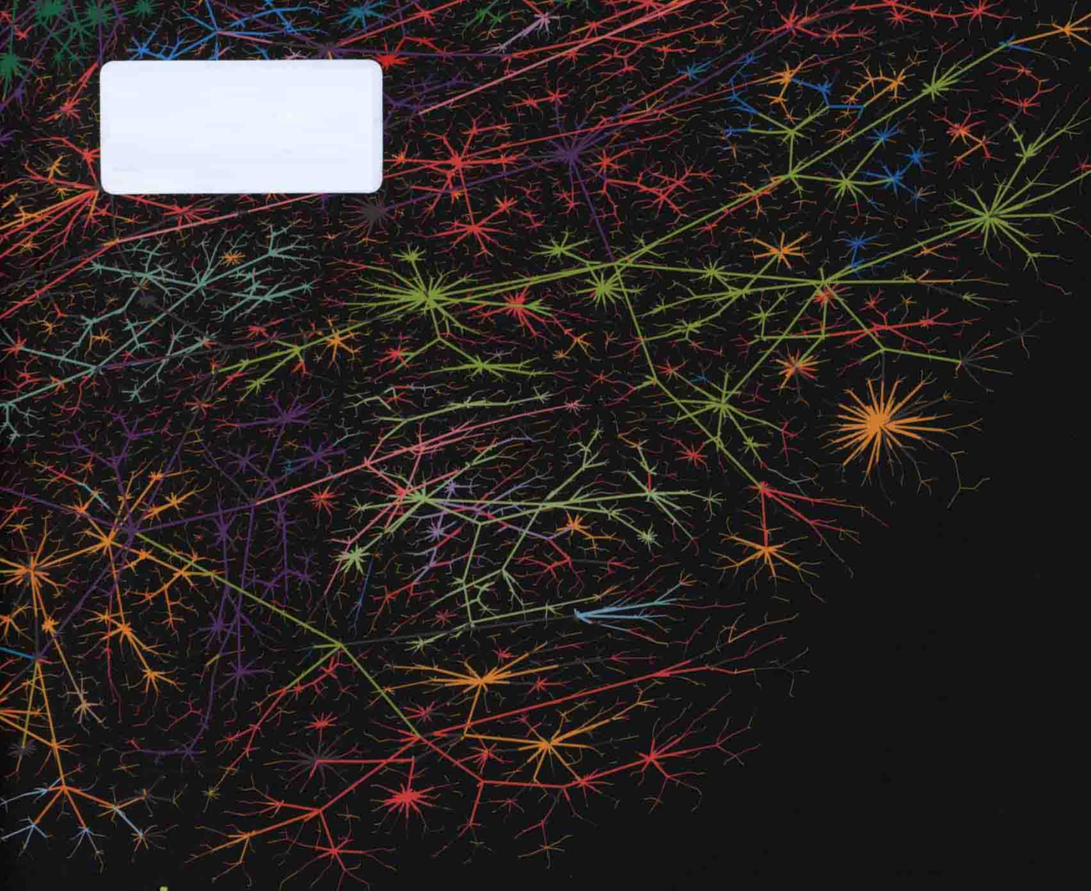

the Intelligent Web

Search, smart algorithms, and big data

GAUTAM SHROFF

the Intelligent Web

Search, Smart Algorithms, and Big Data

GAUTAM SHROFF

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2013
Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2013938816

ISBN 978–0–19–964671–5

Printed in Italy by
L.E.G.O. S.p.A.-Lavis TN

THE INTELLIGENT WEB

To my late father,
who I suspect would have enjoyed this book the most

ACKNOWLEDGEMENTS

Many people have contributed to my thinking and encouraged me while writing this book. But there are a few to whom I owe special thanks. First, to V. S. Subrahamanian, for reviewing the chapters as they came along and supporting my endeavour with encouraging words. I am also especially grateful to Patrick Winston and Pentti Kanerva for sparing the time to speak with me and share their thoughts on the evolution and future of AI.

Equally important has been the support of my family. My wife Brinda, daughter Selena, and son Ahan—many thanks for tolerating my preoccupation on numerous weekends and evenings that kept me away from you. I must also thank my mother for enthusiastically reading many of the chapters, which gave me some confidence that they were accessible to someone not at all familiar with computing.

Last but not least I would like to thank my editor Latha Menon, for her careful and exhaustive reviews, and for shepherding this book through the publication process.

Prologue

POTENTIAL

I grew up reading and being deeply influenced by the popular science books of George Gamow on physics and mathematics. This book is my attempt at explaining a few important and exciting advances in computer science and artificial intelligence (AI) in a manner accessible to all. The incredible growth of the internet in recent years, along with the vast volumes of 'big data' it holds, has also resulted in a rather significant confluence of ideas from diverse fields of computing and AI. This new 'science of *web intelligence*', arising from the marriage of many AI techniques applied together on 'big data', is the stage on which I hope to entertain and elucidate, in the spirit of Gamow, and to the best of my abilities.

* * *

The computer science community around the world recently celebrated the centenary of the birth of the British scientist Alan Turing, widely regarded as the father of computer science. During his rather brief life Turing made fundamental contributions in mathematics as well as some in biology, alongside crucial practical feats such as breaking secret German codes during the Second World War.

Turing was the first to examine very closely the meaning of what it means to 'compute', and thereby lay the foundations of computer science. Additionally, he was also the first to ask whether the capacity of intelligent thought could, in principle, be achieved by a machine that 'computed'. Thus, he is also regarded as the father of the field of enquiry now known as 'artificial intelligence'.

In fact, Turing begins his classic 1950 article[1] with, 'I propose to consider the question, "Can machines think?" ' He then goes on to describe the famous 'Turing Test', which he referred to as the 'imitation game', as a way to think about the problem of machines thinking. According to the Turing Test, if a computer can converse with any of us humans in so convincing a manner as to fool us into believing that it, too, is a human, then we should consider that machine to be 'intelligent' and able to 'think'.

Recently, in February 2011, IBM's Watson computer managed to beat champion human players in the popular TV show *Jeopardy!*. Watson was able to answer fairly complex queries such as 'Which New Yorker who fought at the Battle of Gettysburg was once considered the inventor of baseball?'. Figuring out that the answer is actually Abner Doubleday, and not Alexander Cartwright who actually wrote the rules of the game, certainly requires non-trivial natural language processing as well as probabilistic reasoning; Watson got it right, as well as many similar fairly difficult questions.

During this widely viewed *Jeopardy!* contest, Watson's place on stage was occupied by a computer panel while the human participants were visible in flesh and blood. However, imagine if instead the human participants were also hidden behind similar panels, and communicated via the same mechanized voice as Watson. Would we be able to tell them apart from the machine? Has the Turing Test then been 'passed', at least in this particular case?

There are more recent examples of apparently 'successful' displays of artificial intelligence: in 2007 Takeo Kanade, the well-known Japanese expert in computer vision, spoke about his early research in face recognition, another task normally associated with humans and at best a few higher-animals: 'it was with pride that I tested the program on 1000 faces, a rare case at the time when testing with 10 images was considered a "large-scale experiment".'[2] Today, both Facebook and Google's Picasa regularly recognize faces from among the hundreds of millions contained amongst the billions of images uploaded by users around the world.

Language is another arena where similar progress is visible for all to see and experience. In 1965 a committee of the US National Academy of Sciences concluded its review of the progress in automated translation between human natural languages with, 'there is no immediate or predicable prospect of useful machine translation'.[2] Today, web users around the world use Google's translation technology on a daily basis; even if the results are far from perfect, they are certainly good enough to be very useful.

Progress in spoken language, i.e., the ability to recognize speech, is also not far behind: Apple's Siri feature on the iPhone 4S brings usable and fairly powerful speech recognition to millions of cellphone users worldwide.

As succinctly put by one of the stalwarts of AI, Patrick Winston: 'AI is becoming more important while it becomes more inconspicuous', as 'AI technologies are becoming an integral part of mainstream computing'.[3]

* * *

What, if anything, has changed in the past decade that might have contributed to such significant progress in many traditionally 'hard' problems of artificial intelligence, be they machine translation, face recognition, natural language understanding, or speech recognition, all of which have been the focus of researchers for decades?

As I would like to convince you during the remainder of this book, many of the recent successes in each of these arenas have come through the deployment of many known but disparate techniques working together, and most importantly their deployment at *scale*, on large volumes of 'big data'; all of which has been made possible, and indeed driven, by the internet and the world wide web. In other words, rather than 'traditional' artificial intelligence, the successes we are witnessing are better described as those of '*web intelligence*' arising from 'big data'. Let us first consider what makes big data so 'big', i.e., its *scale.*

* * *

The web is believed to have well over a trillion web pages, of which at least 50 billion have been catalogued and *indexed* by search engines such as Google, making them searchable by all of us. This massive web content spans well over 100 million domains (i.e., locations where we point our browsers, such as <http://www.wikipedia.org>). These are themselves growing at a rate of more than 20,000 net domain additions daily. Facebook and Twitter each have over 900 million users, who between them generate over 300 million posts a day (roughly 250 million tweets and over 60 million Facebook updates). Added to this are the over 10,000 credit-card payments made per *second,*[*] the well-over 30 billion point-of-sale transactions per year (via dial-up POS devices[†]), and finally the over 6 billion mobile phones, of which almost 1 billion are smartphones, many of which are GPS-enabled, and which access the internet for e-commerce, tweets, and post updates on Facebook.[‡] Finally, and last but not least, there are the images and videos on YouTube and other sites, which by themselves outstrip all these put together in terms of the sheer volume of data they represent.

This deluge of data, along with emerging techniques and technologies used to handle it, is commonly referred to today as 'big data'. Such big data is both valuable and challenging, because of its sheer volume. So much so that the volume of data being created in the current five years from 2010 to 2015 will far exceed all the data generated in human history (which was estimated to be under 300 exabytes as of 2007[§]). The web, where all this data is being produced and resides, consists of millions of servers, with data storage soon to be measured in zetabytes.[¶]

[*] <http://www.creditcards.com>.

[†] <http://www.gaoresearch.com/POS/pos.php>.

[‡] <http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats>.

[§] <http://www.bbc.co.uk/news/technology-12419672>.

[¶] petabyte = 1,000 GB, exabyte = 1,000 petabytes, and a zetabyte = 1,000 petabytes.

On the other hand, let us consider the volume of data an average human being is exposed to in a lifetime. Our sense of vision provides the most voluminous input, perhaps the equivalent of half a million hours of video or so, assuming a fairly a long lifespan. In sharp contrast, YouTube alone witnesses 15 million hours of *fresh* video uploaded every year.

Clearly, the volume of data available to the millions of machines that power the web far exceeds that available to any human. Further, as we shall argue later on, the millions of servers that power the web at least match if not exceed the raw computing capacity of the 100 billion or so neurons in a single human brain. Moreover, each of these servers are certainly much much faster at computing than neurons, which by comparison are really quite slow.

Lastly, the advancement of computing technology remains relentless: the well-known Moore's Law documents the fact that computing power per dollar appears to double every 18 months; the lesser known but equally important Kryder's Law states that storage capacity per dollar is growing even faster. So, for the first time in history, we have available to us both the computing power as well as the raw data that matches and shall very soon far exceed that available to the average human.

Thus, we have the *potential* to address Turing's question 'Can machines think?', at least from the perspective of raw computational power and data of the same order as that available to the human brain. How far have we come, why, and where are we headed? One of the contributing factors might be that, only recently after many years, does 'artificial intelligence' appear to be regaining a semblance of its initial ambition and unity.

* * *

In the early days of artificial intelligence research following Turing's seminal article, the diverse capabilities that might be construed to comprise intelligent behaviour, such as vision, language, or logical reasoning, were often discussed, debated, and shared at common forums. The goals exposed by the now famous Dartmouth conference of 1956, considered to be a landmark event in the history of AI, exemplified both a unified approach to all problems related to machine intelligence as well as a marked overconfidence:

> We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.[4]

These were clearly heady times, and such gatherings continued for some years. Soon the realization began to dawn that the 'problem of AI' had been grossly underestimated. Many sub-fields began to develop, both in reaction to the growing number of researchers trying their hand at these difficult challenges, and because of conflicting goals. The original aim of actually answering the question posed by Turing was soon found to be too challenging a task to tackle all at once, or, for that matter, attempt at all. The proponents of 'strong AI', i.e., those who felt that true 'thinking machines' were actually possible, with their pursuit being a worthy goal, began to dwindle. Instead, the practical applications of AI techniques, first developed as possible answers to the strong-AI puzzle, began to lead the discourse, and it was this 'weak AI' that eventually came to dominate the field.

Simultaneously, the field split into many sub-fields: image processing, computer vision, natural language processing, speech recognition, machine learning, data mining, computational reasoning, planning, etc. Each became a large area of research in its own right. And rightly so, as the practical applications of specific techniques necessarily appeared to lie within disparate areas: recognizing faces versus translating between two languages; answering questions in natural language versus recognizing spoken words; discovering knowledge from volumes of documents versus logical reasoning; and the list goes on. Each of these were so clearly separate application domains that it made eminent sense to study them separately and solve such obviously different practical problems in purpose-specific ways.

Over the years the AI research community became increasingly fragmented. Along the way, as Pat Winston recalled, one would hear comments such as 'what are all these vision people doing here'[3] at a conference dedicated to say, 'reasoning'. No one would say, 'well, because we think with our eyes',[3] i.e., our perceptual systems are intimately involved in thought. And so fewer and fewer opportunities came along to discuss and debate the 'big picture'.

* * *

Then the web began to change everything. Suddenly, the practical problem faced by the web companies became larger and more holistic: initially there were the search engines such as Google, and later came the social-networking platforms such as Facebook. The problem, however, remained the same: how to make more money from advertising?

The answer turned out to be surprisingly similar to the Turing Test: Instead of merely fooling us into believing it was human, the 'machine', i.e., the millions of servers powering the web, needed to *learn* about each of us, individually, just as we all learn about each other in casual conversation. Why? Just so that better, i.e., more closely targeted, advertisements could be shown to us, thereby leading to better 'bang for the buck' of every advertising dollar. This then became the holy grail: not intelligence per se, just doing better and better at this 'reverse' Turing Test, where instead of us being observer and 'judge', it is the machines in the web that observe and seek to 'understand' us better for their own selfish needs, if only to 'judge' whether or not we are likely buyers of some of the goods they are paid to advertise. As we shall see soon, even these more pedestrian goals required weak-AI techniques that could mimic many of *capabilities* required for intelligent thought.

Of course, it is also important to realize that none of these efforts made any strong-AI claims. The manner in which seemingly intelligent capabilities are computationally realized in the web does not, for the most part, even attempt to mirror the mechanisms nature has evolved to bring intelligence to life in real brains. Even so, the results are quite surprising indeed, as we shall see throughout the remainder of this book.

At the same time, this new holy grail could not be grasped with disparate weak-AI techniques operating in isolation: our queries as we searched the web or conversed with our friends were *words*; our actions as we surfed and navigated the web were *clicks*. Naturally we wanted to *speak* to our phones rather than type, and the videos that we uploaded and shared so freely were, well, videos.

Harnessing the vast trails of data that we leave behind during our web existences was essential, which required expertise from different fields of AI, be they language processing, learning, reasoning, or vision, to come together and connect the dots so as to even come close to understanding *us*.

First and foremost the web gave us a different way to *look* for information, i.e., web search. At the same time, the web itself would *listen* in, and *learn*, not only about us, but also from our collective knowledge that we have so well digitized and made available to all. As our actions are observed, the web-intelligence programs charged with pinpointing advertisements for us would need to *connect* all the dots and *predict* exactly which ones we should be most interested in.

Strangely, but perhaps not surprisingly, the very synthesis of techniques that the web-intelligence programs needed in order to connect the dots in their practical enterprise of online advertising appears, in many respects, similar to how we ourselves integrate our different perceptual and cognitive abilities. We consciously *look* around us to gather information about our environment as well as *listen* to the ambient sea of information continuously bombarding us all. Miraculously, we *learn* from our experiences, and *reason* in order to *connect* the dots and make sense of the world. All this so as to *predict* what is most likely to happen next, be it in the next instant, or eventually in the course of our lives. Finally, we *correct* our actions so as to better achieve our goals.

* * *

I hope to show how the cumulative use of artificial intelligence techniques at web scale, on hundreds of thousands or even millions of computers, can result in behaviour that exhibits a very basic feature of human intelligence, i.e., to colloquially speaking 'put two and two together' or 'connect the dots'. It is this ability that allows us to make sense of the world around us, make intelligent guesses about what is most likely to happen in the future, and plan our own actions accordingly.

Applying web-scale computing power on the vast volume of 'big data' now available because of the internet, offers the *potential* to create far more intelligent systems than ever before: this defines the new science of *web intelligence*, and forms the subject of this book.

At the same time, this remains primarily a book about weak AI: however powerful this web-based synthesis of multiple AI techniques might appear to be, we do not tread too deeply in the philosophical waters of strong-AI, i.e., whether or not machines can ever be 'truly intelligent', whether consciousness, thought, self, or even 'soul' have reductionist roots, or not. We shall neither speculate much on these matters nor attempt to describe the diverse philosophical debates and arguments on this subject. For those interested in a comprehensive history of the confluence of philosophy, psychology, neurology, and artificial intelligence often referred to as 'cognitive science', Margaret

Boden's recent volume *Mind as Machine: A History of Cognitive Science*[5] is an excellent reference.

Equally important are Turing's own views as elaborately explained in his seminal paper[1] describing the 'Turing test'. Even as he clearly makes his own philosophical position clear, he prefaces his own beliefs and arguments for them by first clarifying that 'the original question, "Can machines think?" I believe to be too meaningless to deserve discussion'.[1] He then rephrases his 'imitation game', i.e., the Turing Test that we are all familiar with, by a *statistical* variant: 'in about fifty years' time it will be possible to program computers . . . so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning'.[1] Most modern-day machine-learning researchers might find this formulation quite familiar indeed. Turing goes on to speculate that 'at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted'.[1] It is the premise of this book that such a time has perhaps arrived.

As to the 'machines' for whom it might be colloquially acceptable to use the word 'thinking', we look to the web-based engines developed for entirely commercial pecuniary purposes, be they search, advertising, or social networking. We explore how the computer programs underlying these engines sift through and make sense of the vast volumes of 'big data' that we continuously produce during our online lives—our collective 'data exhaust', so to speak.

In this book we shall quite often use Google as an example and examine its innards in greater detail than others. However, when we speak of Google we are also using it as a metaphor: other search engines, such as Yahoo! and Bing, or even the social networking world of Facebook and Twitter, all share many of the same processes and purposes.

The purpose of all these web-intelligence programs is simple: 'all the better to understand us', paraphrasing Red Riding Hood's wolf in grandmother's clothing. Nevertheless, as we delve deeper into what these vast syntheses of weak-AI techniques manage to achieve in practice, we do find ourselves wondering whether these web-intelligence systems might end up serving us a dinner far closer to strong AI than we have ever imagined for decades.

That hope is, at least, one of the reasons for this book.

* * *

In the chapters that follow we dissect the ability to connect the dots, be it in the context of web-intelligence programs trying to understand us, or our own ability to understand and make sense of the world. In doing so we shall find some surprising parallels, even though the two contexts and purposes are so very different. It is these connections that offer the potential for increasingly capable web-intelligence systems in the future, as well as possibly deeper understanding and appreciation of our own remarkable abilities.

Connecting the dots requires us to *look* at and experience the world around us; similarly, a web-intelligence program looks at the data stored in or streaming across the internet. In each case information needs to be stored, as well as retrieved, be it in the form of memories and their recollection in the former, or our daily experience of web search in the latter.

Next comes the ability to *listen*, to focus on the important and discard the irrelevant. To recognize the familiar, discern between alternatives or identify similar things. Listening is also about 'sensing' a momentary experience, be it a personal feeling, individual decision, or the collective sentiment expressed by the online masses. Listening is followed eventually by deeper understanding: the ability to *learn* about the structure of the world, in terms of facts, rules, and relationships. Just as we learn common-sense knowledge about the world around us, web-intelligence systems learn about our preferences and