AN INTRODUCTION TO SEARCH ENGINES AND WEB NAVIGATION

MARK LEVENE
Department of Computer Science and Information Systems
Birkbeck University of London, UK

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

AN INTRODUCTION TO SEARCH ENGINES AND WEB NAVIGATION

AN INTRODUCTION TO SEARCH ENGINES AND WEB NAVIGATION

MARK LEVENE
Department of Computer Science and Information Systems
Birkbeck University of London, UK

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Levene, M. (Mark), 1957-
An introduction to search engines and web navigation / Mark Levene.
p. cm.
ISBN 978-0-470-52684-2 (pbk.)
1. Internet searching. 2. Web search engines. I. Title.
ZA4230.L48 2010
025.0425–dc22
2010008435

Printed in Singapore

10 9 8 7 6 5 4 3 2 1

To my wife Sara and three children
Tamara, Joseph and Oren

CONTENTS

PREFACE

MOTIVATION

Searching and navigating the web have become part of our daily online lives. Web browsers and the standard navigation tools embedded in them provide a showcase of successful software technology with a global user-base, that has changed the way in which we search for and interact with information. Search engine technology has become ubiquitous, providing a standard interface to the endless amount of information that the web contains. Since the inception of the web, search engines have delivered a continuous stream of innovations, satisfying their users with increasingly accurate results through the implementation of advanced retrieval algorithms and scalable distributed architectures. Search and navigation technologies are central to the smooth operation of the web and it is hard to imagine finding information without them. Understanding the computational basis of these technologies and the models underlying them is of paramount importance both for IT students and practitioners.

There are several technical books on web search and navigation but the ones I have seen are either very academic in nature, that is, targeted at the postgraduate student or advanced researcher, and therefore have a limited audience, or they concentrate on the user interface and web site usability issues, ignoring the technicalities of what is happening behind the scenes. These books do not explain at an introductory level how the underlying computational tools work. This book answers the need for an introductory, yet technical, text on the topic.

My research into web search and navigation technologies started during the beginning of the 1990s just before the internet boom, when, together with my colleagues, we began looking at hypertext as a model for unstructured (or semistructured) data connected via a network of links, much in the same way web pages are connected. Of particular interest to us was the infamous "navigation problem" when we lose our way navigating (or what has become known as "surfing") through the myriad of information pages in the network. Tackling this problem has provided continued impetus for my research.

In a wider context, the activity of information seeking, that is, the process we go through when searching and locating information in order to augment our state of knowledge, has been of major concern to all involved in the development of technologies that facilitate web interaction.

I have been using browser navigation tools and search engines since their early days, and have been fascinated by the flow of new ideas and the improvements that each new tool has delivered. One of my aims in this text is to demystify the technology underlying the tools that we use in our day-to-day interaction with the web, and another is to inform readers about upcoming technologies, some of which are still in the research and development stage.

I hope that this book will instill in you some of my enthusiasm for the possibilities that these technologies have and are creating to extend our capabilities of finding and sharing information.

AUDIENCE AND PREREQUISITES

The book is intended as an undergraduate introductory text on search and navigation technologies, but could also be used to teach an option on the subject. It is also intended as a reference book for IT professionals wishing to know how these technologies work and to learn about the bigger picture in this area.

The course has no formal prerequisites, all that is required is for the learner to be a user of the web and to be curious to know how these technologies work. All the concepts that are introduced are explained in words, and simple examples from my own experience are given to illustrate various points. Occasionally, to add clarity to an important concept, a formula is given and explained. Each chapter starts with a list of learning objectives and ends with a brief bullet-pointed summary. There are several exercises at the end of each chapter. Some of these aim to get the student to explore further issues, possibly with a reference which can be followed up, some get the student to discuss an aspect of the technology, and others are mini-projects (which may involve programming) to add to the student's understanding through a hands-on approach. The book ends with a set of notes containing web addresses to items mentioned in the book, and an extensive bibliography of the articles and books cited in the book.

Readers should be encouraged to follow the links in the text and to discover new and related links that will help them understand how search and navigation tools work, and to widen their knowledge with related information.

TIMELINESS

I believe that due to the importance of the topic it is about time that such a book should appear. Search and navigation technologies are moving at a very fast pace due to the continued growth of the web and its user base, and improvements in computer networking and hardware. There is also strong competition between different service providers to lock-in users to their products. This is good news for web users, but as a result some of the numerics in the text may be out of date. I have qualified the statistics I have given with dates and links, which can be found in the notes, so the reader can follow these to get an up-to-date picture and follow the trends. I do not expect the core technologies I have covered to radically change in the near future and I would go so far as to claim that in essence they are fundamental to the web's working, but innovation and new ideas will continue to flourish and mold the web's landscape.

If you find any errors or omissions please let me know so that I can list them on the book's web site. I will also be grateful to receive any constructive comments and suggestions, which can be used to improve the text.

ACKNOWLEDGMENTS

First I would like to thank my wife and family who have been extremely supportive throughout this project, encouraging me to put in the extra hours needed to complete such a task. I would also like to thank my colleagues at the Department of Computer Science and Information Systems at Birkbeck, who have read and commented on parts of the book. Special thanks to my editors at Wiley, Lucy Hitz and George Telecki, who have patiently guided me through the publication process. Finally, I would like to thank the reviewers for their constructive comments.

The people who have built the innovative technologies that drive today's web are the real heroes of the revolution that the World Wide Web has brought upon us. Without them, this book could not have been written. Not only in terms of the content of the book, but also in terms of the tools I have been using daily to augment my knowledge on how search and navigation technologies work in practice.

Mark Levene
London, June 2010

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

> "People keep asking me what I think of it now it's done. Hence my protest: The Web is not done!"
>
> — Tim Berners-Lee, Inventor of the World Wide Web

THE LAST two decades have seen dramatic revolutions in information technology; not only in computing power, such as processor speed, memory size, and innovative interfaces, but also in the everyday use of computers. In the late 1970s and during the 1980s, we had the revolution of the personal computer (PC), which brought the computer into the home, the classroom, and the office. The PC then evolved into the desktop, the laptop, and the netbook as we know them today.

The 1990s was the decade of the World Wide Web (the Web), built over the physical infrastructure of the Internet, radically changing the availability of information and making possible the rapid dissemination of digital information across the globe. While the Internet is a physical network, connecting millions of computers together globally, the Web is a virtual global network linking together a massive amount of information. Search engines now index many billions of web pages and that number is just a fraction of the totality of information we can access on the Web, much of it residing in searchable databases not directly accessible to search engines.

Now, in the twenty-first century we are in the midst of a third wave of novel technologies, that of mobile and wearable computing devices, where computing devices have already become small enough so that we can carry them around with us at all times, and they also have the ability to interact with other computing devices, some of which are embedded in the environment. While the Web is mainly an informational and transactional tool, mobile devices add the dimension of being a location-aware ubiquitous social communication tool.

Coping with, organizing, visualizing, and acting upon the massive amount of information with which we are confronted when connected to the Web are amongst the main problems of *web interaction* [421]. Searching and navigating (or surfing) the Web are the methods we employ to help us find information

An Introduction to Search Engines and Web Navigation, by Mark Levene on the web, using search engines and navigation tools that are either built-in or plugged-in to the browser or are provided by web sites.

In this book, we explore search and navigation technologies to their full, present the State-of-the art tools, and explain how they work. We also look at ways of modeling different aspects of the Web that can help us understand how the Web is evolving and how it is being and can be used. The potential of many of the technologies we introduce has not yet been fully realized, and many new ideas to improve the ways in which we interact with the Web will inevitably appear in this dynamic and exciting space.

1.1 BRIEF SUMMARY OF CHAPTERS

This book is roughly divided into three parts. The first part (Chapters 1–3) introduces the problems of web interaction dealt with in the book, the second part (Chapters 4–6) deals with web search engines, and the third part (Chapters 7–9) looks at web navigation, the mobile web, and social network technologies in the context of search and navigation. Finally, in Chapter 10, we look ahead at the future prospects of search and navigation on the Web.

Chapters 1–3 introduce the reader to the problems of search and navigation and provide background material on the Web and its users. In particular, in the remaining part of Chapter 1, we give brief histories of hypertext and the Web, and of search engines. In Chapter 2, we look at some statistics regarding the Web, investigate its structure, and discuss the problems of information seeking and web search. In Chapter 3, we introduce the navigation problem, discuss the potential of machine learning to improve search and navigation tools, and propose Markov chains as a model for user navigation.

Chapters 4–6 cover the architectural and technical aspects of search engines. In particular, in Chapter 4, we discuss the search engine wars, look at some usage statistics of search engines, and introduce the architecture of a search engine, including the details of how the Web is crawled. In Chapter 5, we dissect a search engine's ranking algorithm, including content relevance, link- and popularity-based metrics, and different ways of evaluating search engines. In Chapter 6, we look at different types of search engines, namely, web directories, search engine advertising, metasearch engines, personalization of search, question answering engines, and image search and special purpose engines.

Chapters 7–9 concentrate on web navigation, and looks beyond at the mobile web and at how viewing the Web in social network terms is having a major impact on search and navigation technologies. In particular, in Chapter 7, we discuss a range of navigation tools and metrics, introduce web data mining and the Best Trail algorithm, discuss some visualization techniques to assist navigation, and look at the issues present in real-world navigation. In Chapter 8, we introduce the mobile web in the context of mobile computing, look at the delivery of mobile web services, discuss interfaces to mobile devices, and present the problems of search and navigation in a mobile context. In Chapter 9, we introduce social networks in the context of the Web, look at social network analysis, introduce peer-to-peer networks, look at the technology of collaborative filtering, introduce weblogs as a medium for personal journalism on the Web, look at the ubiquity of power-law distributions on the Web, present effective searching strategies in social networks, introduce opinion mining as a way of obtaining knowledge about users opinions and sentiments, and look at Web 2.0 and collective intelligence that have generated a lot of hype and inspired many start-ups in recent years.

1.2 BRIEF HISTORY OF HYPERTEXT AND THE WEB

The history of the Web dates back to 1945 when Vannevar Bush, then an advisor to President Truman, wrote his visionary article "As We May Think," and described his imaginary desktop machine called *memex*, which provides personal access to all the information we may need [119]. An artist's impression of memex is shown in Fig. 1.1.

The memex is a "sort of mechanized private file and library," which supports "associative indexing" and allows navigation whereby "any item may be caused at will to select immediately and automatically another." Bush emphasizes that "the process of tying two items together is an important thing." By repeating this process of creating links, we can form a *trail* which can be traversed by the user; in Bush's words, "when numerous items have been thus joined together to form a trail they can be reviewed in turn." The motivation for the memex's support of trails as first-class objects was that the human mind "operates by association" and "in accordance to some intricate web of trails carried out by the cells of the brain."

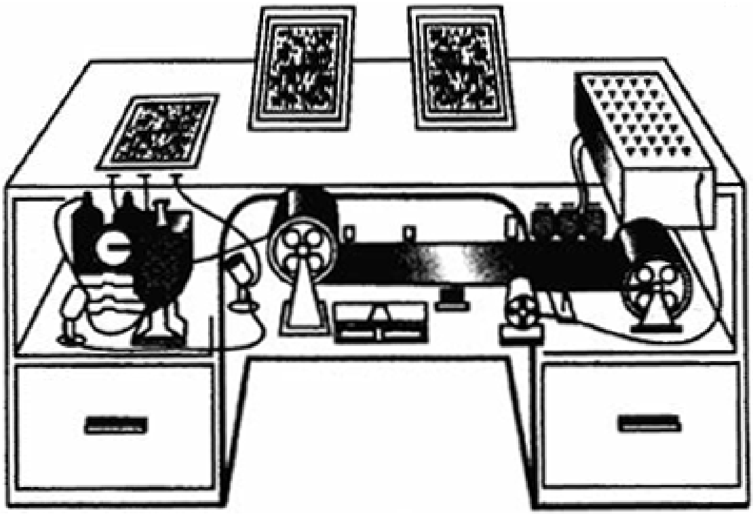


Figure 1.1 Bush's memex. (Source: Life Magazine 1945;9(11):123.)

Bush also envisaged the "new profession of trailblazers" who create trails for other memex users, thus enabling sharing and exchange of knowledge. The memex was designed as a personal desktop machine, where information is stored locally on the machine. Trigg [647] emphasizes that Bush views the activities of creating a new trail and following a trail as being connected. Trails can be authored by trailblazers based on their experience and can also be created by memex, which records all user navigation sessions. In his later writings on the memex, published in Ref. 509, Bush revisited and extended the memex concept. In particular, he envisaged that memex could "learn from its own experience" and "refine its trails." By this, Bush means that memex collects statistics on the trails that the user follows and "notices" the ones that are most frequently followed. Oren [516] calls this extended version *adaptive memex*, stressing that adaptation means that trails can be constructed dynamically and given semantic justification; for example, by giving these new trails meaningful names.

The term *hypertext* [503] was coined by Ted Nelson in 1965 [495], who considers "a literature" (such as the scientific literature) to be a *system of interconnected writings*. The process of referring to other connected writings, when reading an article or a document, is that of *following links*. Nelson's vision is that of creating a repository of all the documents that have ever been written thus achieving a universal hypertext. Nelson views his hypertext system, which he calls *Xanadu*, as a network of distributed documents that should be allowed to grow without any size limit, such that users, each corresponding to a node in the network, may link their documents to any other documents in the network. Xanadu can be viewed as a generalized memex system, which is both for private and public use. As with memex, Xanadu remained a vision that was not fully implemented; a mockup of Xanadu's linking mechanism is shown in Fig. 1.2. Nelson's pioneering work in hypertext is materialized to a large degree in the Web, since he also views his system as a means of publishing material by making it universally available to a wide network of interconnected users.

Douglas Engelbart's o*n*-*l*ine *s*ystem (NLS) [205] was the first working hypertext system, where documents could be linked to other documents and thus groups of people could work collaboratively. The video clips of Engelbart's historic demonstration of NLS from December 1968 are archived on the Web,[1] and a recollection of the demo can be found in Ref. 204; a picture of Engelbart during the demo is shown in Fig. 1.3.

About 30 years later in 1990, Tim Berners-Lee—then working for Cern, the world's largest particle physics laboratory—turned the vision of hypertext into reality by creating the World Wide Web as we know it today [77].[2]

The Web works using three conventions: (i) the URL (unified resource locator) to identify web pages, (ii) HTTP (hypertext transfer protocol) to exchange messages between a browser and web server, and (iii) HTML (hypertext markup language) [501] to display web pages. More recently, Tim Berners-Lee has been

[1] Video clips from Engelbart's demo can be found at http://sloan.stanford.edu/mousesite/1968Demo.html.

[2] A little history of the World Wide Web from 1945 to 1995. www.w3.org/History.html.

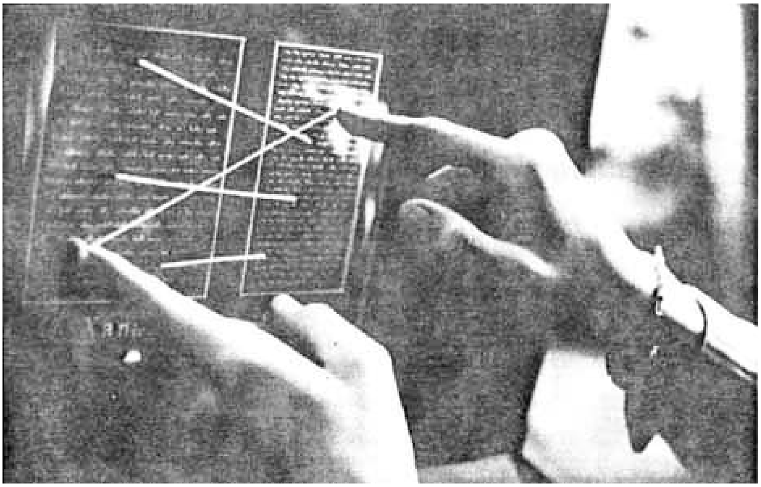
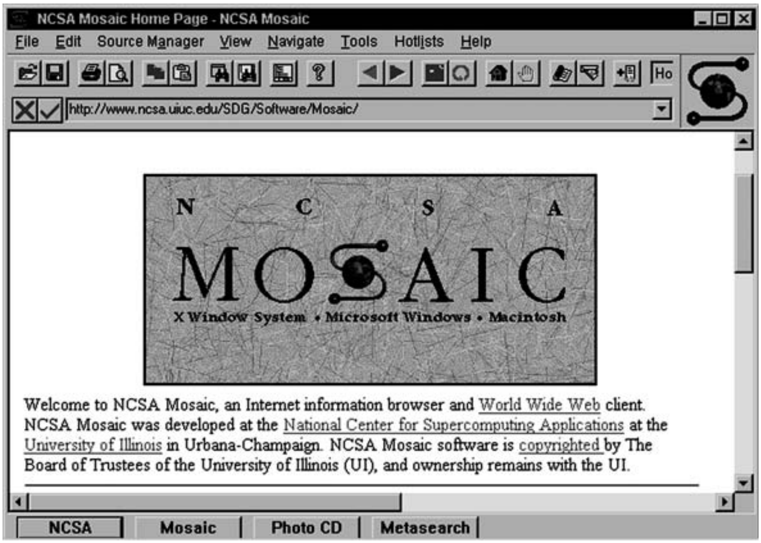


Figure 1.2 Nelson's Xanadu. (Source: Figure 1.3, Xanalogical structure, needed now more than ever: Parallel documents, deep links to content, deep versioning, and deep re-use, by Nelson TH. www.cs.brown.edu/memex/ACM_HypertextTestbed/papers/60.html.)

Figure 1.3 Engelbart's NLS. (Source: Home video of the birth of the hyperlink. www.ratchetup.com/eyes/2004/01/wired_recently_.html.)

promoting the semantic web [78] together with XML (extensible markup language) [259], and RDF (resource description framework) [544], as a means of creating machine understandable information that can better support end user web applications. Details on the first web browser implemented by Tim Berners-Lee in 1990 can be found at www.w3.org/People/Berners-Lee/WorldWideWeb.


NCSA Mosaic Home Page - NCSA Mosaic
File Edit Source Manager View Navigate Tools Hotlists Help
Ho
http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/
N C S A
MOSAIC
X Window System • Microsoft Windows • Macintosh
NCSA
Mosaic
Photo CD
Metasearch

Welcome to NCSA Mosaic, an Internet information browser and World Wide Web client. NCSA Mosaic was developed at the National Center for Supercomputing Applications at the University of Illinois in Urbana-Champaign. NCSA Mosaic software is copyrighted by The Board of Trustees of the University of Illinois (UI), and ownership remains with the UI.

Figure 1.4 Mosaic browser initially released in 1993. (Source: http://gladiator.ncsa.illinois.edu/Images/press-images/mosaic.gif.)

The creation of the Mosaic browser by Marc Andreessen in 1993 followed by the creation of Netscape early in 1994 were the historic events that marked the beginning of the internet boom that lasted throughout the rest of the 1990s, and led to the mass uptake in web usage that continues to increase to this day. A screenshot of an early version of Mosaic is shown in Fig. 1.4.

1.3 BRIEF HISTORY OF SEARCH ENGINES

The roots of web search engine technology are in information retrieval (IR) systems, which can be traced back to the work of Luhn at IBM during the late 1950s [444]. IR has been an active field within information science since then, and has been given a big boost since the 1990s with the new requirements that the Web has brought.

Many of the methods used by current search engines can be traced back to the developments in IR during the 1970s and 1980s. Especially influential is the SMART (system for the mechanical analysis and retrieval of text) retrieval system, initially developed by Gerard Salton and his collaborators at Cornell University during the early 1970s [583]. An important treatment of the traditional approaches to IR was given by Keith van Rijsbergen [655], while more modern treatments with reference to the Web can be found in Refs 45, 68, 453, and 164. More recent developments, which concentrate on web technologies, are the probabilistic perspective on modeling the Web as in Ref. 46 and the data mining perspective on managing web information, which can be found in Refs 128 and 435.

Owing to the massive amount of information on the Web, right from the early days of the Web, search engines have become an indispensable tool for web users. A history of search engines detailing some of the early search services can be found in Ref. 659.[3]

Here, we will be very selective and mention only a few of the early and current search engines; see http://searchenginewatch.com/links and http://en.wikipedia.org/wiki/List_of_search_engines for up-to-date listings of the major search engines. More details on many of the current search engines are spread throughout the book.

- Yahoo (www.yahoo.com), which started up in February 1994, was one of the earliest search services.[4] Initially, Yahoo was only providing a browsable directory, organizing web pages into categories which were classified by human editors. Yahoo continues to maintain a strong brand and has evolved into a full-fledged search engine by acquiring existing search engine technology in mid-2003. (You can get some insight on the latest innovations in Yahoo's search engine from its weblog at www.ysearchblog.com.)
- InfoSeek, which started up in July 1994, was the first search engine that I was using on a regular basis, and as with many of the innovative web tools, users voted with their clicks and its reputation spread by word of mouth. In July 1998, Infoseek merged with Walt Disney's Buena Vista Internet Group to form Go.com, which was ultimately abandoned in January 2001.
- Inktomi, which started up in September 1995, provides search engine infrastructure rather than delivering the service from their web site. Until it was acquired by Yahoo in March 2003, it was providing search services to some of the major search engines.
- AltaVista (www.altavista.com), which started up in December 1995, was the second search engine that I was using on a regular basis. It was initially a research project in Digital Equipment Corporation, and was eventually acquired by Overture in April 2003.
- AlltheWeb (www.alltheweb.com) was launched in May 1999 by Fast Search & Transfer, and in a very short time was able to build a very large and fresh index with fast and accurate search results. It was also acquired by Overture in April 2003.
- Ask Jeeves (www.ask.com) started up in April 1996. It went public in July 1999, and is one of the survivors in the search engine game. Its strong brand and distinctive question answering facility have evolved into a general search service through its acquisition of Teoma in September 2001, which has enabled it to manage a proprietary search service and develop its own search technology. It was acquired by e-commerce conglomerate IAC (InterActiveCorp) in July 2005.

[3] See also, A history of search engines, by W. Sonnenreich. www.wiley.com/legacy/compbooks/sonnenreich/history.html.

[4] The history of Yahoo!—How it all started. http://docs.yahoo.com/info/misc/history.html.

- Overture (www.overture.com) started up as Goto.com in September 1997, and pioneered pay-per-click search engine advertising. It was renamed as Overture in September 2001 and was acquired by Yahoo in July 2003. In April 2005, Overture was rebranded as Yahoo Search Marketing (http://searchmarketing.yahoo.com).
- Bing (www.bing.com) is Microsoft's search engine that went online in June 2009. It replaced Live search, released in September 2006, which replaced MSN search, originally launched in August 1995, coinciding with the release of Windows 95. Initially, MSN search partnered with major search engines to provide the search facility for their site. Realizing the strategic importance of search to Microsoft's core business, Microsoft announced, in 2003, that it would develop its own proprietary search technology. The beta version of the search engine was released by MSN in November 2004, and in February 2005 MSN search was officially delivering search results from its internally developed engine. (You can get some insight on the latest innovations in Bing's search engine from its weblog at www.bing.com/community/blogs/search.)
- Google (www.google.com) was started up in September 1998, by Larry Page and Sergey Brin, then PhD students at Stanford University.[5] Google was the third search engine that I was using on a regular basis and am still using today, although I do consult other search services as well. It became a public company in August 2004, and, as of late 2004, has been the most popular search engine. You will find a wealth of information in this book on the innovative features that Google and other search engines provide. (You can get some insight on the latest innovations in Google's search engine from its weblog at http://googleblog.blogspot.com.)

[5] Google History. www.google.com/corporate/history.html.