Authorship Attribution
Authorship Attribution

Patrick Juola

Department of Mathematics and Computer Science
Duquesne University
Pittsburgh, PA 15282
USA
juola@mathcs.duq.edu

Full text available at: http://dx.doi.org/10.1561/1500000005
Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2006, Volume 1, 4 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.
Authorship Attribution

Patrick Juola

Department of Mathematics and Computer Science, Duquesne University, 600 Forbes Avenue, Pittsburgh, PA 15282, USA, juola@mathcs.duq.edu

Abstract

Authorship attribution, the science of inferring characteristics of the author from the characteristics of documents written by that author, is a problem with a long history and a wide range of application. Recent work in “non-traditional” authorship attribution demonstrates the practicality of automatically analyzing documents based on authorial style, but the state of the art is confusing. Analyses are difficult to apply, little is known about type or rate of errors, and few “best practices” are available. In part because of this confusion, the field has perhaps had less uptake and general acceptance than is its due.

This review surveys the history and present state of the discipline, presenting some comparative results when available. It shows, first, that the discipline is quite successful, even in difficult cases involving small documents in unfamiliar and less studied languages; it further analyzes the types of analysis and features used and tries to determine characteristics of well-performing systems, finally formulating these in a set of recommendations for best practices.
## Contents

1 Introduction  
1.1 Why “Authorship Attribution”? 1  
1.2 Structure of the Review 2

2 Background and History  
2.1 Problem Statement 5  
2.2 Theoretical Background 7  
2.3 Historical Overview 8  
2.4 Methodological Issues in Authorship Attribution 13  
2.5 What Would an Ideal Authorship Attribution System Look Like? 17

3 Linguistic and Mathematical Background  
3.1 Mathematical Linguistics 21  
3.2 Information Theory 25  
3.3 Matrices and Vector Spaces 27  
3.4 A Theoretical Framework 29

4 Linguistic Features  
4.1 Vocabulary as a Feature 31  
4.2 Vocabulary Properties as Features 33  
4.3 Syntactic Properties as Features 34  
4.4 Miscellaneous Features 35
1

Introduction

1.1 Why “Authorship Attribution”?

In 2004, Potomac Books published Imperial Hubris: Why the West is Losing the War on Terror. Drawing on the author’s extensive personal experience, the book described the current situation of the America-led war on terror and argued that much US policy was misguided.

Or did he? The author of the book is technically “Anonymous,” although he claims (on the dust cover) to be “a senior US intelligence official with nearly two decades of experience” as well as the author of the 2003 book Through Our Enemies’ Eyes. According to the July 2, 2004 edition of the Boston Phoenix, the actual author was Michael Scheuer, a senior CIA officer and head of the CIA’s Osama bin Laden unit in the late 1990s. If true, this would lend substantial credibility to the author’s arguments.

But on the other hand, according to some noted historians such as Hugh Trevor-Roper, the author of the 1983 Hitler Diaries was Hitler himself, despite the later discovery that they were written on modern paper and using ink which was unavailable in 1945. Is Imperial Hubris another type of sophisticated forgery? Why should we believe historians...
and journalists, no matter how eminent? What kind of evidence should we demand before we believe?

Determining the author of a particular piece of text has raised methodological questions for centuries. Questions of authorship can be of interest not only to humanities scholars, but in a much more practical sense to politicians, journalists, and lawyers as in the examples above. Investigative journalism, combined with scientific (e.g., chemical) analysis of documents and simple close reading by experts has traditionally given good results. But recent developments of improved statistical techniques in conjunction with the wider availability of computer-accessible corpora have made the automatic and objective inference of authorship a practical option. This field has seen an explosion of scholarship, including several detailed book-length treatments \[39, 41, 44, 89, 98, 103, 105, 111, 112, 150\]. Papers on authorship attribution routinely appear at conferences ranging from linguistics and literature through machine learning and computation, to law and forensics. Despite — or perhaps because of — this interest, the field itself is somewhat in disarray with little overall sense of best practices and techniques.

1.2 Structure of the Review

This review therefore tries to present an overview and survey of the current state of the art. We follow the theoretical model (presented in detail in Section 3.4) of \[76\] in dividing the task into three major subtasks, each treated independently.

Section 2 presents a more detailed problem statement in conjunction with a historical overview of some approaches and major developments in the science of authorship attribution. Included is a discussion of some of the major issues and obstacles that authorship attribution faces as a problem, without regard to any specific approach, and the characteristics of a hypothetical “good” solution (unfortunately, as will be seen in the rest of the review, we have not yet achieved such a “good” solution).

Section 3 presents some linguistic, mathematical, and algorithmic preliminaries. Section 4 describes some of the major feature sets that
have been applied to authorship attribution, while Section 5 describes the methods of analysis applied to these features. Section 6 goes on to present some results in empirical evaluation and comparative testing of authorship attribution methods, focusing mainly on the results from the 2004 Ad-hoc Authorship Attribution Competition [75], the largest-scale comparative test to date.

Section 7 presents some other applications of these methods and technology, that, while not (strictly speaking) “authorship” attribution, are closely related. Examples of this include gender attribution or the determination of personality and mental state of the author. Section 8 discusses the specific problems of using authorship attribution in court, in a forensic setting. Finally, for those practical souls who want only to solve problems, Section 9 presents some recommendations about the current state of the art and the best practices available today.
References


References

References


References

References


References


References


