The History of Information Retrieval Research

Mark Sanderson: School of Computer Science and Information Technology, RMIT University, GPO Box 2476, Melbourne 3001 Victoria, Australia. mark.sanderson@rmit.edu.au; +61 3 992 59675

W. Bruce Croft: Department of Computer Science, 140 Governors Drive, Box 9264, University of Massachusetts, Amherst, MA 01003-9264, USA. croft@cs.umass.edu; +1 413 545-0463

Abstract

This paper describes a brief history of the research and development of information retrieval systems starting with the creation of electro-mechanical searching devices, through to the early adoption of computers to search for items that are relevant to a user's query. The advances achieved by information retrieval researchers from the 1950s through to the present day are detailed next, focusing on the process of locating relevant information. The paper closes with speculation on where the future of information retrieval lies.

Keywords

Information Retrieval, History, Ranking Algorithms

Introduction

The long history of information retrieval does not begin with the internet. It is only in the last decade and a half of the IEEE's one hundred years that web search engines have become pervasive and search has become integrated into the fabric of desktop and mobile operating systems. Prior to the broad public day-to-day use of search engines, IR systems were found in commercial and intelligence applications as long ago as the 1960s. The earliest computer-based searching systems were built in the late 1940s and were inspired by pioneering innovation in the first half of the 20th century. As with many computer technologies, the capabilities of retrieval systems grew with increases in processor speed and storage capacity. The development of such systems also reflects a rapid progression away from manual library-based approaches of acquiring, indexing, and searching information to increasingly automated methods.

An information retrieval (IR) system locates information that is relevant to a user's query. An IR system typically searches in collections of unstructured or semi-structured data (e.g. web pages, documents, images, video, etc.). The need for an IR system occurs when a collection reaches a size where traditional cataloguing techniques can no longer cope. Similar to Moore's law of continual processor speed increase, there has been a consistent doubling in digital storage capacity every two years. The number of bits of information packed into a square inch of hard drive surface grew from 2,000 bits in 1956 to 100 billion bits in 2005[1]. With the growth of digitised unstructured information and, via high speed networks, rapid global access to enormous quantities of that information, the only viable solution to finding relevant items from these large text databases was search, and IR systems became ubiquitous.

This brief review of past work focuses on the algorithms that take a user's query and retrieve a set of relevant documents. This paper opens with a review of the early developments of electro-mechanical and computational devices that searched manually generated catalogues. This is followed by a description of how IR moved to automatic indexing of the words in text and how complex Boolean query languages gave way to simple text queries. The automatic techniques and theories that supported them have continued to be developed for more than forty years, and provided the framework for successful web search engines. This review finishes with a perspective on the future challenges for IR.

Pre-history – mechanical and electro-mechanical devices

Conventional approaches to managing large collections of information originate from the discipline of librarianship. Commonly, items such as books or papers were indexed using cataloguing schemes. Eliot and Rose claim this approach to be millennia old: declaring Callimachus, a 3rd century BC Greek poet as the first person known to create a library catalogue [2]. Facilitating faster search of these physical records was long researched, for example, Rudolph filed a US patent in 1891 for a machine composed of catalogue cards linked together, which could be wound past a viewing window enabling rapid manual scanning of the catalogue. Soper