www.accademiagalileiana.it

Normalmente l'indirizzo Web di una organizzazione richiama il nome o l'obiettivo della organizzazione, ma è un indirizzo costituito da diversi caratteri e non sempre è facile da ricordare. Di molti siti Web non si conosce l'indirizzo, o non si conosce neppure l'esistenza; come si fa a consultare pagine con contenuti di interesse ma delle quali non si conosce l'indirizzo o non se ne conosce l'esistenza? È proprio per venire incontro anche a questo tipo di esigenza informativa che entrano in gioco i motori di ricerca. Anche per utilizzare un motore di ricerca è necessario conoscere il suo indirizzo Web, ma i browser spesso aiutano l'utente, perché spesso vengono inizializzati e impostati nel computer in modo che l'utente trovi già impostato l'accesso ad un motore di ricerca. Di solito l'indirizzo impostato rinvia ad un motore di ricerca di tipo generalista, cioè un motore di ricerca che aiuta l'utente a ritrovare documenti di tipo anche molto diverso. Esistono poi motori di ricerca specialistici che sono motori di ricerca che permettono la consultazione di documenti di contenuto affine, ad esempio di tipo medico. Di solito, però, questi motori di ricerca vengono utilizzati da un utente già più esperto, che naviga nel Web con un motore generalista e lo utilizza per individuare raccolte specialistiche di documenti.

Alcuni esempi di motori di ricerca Web generalisti

Il motore di ricerca di tipo generalista probabilmente più conosciuto e utilizzato è Google Search, o semplicemente Google, che si stima risponda ogni giorno ad alcuni miliardi di domande effettuate da utenti di diverse nazioni. Google raccoglie informazioni su pagine

Web distribuite praticamente in tutto il mondo, pagine Web scritte quindi in diverse lingue e, a seconda della nazione dalla quale l'utente lo utilizza, la risposta che il motore di ricerca fornisce è diversa e l'interfaccia con l'utente è in inglese oppure nella lingua della nazione dalla quale l'utente effettua la domanda.

Un altro motore di ricerca di tipo generalista che opera in modo simile a Google è Microsoft Bing, in breve Bing, proprietà della Microsoft.[9] Una caratteristica di questo motore di ricerca è che ogni giorno presenta una diversa immagine di sfondo.

Un motore di ricerca generalista, ma orientato a soddisfare un'utenza che opera in Cina e che è interessata anche a ritrovare documenti scritti in cinese è Baidu sviluppato dalla società omonima.[10] In questo caso il motore di ricerca è generalista ma interagisce con l'utente direttamente in cinese e ha anche l'obiettivo di fornire all'utente il maggior numero di pagine Web scritte in cinese.

Anche Yandex è un motore di ricerca generalista ma, sviluppato in Russia dalla società omonima,[11] gestisce in modo particolarmente efficace documenti scritti in lingua russa e può essere considerato il corrispondente di Baidu per l'utenza russa.

La figura 3 riporta l'indirizzo Web, il logo e la schermata della home page dei quattro motori di ricerca generalisti che sono stati qui brevemente presentati a titolo di esempio.

Scopo di un motore di ricerca

Si è già detto che lo scopo di un motore di ricerca è quello di rispondere ad una richiesta dell'utente che lo sta utilizzando, fornendogli dei documenti che contengono informazioni pertinenti alla domanda posta; una schematizzazione di come questo scopo viene raggiunto dal motore di ricerca è fornita in figura 4.

Di motori di ricerca ne esistono diversi tipi e quelli che operano nel Web raccolgono e gestiscono documenti disponibili nel Web; pertanto i documenti che un motore di ricerca Web può fornire all'utente

---

([9]) Sito Web della Microsoft: https://www.microsoft.com/

([10]) L'indirizzo Web della società rinvia direttamente alla pagina del motore di ricerca: https://www.baidu.com

([11]) Anche l'indirizzo Web di Yandex indirizza direttamente al motore di ricerca: https://yandex.com

Fig. 3 - Alcuni esempi di motori di ricerca generalisti; di ciascuno viene riportato l'indirizzo Web, il logo e la schermata della home page.

Fig. 4 - Schematizzazione dello scopo e della componente del motore di ricerca Web che interagisce direttamente con l'utente attraverso un browser.

sono sempre e solo documenti che sono stati pubblicati e che sono presenti nel Web.

Una prima domanda che l'utente si può porre riguarda proprio la composizione e la dimensione della collezione di documenti che il motore di ricerca gestisce. Un utente che si reca in una biblioteca fisica sa che la collezione di documenti che può esplorare e utilizzare è quella costituita dai documenti fisicamente presenti nelle sale della biblioteca. L'utente sa anche che la collezione è stata costituita e organizzata nel tempo dal lavoro dei bibliotecari che lavorano nella biblioteca e dagli studiosi che hanno suggerito l'acquisto delle diverse opere o che hanno donato delle opere alla biblioteca. Quindi l'utente di una biblioteca fisica, e così pure l'utente di un archivio o di qualsiasi altra raccolta di 'documenti' fisici, sa che la sua ricerca di informazioni è limitata alle informazioni contenute nei documenti effettivamente presenti all'interno della biblioteca. Ma un utente che utilizza un motore di ricerca non sa che documenti contiene effettivamente la collezione di documenti dalla quale il motore di ricerca attinge per fornire risposta alle sue domande. L'utente non sa neppure da chi e in che modo sia stata costituita la collezione dei documenti Web e come essa venga arricchita e aggiornata nel tempo. Quindi l'utente non conosce le dimensioni e le caratteristiche della collezione, e non sa se documenti di suo interesse, pur esistendo effettivamente nel Web, non sono presenti nella collezione che il motore di ricerca ha a disposizione per la sua operatività e per tale motivo non riesca ad ottenerli in risposta, oppure non riesca ad ottenerli in risposta, anche se presenti nella collezione, perché il motore di ricerca non è in grado di 'capire' appieno la sua domanda e quindi non è in grado di individuare dei documenti utili e presenti nella collezione gestita.

Una seconda domanda importante che l'utente si può porre riguarda le operazioni che devono essere fatte per rendere la collezione di documenti fruibile all'utente alla velocità con la quale il motore risponde alle domande che gli vengono poste. Infatti l'utente si può domandare come faccia un motore di ricerca a rispondere e come lo riesca a fare così velocemente, rispondendo con elenchi costituiti spesso da decine di migliaia, ma a volte anche da milioni di riferimenti a documenti di interesse.

Le due sezioni successive sono dedicate a fornire risposta a queste due domande.

La costruzione della collezione di documenti gestita da un motore di ricerca Web

Si è già specificato che dal punto di vista informatico un motore di ricerca è un'applicazione software, quindi è un insieme di programmi scritti in un linguaggio di programmazione idoneo a istruire un calcolatore elettronico a svolgere operazioni di ricerca e recupero di informazioni e/o documenti digitali disponibili nel Web. Guardando ora più da vicino la struttura di un motore di ricerca, questo insieme di programmi è in effetti strutturato in due parti o componenti principali: una prima componente è quella con la quale l'utente interagisce direttamente attraverso un browser per porre le sue domande ed ottenere informazioni e/o documenti in risposta – ed è quella rappresentata schematicamente in figura 4 – e una seconda componente, della cui esistenza l'utente normalmente non è consapevole, che è quella deputata alla costruzione della collezione di documenti Web che la componente del motore di ricerca che interagisce con l'utente utilizza per fornire risposte all'utente. Quindi, se si guarda la struttura – o anatomia – di un motore di ricerca più da vicino, essa è costituto da due insiemi di programmi: uno che, dal punto di vista temporale, opera preliminarmente e costruisce la collezione di documenti Web, e un altro che utilizza la collezione di documenti Web per fornire risposte agli utenti.

L'insieme di programmi che fornisce le risposte agli utenti è il motore di ricerca propriamente detto; l'altro insieme di programmi costituisce uno strumento software che è programmato per comportarsi come un 'vagabondo' del Web, cioè uno strumento che, utilizzando i collegamenti esistenti fra i documenti dell'ipertesto che costituisce il Web, è istruito per vagabondare nel Web, esplorarne le pagine e raccogliere e collezionare quelle più utili e coerenti con le istruzioni ricevute. Le pagine, man mano che vengono raccolte, formano la collezione dei documenti che così viene costruita per poi essere utilizzata dalla componente del motore di ricerca che elabora e risponde alle domande dell'utente.

Il vagabondo è programmato per operare in maniera simile ad un utente che naviga nel Web con un browser per individuare delle pagine con contenuti di interesse; però questo programma opera su vasta scala e con una velocità molto maggiore di quella che un qualsiasi utente può raggiungere. Il vagabondo segue i collegamenti da pagina a pagina e, quando individua una pagina di interesse, la scarica e la inserisce nella collezione; poi estrae i collegamenti presenti nella pagina, e itera l'operazione con i nuovi collegamenti andando a visitare

le pagine collegate e procedendo nella individuazione e acquisizione di nuove pagine di interesse. Ciascuna di queste operazioni viene eseguita in breve tempo, molto più velocemente di quanto un qualsiasi utente potrebbe fare.

Il vagabondo è uno strumento software che fa parte dell'insieme di strumenti informatici denominati 'agenti software' quindi, in questo caso, agenti software Web, perché sono programmati per operare specificatamente nel Web; molti sono i nomi con i quali questi strumenti vengono chiamati in inglese; alcuni di questi sono anche suggestivi, perché ne richiamano qualche specifica caratteristica operativa, come ad esempio: *Web wanderer* – un vagabondo del Web già sopra descritto; *Web crawler* – uno strumento software che si comporta come una cosa che striscia o si muove lentamente come un insetto, strisciando da pagina a pagina; *spider* - uno strumento software che agisce come un ragno che si muove sulla ragnatela Web; *robot* – uno strumento che fa delle azioni simili a quelle che fa un essere umano, quindi in grado di replicare automaticamente le funzioni che le persone fanno quando navigano nel Web; *bot* (forma abbreviata di robot) – un programma che si comporta in modo simile ad un robot.

A seconda delle informazioni che si desidera avere a disposizione nei documenti della collezione, quindi a seconda delle domande che ci si aspetta che gli utenti siano interessati a porre al motore di ricerca e alle quali si vuole che il motore di ricerca risponda in modo coerente e utile, si programma il vagabondo in modo tale che, seguendo i collegamenti fra le pagine Web, sia in grado di scegliere e raccogliere quelle più coerenti. Ad esempio, se si vuole che il vagabondo costruisca una collezione di pagine che trattano i diversi argomenti che sono trattati nelle pagine pubblicate nel Web, si deve costruire una collezione di tipo 'generalista', programmando il vagabondo perché raccolga tutte le pagine che man mano visita. Se, invece, si vuole costruire una collezione specialistica, quindi una collezione di documenti che trattano di un argomento specifico, che potrebbe essere uno specifico argomento medico quale la malattia da coronavirus COVID-19, il vagabondo sarà programmato per raccogliere solo documenti che trattano di questo specifico argomento.

Per fare in modo che il vagabondo sia in grado di iniziare a esplorare il Web e raccogliere pagine di interesse, all'avvio gli viene fornito un elenco di indirizzi di siti Web che contengono sicuramente pagine coerenti con i contenuti della collezione che si vuole costruire. Se, ad esempio, si volesse costruire una collezione specialistica con le informazioni dei diversi corsi di studio attivati nelle università italiane, l'elenco sarebbe costituito dagli indirizzi dei siti Web delle universi-

tà italiane e il vagabondo inizierebbe il suo lavoro di esplorazione e raccolta proprio da questi indirizzi. Gli indirizzi contenuti in questo elenco sono chiamati indirizzi *seed*, perché costituiscono i semi che permettono l'avvio del processo di nascita e costruzione della collezione di interesse.

Per costruire la collezione di documenti di una biblioteca o di un archivio di solito occorrono decenni o anche secoli; a questo punto ci si potrebbe chiedere quanto tempo impiega un vagabondo Web a raccogliere e costruire la collezione dei documenti da fornire al motore di ricerca per la sua operatività. Per rispondere a questa domanda sarebbe utile conoscere quante pagine sono presenti nel Web perché, anche ad uno strumento software veloce come un vagabondo occorre il tempo necessario ad individuare la pagina di interesse, scaricarla e inserirla nella collezione. Quindi, conoscendo il tempo necessario per effettuare queste operazioni e poi il numero delle pagine per le quali è necessario effettuarle, si potrebbe conoscere il tempo complessivo necessario a costruire la collezione. Ma non è di fatto possibile conoscere la reale dimensione del Web, anche se vari studiosi hanno proposto dei metodi per effettuare il calcolo delle pagine Web che già qualche anno fa erano stimate in decine di miliardi; ma anche se il numero di pagine oggi esistenti potesse essere esattamente determinato, quel numero sarebbe immediatamente superato, perché l'ipertesto Web è molto dinamico e in continua evoluzione, dato che le pagine che lo compongono vengono costantemente aggiornate, aggiunte e/o rimosse.

Un altro aspetto da tenere presente è che, anche se si conoscesse il numero delle pagine Web, essendo questo numero molto elevato, nessun motore di ricerca potrebbe avere a disposizione gli strumenti fisici necessari ad archiviare così tanti documenti. Quindi, chi programma un vagabondo adotta usualmente qualche criterio per limitare la dimensione della raccolta dei documenti da costruire e, comunque, la collezione risulta in ogni caso ampia. Per costruirla è talora utile mettere più programmi vagabondo ad operare in parallelo su porzioni diverse dell'ipertesto Web; con questi e anche altri accorgimenti specifici la collezione può essere costruita in qualche settimana di lavoro. Quando la collezione è stata costruita, prima di essere resa disponibile al motore di ricerca, deve essere ripulita, ad esempio da eventuali pagine duplicate (spesso siti diversi contengono pagine che sono uguali perché vengono copiate da sito a sito) e da pagine che contengono contenuti offensivi o inattendibili. Dopo tutte queste attività la collezione viene finalmente resa utilizzabile all'utenza, ma il lavoro che il vagabondo o l'insieme di vagabondi fa, praticamente non finisce mai, perché le pagine della collezione devono anche essere mantenute ag-

giornate: quelle che sono state modificate dai loro autori devono essere sostituite con le loro nuove versioni e devono essere raccolte pagine nuove che vengono via via pubblicate su argomenti pertinenti; tutte queste attività di aggiornamento della collezione solitamente vengono fatte senza che l'utente del motore di ricerca se ne accorga e sono fatte a sua insaputa, ma a suo beneficio, perché, in questo modo, la collezione dei documenti viene mantenuta costantemente aggiornata con informazioni utili per essere fruite dall'utente.

Come fa un motore di ricerca a rispondere velocemente

Una volta che la collezione di documenti è stata preparata, essa viene fornita al motore di ricerca propriamente detto, che deve operare per preparare la collezione in modo tale da rispondere velocemente alle domande dell'utente.

I documenti della collezione Web, raccolti dal vagabondo, vengono forniti al motore di ricerca senza che siano stati organizzati in un modo appropriato per poi essere utilizzati per rispondere in modo efficace ed efficiente alle domande dell'utente. Quindi ci si trova ad avere una vasta collezione di documenti, che però non è organizzata e non è predisposta per la ricerca. Riprendendo l'analogia con la collezione dei documenti fisici che viene gestita in una biblioteca, ci si trova ad avere a disposizione una raccolta di documenti che non sono organizzati per autore, per soggetto o in base a qualche altra classificazione e organizzazione che rispecchia gli argomenti trattati nei documenti; occorre quindi inventare preliminarmente o, se disponibili, adottare dei metodi di rappresentazione del contenuto informativo dei documenti che permetta poi di organizzarli e gestirli per renderli fruibili all'utente. Quindi occorre avere a disposizione dei metodi di rappresentazione della conoscenza raccolta nei documenti corrispondenti a quelli manuali che sono stati studiati e sviluppati nei secoli dalla biblioteconomia e dall'archivistica, per renderli disponibili ai bibliotecari e agli archivisti per il loro lavoro di catalogazione, anche semantica, del materiale gestito nelle biblioteche e negli archivi. E nella collezione di documenti Web, come nelle collezioni delle biblioteche e degli archivi, sono disponibili documenti sostanzialmente testuali, nonché documenti testuali corredati da immagini, documenti sonori e anche documenti video. Quindi la collezione Web è per sua natura una collezione multimediale e questo fatto ha di conseguenza reso necessario mettere a disposizione metodi di rappresentazione

del contenuto informativo specifico per ciascun medium informativo. Dei metodi utili sono stati inventati e sviluppati negli anni, ma ancora oggi la ricerca di metodi sempre migliori ed efficaci è molto attiva e metodi e strumenti via via più efficaci vengono proposti. Nel seguito si concentra l'attenzione soltanto su alcuni dei metodi più efficaci e consolidati che sono stati inventati e sviluppati per la rappresentazione del contenuto informativo dei documenti testuali.

I metodi che sono stati sviluppati per gli altri media non vengono qui trattati sia per motivi di spazio ma anche perché essi sono solo in parte specifici per ogni singolo medium, poiché tutti utilizzano anche metodi di rappresentazione testuale, facendo sempre uso della rappresentazione dei contenuti informativi dei testi descrittivi di cui sono corredate, ad esempio, le immagini, le registrazioni sonore e quelle video.

Analisi e rappresentazione automatica del testo con approccio statistico

Per analizzare e rappresentare il contenuto informativo dei testi, è stato necessario inventare e sviluppare nuovi strumenti metodologici da utilizzare per fare l'analisi e la rappresentazione dei contenuti informativi di un testo in modo automatico, perché la quantità di documenti che compone una collezione di documenti Web è talmente elevata da rendere improponibile pensare di poter effettuare le operazioni di rappresentazione in modo manuale. Negli anni sono stati proposti metodi diversi, ma quelli che fino ad ora si sono dimostrati più efficaci per rendere operativo un motore di ricerca sono quelli che si basano su approcci di tipo statistico, che studiano le distribuzioni delle frequenze dei caratteri e delle parole presenti nei documenti testuali.

I metodi non esistevano già e si sono dovuti inventare appositamente, però, come sempre accade in tutti i settori scientifici, ci si è potuti basare anche su risultati già raggiunti da altri studiosi. Si sono dimostrati particolarmente utili i risultati che aveva raggiunto George Kingsley Zipf[12] a seguito dei suoi studi di linguistica e di filologia, che hanno fornito le prime formulazioni e dimostrazioni di varie leggi em-

([12]) Una presentazione degli argomenti affrontati da George Kingsley Zipf (1902-1950) si può trovare all'indirizzo Web https://www.treccani.it/enciclopedia/george-kingsley-zipf/ e un'altra presentazione, corredata anche da note bibliografiche e riferimenti, si trova all'indirizzo Web https://en.wikipedia.org/wiki/George_Kingsley_Zipf

piriche che mettono in rapporto la frequenza di una parola con la sua forma e il suo significato, o contenuto informativo. Gli studi di Zipf evidenziano l'equilibrio o 'economia' con cui sono costruiti i testi, composti di pochi tipi di parole d'altissima ricorrenza e molti tipi di parole rare. Se si osserva la distribuzione di frequenza delle parole in un testo, la loro distribuzione non è uniforme, perché poche parole sono molto frequenti e molte altre sono poco frequenti o rare; quindi la distribuzione delle parole messe in ordine decrescente di frequenza è asimmetrica.

Lo studio della distribuzione delle parole nei testi ha consentito a Zipf di formulare quella che oggi è chiamata 'Legge di Zipf' e che afferma: dato un ampio campione di testi e calcolata la frequenza *f* delle parole, una volta che le parole sono state ordinate in maniera decrescente di frequenza, cioè in base al rango *r*, la distribuzione che si ottiene ha un andamento assimilabile ad una iperbole. In particolare si ha che *r* x *f* è uguale ad una costante e quindi, se la costante è denominata *k*, la distribuzione che se ne ottiene è $f = k/r$, rappresentata in figura 5.

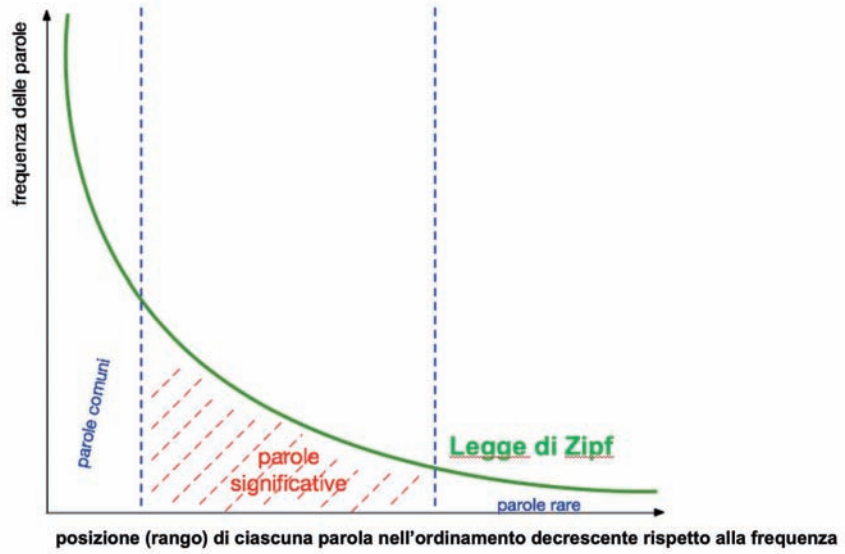

Fig. 5 - Legge di George Kingsley Zipf.

La legge di Zipf è importante per la generalità della sua struttura statistica, perché l'ordine delle parole può cambiare a seconda dei documenti testuali che compongono la collezione di documenti in esame e il valore che può assumere la costante *k* può variare a seconda della lingua in cui sono scritti i documenti, ma l'andamento della distribuzione rimane sempre lo stesso.

Zipf ha condotto i suoi studi prima del 1950, anno in cui è morto, prima che i calcolatori cominciassero ad essere resi disponibili, e quindi non ha conosciuto l'informatica moderna e ha condotto i suoi studi in modo molto rigoroso ma solamente manuale; nonostante questo, Zipf è riuscito a dimostrare che le parole che vengono utilizzate in collezioni ampie di documenti testuali assumono delle distribuzioni statistiche particolari in quanto le parole comuni sono molto frequenti, ma trasmettono un significato estremamente limitato, mentre le parole significative per la trasmissione di contenuti sono quelle che si trovano al centro della distribuzione, come viene illustrato nel grafico di figura 5. Questi risultati erano stati pubblicati in lavori di linguistica e di statistica della lingua, di conseguenza in un qualche modo è stato necessario riscoprirli e avere la capacità di intuire che potevano essere utilmente applicati anche per la rappresentazione automatica del contenuto informativo di documenti testuali.

È il lavoro dello studioso Hans Peter Luhn[13] che, utilizzando i risultati di Zipf verso la fine degli anni '50 dello scorso secolo, indica la strada per effettuare un'analisi automatica dei documenti testuali utile alla rappresentazione del loro contenuto informativo. Il contributo significativo di Luhn può essere così sintetizzato: le frequenze, con le quali le parole appaiono in un testo, possono essere utilizzate per selezionare le parole utili a rappresentare il contenuto di un documento testuale; quindi l'ipotesi di partenza di Luhn è che i dati di frequenza possano essere utilizzati per estrarre da un documento testuale le parole utili a rappresentarne il contenuto informativo. Poi Luhn specifica due valori limite, uno superiore e uno inferiore (rappresentati in figura 5 dalle due linee blu tratteggiate) per escludere le parole comuni e quelle rare, che sono entrambe poco capaci di rappresentare il contenuto informativo di un documento. Si concentra, invece, sulle parole presenti nella parte centrale della distribuzione, che hanno un più elevato 'potere risolutivo' – *resolving power* – quindi sono parole capaci

(13) Una biografia di Hans Peter Luhn (1896-1964), corredata da note bibliografiche e riferimenti, si può trovare all'indirizzo Web https://en.wikipedia.org/wiki/Hans_Peter_Luhn.

di discriminare il contenuto informativo dei documenti. La curva di distribuzione empirica del potere risolutivo è rappresentata in figura 6 con una linea tratteggiata nera. La determinazione dei due valori limite comporta una certa arbitrarietà, perché devono essere stabiliti in modo empirico, quindi mediante tentativi e individuazione di errori. È importante sottolineare come queste idee siano state e siano ancor oggi alla base di gran parte del lavoro relativo alla rappresentazione del contenuto informativo dei documenti testuali nei motori di ricerca.

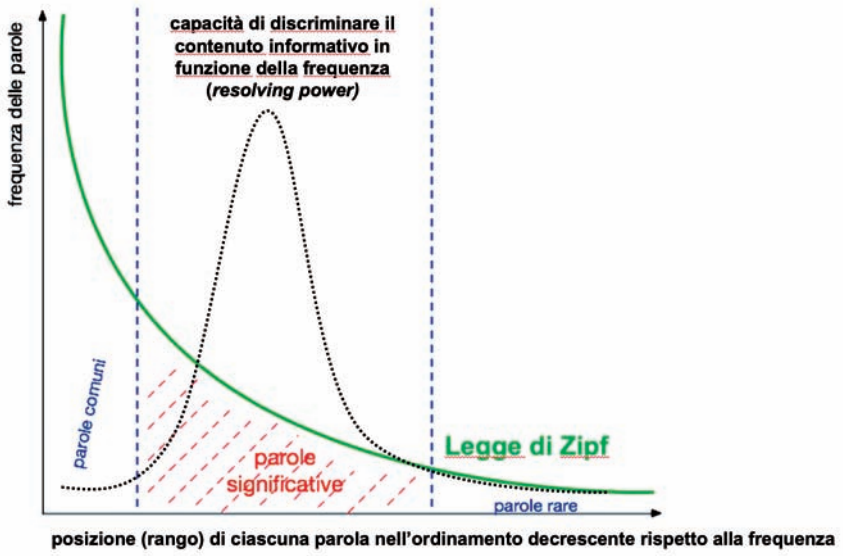

Fig. 6 - *Resolving power*: capacità di discriminare il contenuto informativo di una parola in funzione della sua frequenza.

Costruzione degli indici e risposta alla domanda dell'utente

A seguito del lavoro di rappresentazione del contenuto informativo dei documenti della collezione di interesse, si prepara un elenco delle parole significative che sono state estratte dai documenti; per ogni parola inserita nell'elenco si mantiene memoria dei numeri identificativi degli specifici documenti all'interno dei quali la parola è stata

utilizzata, in modo tale da ricordare i documenti per i quali questa parola svolge un ruolo di rappresentazione del contenuto informativo.

Le singole parole vengono inserite all'interno dell'elenco in ordine lessicografico, in modo tale da poter poi utilizzare, nella fase di preparazione della risposta ad una specifica domanda dell'utente, dei metodi di calcolo molto efficienti nel ritrovare le specifiche parole che un utente ha utilizzato nella frase di ricerca. Infatti occorre tenere presente che un motore di ricerca deve consultare elenchi costituiti da centinaia di migliaia di parole che sono state estratte dai miliardi di documenti che costituiscono la collezione di interesse.

Una volta che le parole della frase di ricerca sono state individuate nell'elenco, ad esse sono associati i numeri identificativi dei documenti che le contengono. A questo punto il motore di ricerca utilizza dei metodi di calcolo utili a preparare un elenco dei documenti, da dare in risposta alla domanda dell'utente, ordinato in base ad una valutazione della potenziale rilevanza di ciascun documento alla domanda dell'utente. Sono proprio questi metodi di calcolo, associati ai metodi di rappresentazione dei documenti utilizzati, che distinguono il modo di operare di uno specifico motore di ricerca rispetto ad un altro.

L'elenco delle parole costituisce un indice che svolge una funzione analoga a quella dell'indice analitico di un libro che rinvia alle parti di un testo che trattano l'argomento individuato dalla parola.

Nei motori di ricerca moderni, nella fase di rappresentazione ed estrazione delle parole significative da un testo, oltre all'elenco delle singole parole, si estraggono spesso le informazioni utili a costruire anche dei termini di interesse composti da due o più parole in modo da poter poi utilizzare dei metodi più sofisticati per rispondere in modo più completo alle domande degli utenti. Quindi, oltre alle parole, gli indici contengono dei termini che rinviano in modo più preciso ad argomenti specifici trattati nei documenti. Gli indici costruiti dal motore di ricerca svolgono, quindi, in modo automatico, una funzione corrispondente a quella che viene svolta dai sistemi di soggettazione e/o classificazione che vengono preparati nelle biblioteche e negli archivi. Gli indici prendono il nome di indici trasposti – *inverted index* – perché permettono, partendo da una parola o da un termine, di risalire ai documenti che contengono informazioni che illustrano il significato della parola o del termine.

L'effettiva funzionalità degli indici, nel supportare il reperimento dell'informazione, può essere apprezzata implicitamente dall'utente che in un tempo molto breve può ricevere, in risposta ad una sua domanda, numerosi documenti; ad esempio, la risposta, riportata in figura 2, alla domanda dell'utente 'Accademia Galileiana in Padova' è

stata fornita in 0,74 secondi rendendo disponibili 45.400 riferimenti a documenti Web che, quando la domanda è stata posta, erano presenti nella collezione di documenti gestita dal motore di ricerca Google.

La nascita del settore del reperimento automatico dell'informazione e dei motori di ricerca in Italia

Gli studiosi sopra citati, hanno operato negli Stati Uniti d'America o in Gran Bretagna. Ora è il momento di domandarsi che cosa è successo nel settore in Italia. Nel nostro Paese c'è stata una attenzione molto precoce alle problematiche del reperimento automatico dell'informazione fin da quando, nel 1961, nasce l'associazione italiana di informatica, denominata allora Associazione Italiana per il Calcolo Automatico (AICA), ma che cambierà successivamente il suo nome in Associazione Italiana per l'Informatica e il Calcolo Automatico; in figura 7 viene riportato l'atto costitutivo insieme all'elenco dei firmatari.

Nel 1962, e dunque nell'anno immediatamente successivo alla sua costituzione, visto che l'informatica si stava articolando in sotto aree di interesse industriale e di ricerca, l'AICA decide di organizzarsi in alcuni gruppi di lavoro e, in particolare, viene fondato il gruppo di lavoro AICA-GLIR, già con il nome inglese Information Retrieval. In figura 8 viene riportata la circolare con la quale tutti i membri dell'AICA vengono informati dell'intenzione di attivare il gruppo di lavoro; successivamente il Consiglio Direttivo dà l'autorizzazione a procedere e poco dopo iniziarono le attività dell'AICA-GLIR.

Alcuni dei fondatori del gruppo di lavoro sono degli studiosi che contribuiranno a sviluppare in Italia il settore dell'IR, ma anche l'informatica in generale. Infatti alcuni di loro sono: Padre Roberto Busa,[14] Eduardo Caianiello,[15] Alfonso Caracciolo,[16] A. Zampoldi –

([14]) Padre Roberto Busa (1913–2011) ha ideato e realizzato l'Index Thomisticus sviluppando e utilizzando per primo alcuni metodi automatici di analisi del lessico, nel suo caso del lessico di Tommaso d'Aquino. Una biografia di Padre Busa è consultabile all'indirizzo Web https://it.wikipedia.org/wiki/Roberto_Busa

([15]) Eduardo Renato Caianiello (1921–1993) è stato uno studioso centrale nello sviluppo della cibernetica e dell'informatica come illustrato in dettaglio all'indirizzo Web https://it.wikipedia.org/wiki/Eduardo_Caianiello

([16]) Di Alfonso Caracciolo alcune informazioni sono disponibili agli indirizzi Web http://www.cep.cnr.it/storia03.html e https://it.wikipedia.org/wiki/Associazione_italiana_per_l%27informatica_ed_il_calcolo_automatico

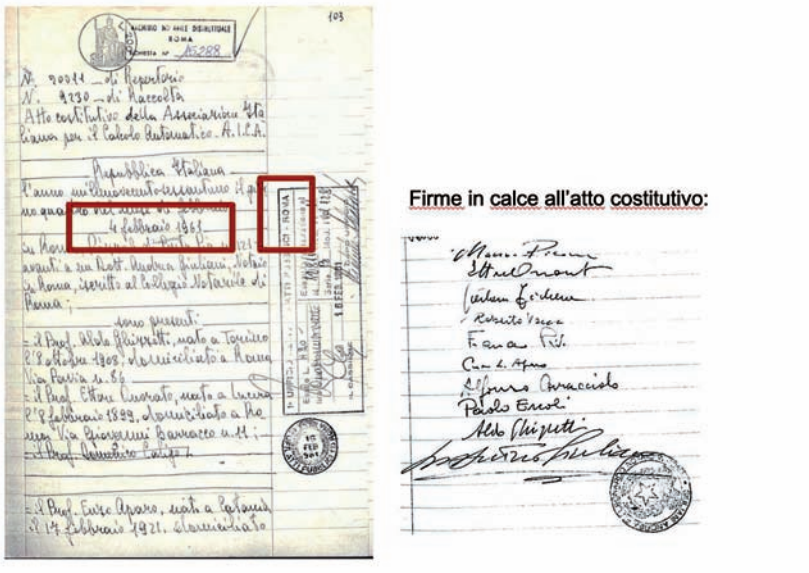

Fig. 7 - 4 febbraio 1961: nasce AICA, l'Associazione italiana di informatica.

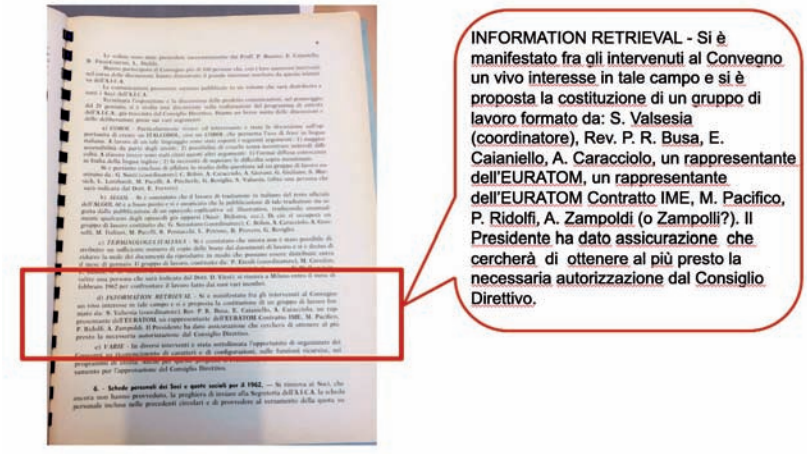

Fig. 8 - Circolare n.7 dell'AICA del 12 febbraio 1962 inviata a tutti i membri della associazione con la proposta di lanciare il Gruppo di lavoro in Information Retrieval (AICA-GLIR).

nome riportato con un refuso, perché probabilmente lo studioso è Antonio Zampolli,[17] uno dei fondatori della linguistica computazionale in Italia. Quindi vediamo che già dalla nascita dell'AICA-GLIR c'è un'attenzione multidisciplinare, perché chi opera nel settore del reperimento dell'informazione non può operare in maniera isolata rispetto ad altri settori, dovendo padroneggiare le conoscenze di altre discipline che concorrono alla preparazione, rappresentazione e gestione dei documenti digitali; fra queste discipline, per citarne solo alcune, vi sono la statistica, il calcolo delle probabilità, la matematica, e le discipline che si occupano della rappresentazione dell'informazione e della conoscenza. Quindi l'IR è un'area interdisciplinare che richiede una sistematica collaborazione fra aree diverse.

All'inizio dello sviluppo dell'IR in Italia c'è stato un interesse sia accademico che industriale ma poi, per qualche tempo, l'interesse e le attività hanno riguardato maggiormente il comparto industriale, forse anche perché, all'inizio degli anni '70, erano stati resi disponibili, da parte di diverse aziende di informatica, i primi sistemi software industriali di reperimento dell'informazione; fra questi il sistema STAIRS[18] dell'IBM e il sistema FIND della Sperry Univac che erano anche stati adottati per la realizzazione di applicazioni di reperimento dell'informazione funzionanti in importanti organizzazioni a carattere industriale, ad esempio all'interno del centro di documentazione della FIAT di Torino. Invece all'interno del mondo universitario e della ricerca l'interesse si affievolisce, probabilmente anche per la complessità dei problemi che quest'area disciplinare e interdisciplinare costringe ad affrontare e per la difficoltà di collocare le relative ricerche in un ben definito Settore Scientifico Disciplinare di interesse accademico.

Dal 1984 la collaborazione industria/università riprende e già nel 1986 il gruppo di lavoro AICA-GLIR collabora attivamente alla realizzazione della prima edizione del congresso internazionale più importante del settore; infatti la nona edizione dell'*International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1986)* viene organizzata in settembre a Pisa. Nel 1987 viene fondato in Italia, a Padova, il primo gruppo di ricerca che si

---

([17]) Antonio Zampolli (1937-2003) uno degli studiosi che ha contribuito a fondare la linguistica computazionale, alcune informazioni sono disponibili all'indirizzo Web https://it.wikipedia.org/wiki/Antonio_Zampolli

([18]) Alcune informazioni sul sistema *Storage and Information Retrieval System* – STAIRS dell'IBM sono disponibili all'indirizzo Web https://en.wikipedia.org/wiki/IBM_STAIRS.

occupa a livello scientifico delle tematiche del settore; il gruppo viene fondato contestualmente al Dipartimento di Elettronica e Informatica dell'Università degli Studi di Padova e prende il nome di *Information Management Systems (IMS) research group*.[19] Il gruppo IMS da allora è sempre stato attivo e da allora opera a livello nazionale e internazionale; negli anni ha ottenuto risultati di ricerca di rilievo anche per l'ideazione e lo sviluppo di motori di ricerca. In anni più recenti sono stati attivati in Italia anche altri gruppi di ricerca che operano a Pisa, a Milano, a Roma e a Bari. Dunque quest'area disciplinare è stata avviata come area di ricerca a livello nazionale nel 1987 nell'allora Dipartimento di Elettronica e Informatica (DEI),[20] ma successivamente si è sviluppata e i diversi gruppi di ricerca presenti a livello nazionale hanno ora una meritata visibilità anche a livello internazionale.

Arte dei motori di ricerca

Ora occorre affrontare un ultimo aspetto che viene richiamato nel titolo della memoria dalla parola 'arte'. La parola arte viene introdotta in informatica dal grandissimo studioso Donald Ervin Knuth[21] che, giovanissimo, negli anni '60 dello scorso secolo comincia a scrivere una enciclopedia dell'informatica; sono anni in cui si stava fondando la disciplina e stava esplodendo l'interesse per l'informatica. Infatti gli anni '60 sono molto importanti, ma non ci sono ancora volumi o trattati di riferimento perché la disciplina si sta sistematizzando solo allora. La disciplina trae beneficio da studi precedentemente sviluppati in tante altre aree, quali la matematica e la fisica, e dai risultati della teoria della computabilità definita da Alan Mathison Turing[22] negli anni '30. Non esistono ancora trattati sui fondamenti dell'informatica e allora Knuth comincia a scrivere la sua enciclopedia di informatica e nel 1968 ne pubblica il primo volume, usando la parola arte nella

---

([19]) Indirizzo del sito Web del gruppo IMS http://ims.dei.unipd.it/

([20]) Nel 2002 il Dipartimento di Elettronica e Informatica ha cambiato la sua denominazione in Dipartimento di Ingegneria dell'Informazione, mantenendo l'acronimo identificativo DEI; il sito Web del dipartimento è raggiungibile all'indirizzo: https://www.dei.unipd.it/

([21]) Una biografia di Donald Ervin Knuth (1938-) può essere consultata all'indirizzo Web https://en.wikipedia.org/wiki/Donald_Knuth

([22]) Una bibliografia di Alan Mathison Turing (1912-1954) può essere consultata all'indirizzo Web https://en.wikipedia.org/wiki/Alan_Turing

denominazione: '*The Art of Computer Programming*'. Knuth possiede uno stile di scrittura molto rigoroso e asciutto e spiega molto sinteticamente e solo implicitamente il perché dell'uso della parola *Art* nella prefazione del primo volume:

The process of preparing programs for a digital computer is especially attractive, not only because it can be economically and scientifically rewarding, but also because it can be an aesthetic experience much like composing poetry or music.

Knuth assimila, quindi, il processo di preparazione di programmi per un calcolatore ad un'esperienza artistica quali sono la composizione di poesie e musica.

Per preparare un programma prima di tutto occorre immaginare le azioni e le attività che si vuole che il programma realizzi al nostro posto, in modo simile allo scultore che davanti ad un nuovo blocco di marmo immagina quale sarà la statua che riuscirà a far emergere dal blocco.

Poi, dalla lettura dell'intera prefazione, si coglie anche la fierezza dello studioso che fa ogni sforzo per raggiungere una competenza nell'arte della programmazione di altissimo livello sia teorico che sperimentale.

Analogamente, per rendere oggi disponibili i motori di ricerca è stato necessario prima di tutto immaginare uno strumento che non esisteva, ma che allora si è pensato di poter ideare e realizzare per ottenere un aiuto nella ricerca di informazioni utili tra enormi moli di documenti digitali. E i motori di ricerca basano la loro operatività sulla preparazione di programmi scritti a regola d'arte.

Ringraziamenti

Ringrazio la signora Emanuela Scalzotto, della segreteria della Associazione italiana per l'informatica e il calcolo automatico (AICA), per il suo aiuto nella ricerca dei documenti d'archivio di interesse per la storia del reperimento automatico dell'informazione e dei motori di ricerca in Italia. Documenti che, purtoppo, non sono oggi reperibili attraverso un motore di ricerca.

I would like to thank Ms Emanuela Scalzotto, from the secretariat of the Italian Computer Science Society (AICA), for her help in finding the archive documents of interest for the history of modern automatic information retrieval and search engines in Italy. Documents that, unfortunately, are not available today through a search engine.

BIBLIOGRAFIA

Stefan Büttcher, Charles L.A. Clarke, Gordon V. Cormack. *Information retrieval: implementing and evaluating search engines*. Cambridge, Massachusetts, MIT Press, 2010.

W. Bruce Croft, Donald Metzler, Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Pearson Education, 2010.

Antonio Gulli, Alessio Signorini. The indexable Web is more than 11.5 billion pages. In: Allan Ellis, Tatsuya Hagino (Eds), *Proceedings of the 14th International Conference on World Wide Web (WWW2005)*, Chiba, Japan, May 10-14, 2005 - Special interest tracks and posters, pp. 902-903.

Donald E. Knuth. *Volume 1 / Fundamental algorithms*. Addison-Wesley, Reading, Mass., USA, 1968. Fa parte di: *The art of computer programming*; vol. 1.

Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2, 159-165 (1958).

Cornelis J. "Keith" van Rijsbergen. *Information Retrieval*, 2nd ed. Butterworth, London, 1979 - anche disponibile come documento Web all'indirizzo:

http://www.dcs.gla.ac.uk/Keith/Preface.html