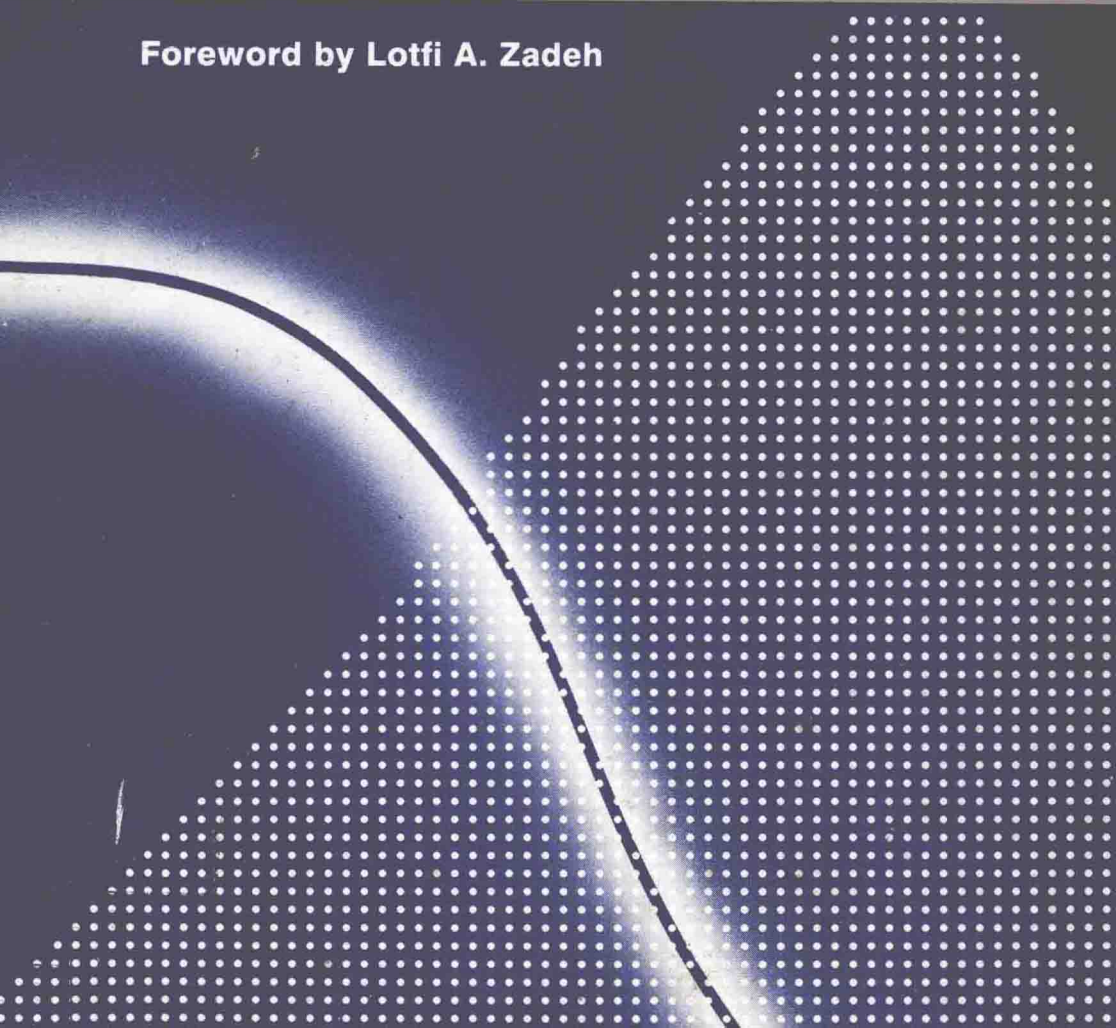
FUZZY SETS, NATURAL LANGUAGE COMPUTATIONS, AND RISK ANALYSIS

KURT J. SCHMUCKER

Foreword by Lotfi A. Zadeh



FUZZY SETS, NATURAL LANGUAGE COMPUTATIONS, AND RISK ANALYSIS

KURT J. SCHMUCKER
The George Washington Universitv

Foreword by Lotfi A. Zadeh

COMPUTER SCIENCE PRESS

Computer Science Press
11 Taft Court
Rockville, Maryland 20850

1 2 3 4 5 6 Printing Year 89 88 87 86 85 84

Library of Congress Cataloging in Publication Data

Schmucker, Kurt J.
Fuzzy sets, natural language computations, and risk analysis.

Bibliography:
Includes indexes.
1. Fuzzy sets. 2. Risk. 3. Language and languages.
I. Title.
QA248.S345 1984 511.3'22 82-23648
ISBN 0-914894-83-8

FOREWORD

The traditional approaches to risk analysis are based on the premise that probability theory provides the necessary and sufficient tools for dealing with the uncertainty and imprecision which underlie the concept of risk in decision analysis.

The theory of fuzzy sets calls into question the validity of this premise. More specifically, it suggests that much of the uncertainty which is intrinsic in risk analysis is rooted in the fuzziness of the information which is resident in the database and, more particularly, in the fuzziness of the underlying probabilities. Viewed in this perspective, then, it is the failure of classical probability theory to come to grips with the issue of fuzziness of data that limits its effectiveness in dealing with a wide variety of problem areas—including risk analysis—in which some of the principal sources of uncertainty are nonstatistical in nature.

In applying the theory of fuzzy sets to the analysis of real-world problems, it is natural to adopt the view that imprecision in primary data should, in general, induce commensurate imprecision in the results of the analysis. It is, basically, this view that motivated the introduction of the concept of a linguistic variable, that is, a variable whose values are not numbers but words or sentences in a natural or synthetic language. The theory of fuzzy sets provides a framework for dealing with such variables in a systematic way and thereby opens the door to the application of the linguistic approach in a wide variety of problem areas which do not lend themselves to precise analysis in the classical spirit.

Professor Lance Hoffman and his associate Don Clements were the first to explore the application of the theory of fuzzy sets—and, more particularly, the linguistic approach—to privacy, security and risk analysis. The present monograph is an outgrowth of this effort. It serves to introduce the reader to the theory of fuzzy sets and explains clearly and with many examples the use of the linguistic approach. Mr. Schmucker deserves to be complimented for presenting a coherent and self-contained account of a body of concepts and techniques which are of considerable relevance to risk analysis and natural language computations, and for contributing many insights which facilitate their application to the solution of practical problems.

L. A. Zadeh
Berkeley
April 1982

AUTHOR'S PREFACE

The intellectual task of analyzing the risk present in any large undertaking is an endeavor that abounds both with inherent imprecision and with a scarcity of historical data. Traditional mathematical and computational methods offer little to aid the analyst in work beset with either of these two difficulties, let alone work that is plagued by both of them. This is because the basic philosophical system upon which our mathematics and computer science is based is discrete and adheres strictly to the principle of the excluded middle: a statement must either be true or false. Unfortunately, this is rarely the case in risk analysis.

Fortunately, there is an alternative to this philosophy. This alternative, fuzzy set theory, is aimed at the development of tools for the solution of problems too complex or too ill-defined to be susceptible to analysis by conventional methods. This text provides the reader with an introduction to fuzzy set theory and explains one example of the use of that theory in risk analysis: the use of natural language expressions for the estimation of risk. An existing experimental automated risk analyzer which embodies these techniques is also described in some detail and future research directions are outlined.

This text will be useful to both students and professionals in a variety of disciplines and occupations: to the computer scientist it presents a readable introduction to a current topic in computer security and risk analysis and presents an application of the principles of abstract data structures considerably more involved than the stacks and queues usually presented in introductory courses; to the mathematician it presents an application of the results of an esoteric branch of mathematics, fuzzy set theory, to a practical problem that is becoming increasingly more important today; to the linguist it presents an application of the "linguistic approach" to a problem traditionally the forte of numerical scientists, as well as presenting a technique for the modeling of natural language expressions—a modeling which is both theoretically sound and experimentally verified. The common theoretical underpinning of these diverse fields is mathematics, and it is assumed that all readers will possess the mathematical maturity that is gained, for instance, from an undergraduate education in engineering or the physical sciences.

It is hoped that this text will be an introduction to the idea of automatic risk analysis utilities for those who require only an overview of this current research area, as well as a gentle introduction to the supporting literature for those who would extend the research frontiers. Both groups need to see "the big picture." Hopefully, this little text presents such a view.

K.J.S.
Washington, D.C.
November 1982

ACKNOWLEDGMENTS

The author wishes to acknowledge the contribution of both the scholarship and the camaraderie of the George Washington University Computer Security Research Group to the production of this text. That group provided an exceptionally fertile ground for the discussion of future plans, the correction of old errors, and the presentation of these ideas to the uninitiated end user. Special mention must also be made of the assistance of the group's leader, Lance Hoffman, and the group's other mathematician, Jerry Gaskill. Both suffered through the many drafts and supplied the author with numerous corrections and words of encouragement. The students of CSCI 229 ("Security & Privacy in Computer Systems") at The George Washington University used the manuscript version of this text in their course and provided me with many corrections and suggestions. Rich Atkinson spent a Christmas vacation poring over the final draft and suggested many rewordings and additions. Terry Ireland checked the syntax of the Pascal procedures and advised me on the best format style. My thanks to all for their assistance.

INTRODUCTION

The rigorous determination of the amount of risk associated with a particular proposed endeavor is a topic of both great practical and theoretical interest. It is of theoretical interest because it is an unsolved and difficult problem. It is of practical interest because the risks associated with many important projects are potentially serious and have ramifications throughout our society. Two examples where the determination of risk is both difficult and important are: (1) the risk to human life in a space launch and (2) the risk of compromise of personal data stored in a computer system. To be able somehow to reasonably estimate the risk to human life associated with as complex and multi-faceted an undertaking as a space launch would aid technicians and managers alike in evaluating tradeoffs for safety that must be made both before and during the launch. Similarly, to be able to meaningfully estimate the risk of compromise to confidential personal data in a computer system is essential in deciding on the security measures to be installed on the system and the security practices to be followed by its users.

At first glance, the determinations needed in these two examples appear extremely difficult, if not impossible. This difficulty lies in two completely separate phenomena: overall complexity and inherent imprecision. In the space launch problem, for example, suppose that you have been asked to estimate the risk associated with the launch *in toto* and that afterwards you have to decide how much safety equipment to purchase. The overall environment of a space launch is a complex arrangement of dependent interlocking events. The cognitive overload on a person who must estimate some important quantity based on data for the entire system is staggering. More often than not, a human is forced to neglect many facets of the total problem in order to delimit a manageable set of the data. Unfortunately, this can result in the ignoring of data ultimately important to the overall result, thereby providing a suboptimal (or even a totally wrong!) estimate. Such a suboptimal estimate can result in a considerable danger being overlooked or, alternatively, can force the use of unnecessary and costly safety equipment and procedures.

Even if the complexity problem was solved, the other problem of inherent imprecision remains to complicate the task of estimating risk. Suppose in the computer security example mentioned above that you have been asked to estimate the probability of one specific type of security failure: the unauthorized access of an intruder to the main computer room—a room whose only entrance is equipped with a cypher lock. (For those unacquainted with such a device, a

cypher lock is a mechanical device consisting of an array of ten buttons or toggle switches which, when a certain sequence of five buttons is pressed, opens a door. Such a device is used to limit access to a restricted area that has a heavy flow of traffic in and out. It is functionally similar to the more common key lock with the advantage that it is much easier to change the combination of a cypher lock than it is to change the key for a key lock.) If one forgets for a moment the case where one of the authorized individuals knowingly lets an intruder into the facility, the only case you have to consider is that in which somehow the intruder was able to get by the electro-mechanical device controlling access, the cypher lock. Since the cypher lock has 10 buttons, and since any combination that activates the cypher lock (thereby opening the door) is a certain sequence of five buttons, you could estimate that there are 100,000 possible combinations and that, therefore, the probability that an outsider might guess the right combination in one try is 1/100,000, in two tries 2/100,000, etc. Since your particular cypher lock allows only two incorrect tries before covering the presumed imposter in seven gallons of indelible yellow foam and sounding an alarm that would wake the dead, you feel pretty secure. Unfortunately, this estimate (and this entire methodology for estimating) neglects the time when someone spilled a Coke into the cypher lock and *any* combination opened the door, as well as the time when a disgruntled computer operator, having been covered in yellow foam the week before, painted the correct combination on the wall just above the lock *and* it took your ever vigilant security office more than two weeks to notice it!

The problem with your precise estimate of the probability of an intruder gaining access is that it possesses only a pseudo-accuracy—it looks great to the casual observer, but it fails to take into account perturbations that are *possible* in the real world—perturbations that are in some sense likely, taking into account Murphy's Second Law! ("If things can possibly go wrong, they will; if they can't possibly go wrong they still will—and in spades!" (This is also known as the Titanic effect.)) These real-world events are ignored in the "precise" analysis because it is unrealistic to calculate the *probability* of an intruder gaining access—there just isn't sufficient data for such a mathematically precise estimate. All one can reasonably estimate is the *possibility* or the *plausibility* of such an event taking place, given the information that you can have on hand or can reasonably assemble. Realizing this inherent lack of precise and complete data, it would seem (at least at first glance) that rather than estimating the probability of an intruder gaining access to your facility as .00162, it is really more accurate to say that the intruder's chance of success is 'EXTREMELY LOW'. In making this replacement of 'EXTREMELY LOW' for .00162, we are sacrificing the "precision" of the numerical estimate to gain the believability and confidence of an inexact, "fuzzy" estimate that is both more realistic and easier to interpret.

These two limitations to risk analysis, overall complexity and inherent imprecision, can be overcome to some degree if one has access to an *automated* risk

analysis utility that allows estimates of risk to be stated in *natural language terms* like 'VERY LOW,' 'MEDIUM TO HIGH,' or 'SOMEWHAT HIGH.' Both problems of overall complexity and inherent imprecision then become manageable. The automation present in this utility allows one to "simultaneously" consider a very large number of factors, something that would not be possible if you had to keep everything "in your head" or if you had to work by hand with the 2000 estimates for the risks associated with each of the 2000 components of the system. Thus, the apparent overall complexity of the risk analysis task is reduced to the job of estimating the risk of the individual components and then allowing the utility to combine these many estimates to produce the risk of the entire system. The other feature of the utility, that of accepting estimates in natural language terms, allows one to avoid the false precision that numerical estimates can provide, and it also allows one to form more reasonable estimates even with a paucity of data.

All of these benefits require that the automated risk analysis utility deal in an algorithmic fashion with natural language expressions in a way that is consistent with their use in ordinary discourse. This text describes such an automated risk analysis tool and provides the reader with the mathematical background necessary to understand the algorithms used to manipulate the natural language expressions. With this background and with the knowledge that psychological studies have demonstrated the "reasonableness" of this approach, the reader will then be prepared to knowledgeably use this automated risk analyzer.

CONTENTS

LIST OF ILLUSTRATIONS

Chapter 1

REVIEW OF SET THEORY

The theoretical foundations for the automated risk analysis utility that is to be described are in a rather specialized branch of modern mathematics: the theory of fuzzy sets. While one can study this theory at a very deep and mathematically sophisticated level, it is also possible to gain a great deal of useful insight at a more introductory and expository level. At this more elementary level, one can consider fuzzy set theory to be a generalization of ordinary set theory: the theory of collections of things. Much of the fundamentals of ordinary set theory are (or were!) the basis for the so-called "modern math" approach in elementary and secondary education and, therefore, are familiar to large numbers of people and can be grasped without much effort. For those who may have been away from such concepts for some time, as well as to gracefully ease into what will be for most a new topic, let us review some of the basic terms and ideas of ordinary set theory.

As is now known to most students of mathematics, one can consider the notion of a *set* as one of the most basic in modern mathematics. For our purposes we need only recall that a set is a collection of objects from some universe, U. If the universe is the natural numbers, we can form a set, A, composed of the numbers, 6, 222, and 376458. We would then write $A = \{6, 222, 376458\}$ and it would be exactly clear which numbers in the universe were in the set A (the *elements* of A) and which were not. When we specify a set by listing out all its elements as we did above for the set A, we have specified the set by *roster*. The only alternative is to specify the set by *rule*, i.e., to describe all its elements by some property or formula. A rule description of the set A would be the set of numbers that describe the length of this text in terms of the number of chapters, the number of pages, and the number of characters in the original manuscript. The rule method of specification is usually preferred in the case of infinite sets.

We could lump together the elements of two sets, taking their *union*, or we could examine the elements held in common by two sets, taking their *intersection*. We could also consider all the elements not in a set by taking its *complement*. It is assumed that the reader is quite familiar with such notions and with the conventional Venn diagrams for depicting these operations shown in Figure 1.1.

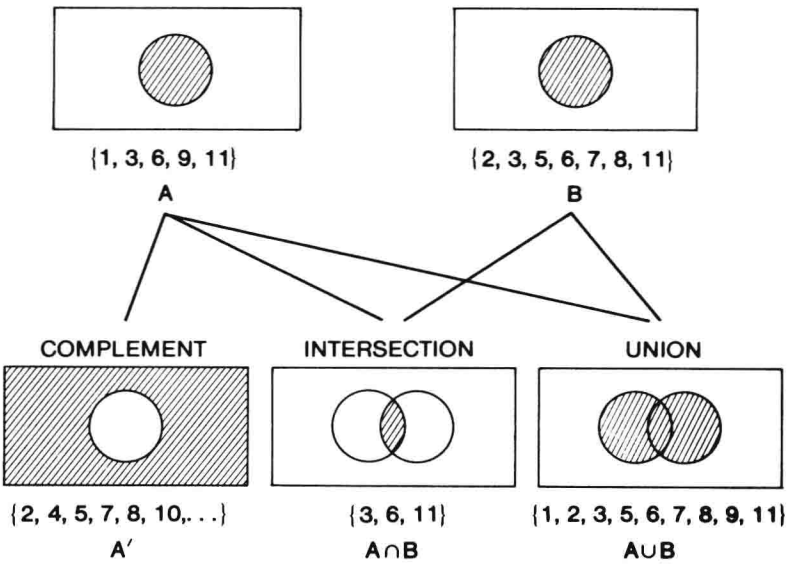


Figure 1.1 Set Union and Intersection.

One idea that may not be so familiar to the reader is a particularly precise specification for a set and the manner in which this specification can link set theory and logic. "It is common for logicians to give truth conditions for predicates in terms of classical set theory. 'John is tall' (or 'TALL(j)') is defined to be true just in case the individual denoted by 'John' (or 'j') is in the set of tall men" [Lakoff, 1973]. Hence, questions concerning logical reasoning can be reduced to a determination of set membership. The question of whether something is in a particular set can be answered through the use of a set specification method different from both the roster and the rule methods. We can, for any set, A, describe a function which determines for any element of the universe, whether that element is a member of A. Such a function is called the *characteristic function* of A, and is defined by:

$$char_A(x) = \begin{cases} 0 \text{ if } x \text{ is not in the set } A \\ 1 \text{ if } x \text{ is in the set } A \end{cases}$$

This function is defined for all the elements of the universe. It is a function mapping the whole of the universe U to the set of two elements $\{0, 1\}$. (We usually write this as $char_A(x)$: $U \rightarrow \{0\ 1\}$.)

With an identification of $\{0, 1\}$ and $\{false, true\}$, this characteristic function can also play a role in assigning truth values to statements about A. The most

elementary statement about A is one of the form "x is an element of A." In this case, the characteristic function also acts as a truth function: if x is an element of A, then $char_A\ (x) = 1 = true =$ truth_value ("x is an element of A").

These notions from set theory form the basis of modern mathematics, yet they seem rather inappropriate to our needs in risk analysis: they require too much precision - precision which we do not have and cannot obtain. It would also seem that any mathematical tools built upon this foundation would also inherit these deficiencies. We must use theories built on an entirely different base.

Chapter 2

FUZZY SET THEORY

With these preliminaries of set theory reviewed, let us propose a generalization of that theory. This generalization will be accomplished by suitably modifying the notion of *membership* in a set. What if an element was not completely *in* a set and was also not completely *out* of a set, but rather was half in and half out? Consider the following example:

$$A = \{x \mid x \text{ is a natural number and Mary's car can hold } x \text{ adult passengers}\}$$

and suppose that Mary's car is a Pinto. Then it seems safe to state that 0, 1, 2, and 3 are all elements of A and it seems equally safe to state that 7, 8, 9, ... are not elements of A. But what about 4, 5, and 6? Intuitively, 4 is *more* in A than 6 is, or more precisely, it is more plausible that 4 is an element of A than it is that 6 is an element of A. This notion of the plausibility of set membership (as distinguished from the probability of set membership [Kaufmann, 1977], [Zadeh, 1980]) leads to the generalization of the *degree of membership* in a set, and from this generalization comes a variant of the set theory discussed earlier; this variant is called *fuzzy set theory*.

A fuzzy subset of some universe U is a collection of objects from U (the *set* part) such that with each object is associated a degree of membership (the *fuzzy* part). The degree of membership is always a real number between zero and one, and it measures the extent to which an element is in a fuzzy set, or in ordinary set-theoretic terms, it measures the plausibility of an element being in a particular set. A degree of membership of 0 for an element of a fuzzy set corresponds to an element that is not in an ordinary set, and a degree of membership of 1 corresponds to an element which is in an ordinary set. Therefore, if the universe is the set $\{a, b, c, d, e, f\}$, then a fuzzy subset, A, of this universe could be defined as

a is present with degree of membership 1.0
b is present with degree of membership .9
c is present with degree of membership .2

d is present with degree of membership .8
e is present with degree of membership 1.0
f is present with degree of membership 0

Equivalently, A could be written

$$\{1/a, .9/b, .2/c, .8/d, 1/e\}$$

where the degree of membership is juxtaposed next to each element and elements with 0 degree of membership are omitted.

The exact relationship of the notion of a fuzzy set to that of an ordinary set can be seen most clearly when one recalls the definition of the characteristic function of a set. For an ordinary set A, the characteristic function is of the form

$$char_A(x) : U \rightarrow \{0, 1\}$$

but for a fuzzy subset A, it is

$$char_A(x) : U \rightarrow [0, 1]$$

where here the degree of membership function is the characteristic function. The characteristic function of a fuzzy subset, instead of mapping to the set of two elements (a binary choice of either being in or out of a set), is a mapping to a portion of the real line, allowing a continuum of possible choices. If the range of the characteristic function of a fuzzy set, A, (i.e., its degree of membership function, $char_A(x)$: $U \rightarrow [0, 1]$), is in fact restricted to just the two values of 0 and 1, then this function reduces to an ordinary characteristic function and A reduces to an ordinary, non-fuzzy set. We see then that fuzzy set theory contains ordinary set theory as a special case.

Before we continue in our discussion of the principles of fuzzy sets and in the extension of results from ordinary set theory to fuzzy set theory, it is worthwhile to examine the motivation for making this extension. While it is certainly sufficient in this regard to say that fuzzy sets are studied for the same reason n-dimensional, non-Euclidean geometry or any other branch of higher mathematics is studied, because it is there, the manipulation of fuzzy sets represents something more than mental gymnastics. The originator of the notion of fuzzy sets, Lotfi A. Zadeh, has stated:

> One of the aims of the theory of fuzzy sets is the development of a methodology for the formulation and solution of problems which are too complex or ill-defined to be susceptible to analysis by conventional techniques [Zadeh, 1980].