INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

EVALUATION OF INFORMATION RETRIEVAL SYSTEMS USING FUZZY SET TECHNIQUES

by

MORRIS YARMISH

A dissertation submitted to the Graduate Faculty in Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

1997

This manuscript has been read and accepted for the Graduate Faculty in Computer Science in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

July 14, 1997
Date

Jack Schwartz
Chair of Examining Committee

July 14 '97
Date

Executive Officer

Professor Linda Friedman

Professor Jerry Waxman

Professor Valery Frants

Supervisory Committee

The City University of New York

Abstract

EVALUATION OF INFORMATION RETRIEVAL SYSTEMS USING FUZZY SET TECHNIQUES

by

Morris Yarmish

Advisor: Professor Jacob Shapiro

In this thesis we explore an area that has only been touched upon until now - namely, the area of non-binary evaluation of information retrieval systems. We first explore it in terms of the traditional measures of R and P. In this part of the analysis, we describe carefully how to calculate R and P in the non-binary context, something that has never been explicitly addressed before in the literature. We suggest two separate methods of calculating R and P, and explain the differences between the two methods.

We then take the analysis further into the area of *composite* non-binary evaluation of information systems. We use fuzzy set techniques to address this problem and develop four non-binary composite

measures. The first measure is an extension of a suggestion by Voiskunskii to use a cos measure in the binary case to calculate to what degree two vectors coincide with each other. However, he shows that his suggestion is not extendible to the non-binary case. We use fuzzy set techniques and successfully make the transition to the non-binary case. The others are based on the fuzzy set technique of subsethood. We develop three measures based upon this concept. The measures are fully explored and contrasted in terms of their characteristics and advantages and disadvantages, especially in terms of the important order preservation property characteristic.

Acknowledgments

ברוך שהחיינו וקיימנו והגיענו לזמן הזה הטוב והטהיב

The author wishes to acknowledge all the assistance given to him by the members of the Supervisory Committee and the Executive Officer of the Phd Program in Computer Science, Professor Stanley Habib.

I would like first to express my deepest thanks to my mentor, Professor Jacob Shapiro, who gave willingly and unsparingly of his time, expertise, and wisdom, meeting with me on a weekly basis over a long period of time, to help me complete this project. Without his help this thesis would never have been written. Equally, my deepest appreciation to Professor Linda Friedman, who encouraged me not to give up when the going got rough, and who took time from her busy schedule to keep up with the work as it progressed, reading, commenting, and offering excellent advice at each step of the way. Without her input, this project would not have been completed. Also, my many thanks to Professor Waxman, who gave of his time to help. His encouragement and comments were a major asset to the project. Certainly no less than the others, my deepest appreciation to Professor Stanley Habib, Executive Officer of the Phd Program in Computer Science, whose encouragement and help throughout the time I spent in the Department, particularly at crucial times when it was most needed, enabled me to complete my Phd.

Lastly, and most of all, I would like to thank my wife, for her encouragement, patience and good sense in all aspects of our lives, and in this project in particular, and my children, who had to cope these last few years with a father whose attention was elsewhere.

To all the above, my lasting gratitude.

TABLE OF CONTENTS

LIST OF TABLES

Chapter I

Introduction

Information retrieval techniques have not been perfected to the point where a query yields all relevant documents and only relevant documents. It is unclear whether it is at all possible to achieve perfection in this area. Since we continue to have imperfect systems, we must evaluate to what degree each attains effectiveness, so as to compare them and choose the best one.

Traditionally, we have used recall to measure the percentage of relevant documents retrieved, and precision as an error measure, to measure the percentage of retrieved documents that are indeed relevant. The problem is that this pair of measures does not yield consistent results in that the two of them are inversely related - i.e., when R is high, P is low and vice versa. So , given two systems, one that has higher recall and the other that has higher precision, we cannot necessarily decide which is better. So one problem addressed in the literature, but never satisfactorily solved, is to find one composite measure, which will yield unequivocal results. There have indeed been a number of suggestions (Van Risjbergen, 1979, Voiskunskii, 1997), but none have been universally accepted.

In addition, relevance has always been evaluated in a binary manner - i.e., a document is either considered to be completely relevant to a query, or not relevant at all - there is no middle ground. As far as the system is concerned also, when a document is retrieved, the system either considers it totally relevant, or not relevant at all. In calculations of recall and precision, standard practice is to use binary relevance, and binary retrieval. This means that the denominator of the recall fraction, total number of relevant documents in the collection, will be a whole number - the number of documents in the collection that are relevant. So too the denominator of the precision fraction, the total number retrieved, will be the number of the documents retrieved by the system as being relevant. The literature (Miyamoto, 1990, Voiskunskii, 1997) has mentioned the possibility of evaluating relevance in a non-binary manner - that is, when evaluating the relevance of documents in the collection, having the latitude of being able to assign partial, or fractional, relevance to a document. The system would also be given this latitude, and upon its retrieving a document it would be able to assign partial relevance/retrieval to the document. In that case, the denominators of R and P would be summations of fractional values. However, considerations of this possibility have been few indeed, and this area has not been analyzed in detail at all. Moreover, when it comes to composite measures in the non-binary mode, the one attempt

(Voiskunskii, 1997) to adapt a binary evaluation measure to the non-binary case was unsuccessful.

It is these two problems that this thesis addresses. We will investigate the entire area of non-binary evaluation. First we will carefully analyze non-binary R and P. Then we will address the difficult problem of non-binary composite measurement, and suggest a number of solutions.

The structure of the paper is as follows. In Chapter II we explore in detail the background problems associated with information retrieval in general and evaluation of information retrieval systems. In Chapter III we discuss the traditional measures of R and P, the issue of composite measurement, and Voiskunskii's (Voiskunskii, 1997) composite measure. Chapter IV explains fuzzy logic concepts and the fuzzy mathematical techniques necessary for an understanding of our subsequent analysis. Chapter V deals with non-binary evaluation in general and describes in detail how non-binary R and P should be calculated (which is something that we believe has never explicitly been analyzed before in the literature). Chapter VI constitutes the major contribution of this thesis. It is an exhaustive exposition of non-binary composite measurement, including Voiskunskii's unsuccessful attempt to adapt his binary measure to the non-binary case, and our suggestions to solve the problem with four possible non-binary composite measures. We end the chapter with a discussion of the order preservation property

and which of our four measures possesses this important property. In the final chapter, Chapter VII, we conclude by summarizing our findings.

CHAPTER II

Background

Section 1

Basic Concepts - Information Retrieval and Information Retrieval Problems

Information systems can be divided into a number of different types. Database systems are one type. In database systems, information is inputted in a certain fixed structure, and, knowing the structure, or learning the structure from the system itself, the user can retrieve the information or facts he is looking for. If the system does not contain the facts he is looking for, i.e., if the fact is missing for some reason, the user will be able to immediately ascertain that this is the case, because that piece of information will not be in its proper place in the overall structure, and this is the only relevant location to look for it. In this case, evaluation of effectiveness of the information retrieval system is very simple. Indeed, there is really no question of the effectiveness of the retrieval mechanism itself at all. The retrieval

mechanism is a very simple one. It looks for a particular file, and a particular field in the file, for the information it seeks. If it cannot do that, it is totally ineffective. If it can successfully do that each and every time, and is still unsuccessful in retrieving the required information, then the fault lies not in the information retrieval system mechanism and its effectiveness, but in the data base itself - i.e., in the information base which is being accessed. The information was never provided to the data base and never inputted into the system. So, assuming the data base contains all the data necessary, a data base system is evaluated as either working or not working at all.

Moving on to less structured or unstructured systems, there are information systems which are referred to as "factographic information systems" (Voiskunskii, 1997). In these systems, the data is not inputted in a fixed structure. Yet they are designed to be used for searching for specific facts. In these systems, the users "information need" (IN) is a "concrete information need" (CIN) (Frants, 1988, Voiskunskii,1997). Here it is not the case that not finding a piece of information indicates that it is simply not there. The underlying base of information has no structure, and the user really does not know where to search for each piece of data. So the fact that a piece of data was not found may indeed be the result of a defect in the information retrieval mechanism and the technique and algorithm being used. However, evaluation of

effectiveness of the retrieval mechanism is still very simple. We use the "search success concept" (Voiskunskii, 1997). If the required information was found, then the search was successful, and if not, then the search was unsuccessful. There is no need for the system to find many instances of the same fact or every instance of the same fact - indeed, it is better if only one instance is retrieved, as one instance is perfectly sufficient and further instances are totally superfluous.

Lastly, there are systems referred to as document retrieval systems (Voiskunskii, 1997) designed to address the information need of a user having a "problem oriented information need" (POIN). The user has a problem and the problem needs information to help solve it. Again, as in factographic systems above, there is no designated location in any data base where the user knows he can find the information. So these systems are characterized by an unstructured information base. However, in this case, the user is not simply looking for one fact, where one instance of the fact is sufficient. On the contrary, the user has a problem oriented information in the sense that he has a problem and knows that information will help solve the problem. He is interested in marshaling as many facts and as much information as he can to shed light on the problem.

Additionally, problem oriented information need means that the search query is unstructured as well; unstructured in the following

sense. One aspect is that the user need does not have exact boundaries and hence cannot be expressed precisely in natural language (unlike the CIN case above, where the request is exact) (Frants and Brush, 1988, Frants and Shapiro, 1991). Secondly, the user is obligated to guess the best combination of descriptors, from the search point of view, to query the system, based on his experience and intuition, and generally without intimate knowledge of the information system and its structure and mechanism (Frants and Shapiro, 1991). These two thing degrade the clarity of the query. This, in conjunction with an unstructured information base, means that a search will never be perfect. This is especially the case if perfect means retrieval of all documents that are relevant to the user, and only those documents that are relevant to the user. Since the query itself is not perfectly clear, and the database is unstructured as well, it is not at all possible that perfect results can be obtained. Being that it is the case that results can never be perfect, we must then be able to evaluate the results in terms of how close to perfection they are - hence, the problem of evaluation.

These document retrieval systems are very common and very important. It is these systems, that is, the evaluation of these systems, which this thesis will discuss.

Section 2

General Evaluation

The question of evaluation has been extensively discussed in the literature (Cleverdon, 1970, Saracevic, 1995, Van Rijsbergen, 1979, Voiskunskii, 1997). We will follow Van Rijsbergen and introduce the entire subject from the standpoint of three questions. Why evaluate? What to evaluate? How to evaluate?

The answer to the first question is as follows. Whenever one uses a system, whether it is new or old, or contemplates implementing a new system, it is important, indeed imperative, to know how well the system works, and to what extent desirable results are being obtained from the system. This is important in terms of the system in question itself and also in terms of being able to compare the results from two systems to decide which is better. Closely related to this is the question of costs. Part of why we evaluate systems is to ascertain the benefits, and consider whether the desired benefits are worth the costs. So in evaluating information retrieval systems, we are mainly concerned with providing data so that users can make a decision as to whether the system provides reasonable benefits, and whether it will be worth the

cost. Any methods of evaluation that we use will also be used to measure whether certain changes will lead to an improvement in performance. When a claim is made for a particular search strategy, the evaluation methodology can be applied to determine whether the claim is valid. The second question, what to evaluate, is best answered from the standpoint of the user. That is, we will seek to measure how well the system meets the user's needs - i.e., those things which reflect the ability of the system to satisfy the user. The literature cites a number of measurable quantities (Cleverdon, 1966). Among them are:

- Time lag: The time lag is the average interval between the time the search request is made and the time the answer is given.
- Effort: the effort involved on the part of the user to obtain answers to his search request.
- Presentation: Presentation is the form of presentation of the output
- User friendliness: User friendliness is closely related to effort; how easy the system is to use.

All of these measures are important. Some of the measures are quantitative (time lag), and some are subjective (the others). Systems are indeed often evaluated in light of the above criteria. However, there are two other measures in terms of which systems are always evaluated - namely, recall and precision. We will elaborate on these measures below, but briefly, recall is the proportion of relevant material

retrieved in answer to a search request, and precision is the proportion of retrieved material that is actually relevant. Both these measures together constitute the effectiveness of the system. It is clear why these measures are by far the most important and the most commonly used. The main goal of the user is to retrieve the information he is searching for. If our goal is to measure how well the system meets the user's goals, then recall and precision are indeed those measures which evaluate to what degree the main goal of the user has been met. All the other measures are measures of auxiliary goals and secondary needs on the part of the user. In conclusion, what we measure mainly is the effectiveness of the system, traditionally represented by recall and precision. It is this effectiveness of the system that we will concentrate on in this paper as well.

The final question, how to evaluate, the basic way being by calculating recall and precision, is the concern of this paper. We address the major problem with using recall and precision (see below, Chapter III), and describe the need for another (composite) measure. Then we go on to address the need to evaluate information retrieval systems in a non-binary manner, which is the starting point of our research. So this thesis is concerned with the question of how to evaluate information retrieval systems in a non-binary manner via a composite measure.

CHAPTER III

Traditional Measures of Search Results - Binary Evaluation

Section 1

Precision and Recall

As we said, the standard way of measuring search results is in terms of two measures, recall and precision. Recall is defined as the proportion of relevant material retrieved, while precision is the proportion of retrieved material that is relevant. In other words, recall measures the proportion of the relevant items in the collection that were actually retrieved by the system, and is therefore a measure of the efficiency of the retrieval ability of the system. Precision, on the other hand, takes the total number of documents retrieved and measures the percentage of those retrieved documents that are actually relevant. It is therefore a measure of how accurate the system is in recognizing and retrieving only relevant documents. In keeping with these explanations, recall and precision are defined as follows:

$$R = \frac{number\ of\ items\ retrieved\ and\ relevant}{total\ \ relevant\ in\ collection} \qquad (3.1.1)$$

$$P = \frac{number\ of\ items\ retrieved\ and\ relevant}{total\ retrieved} \qquad (3.1.2)$$

Recall and precision are normally inversely related (Salton, 1983, Cleverdon, 1979). A system which exhibits a high recall, will have relatively low precision, and a system which exhibits high precision, will have relatively low recall. It is the same for queries: broad queries will have high recall and low precision, and narrow, specific queries will have high precision and low recall. It is herein that the problem lies.

How can we compare systems to determine which one is better? The system with high recall will have low precision - e.g., R might be .7 and P might be .2. How does this compare with a system with R=.4 and P=.75? Which is better? The first shows a higher R, and the second shows a higher P - what should be the proper tradeoff between R and P? It is because of this problem that researchers have attempted to develop some sort of "composite" measure, or in Voiskunskii's terms (Voiskunskii, 1997), some "complex search characteristic" (CSC). a single measure that would enable comparison of systems. A number of attempts have been made; none have been universally accepted. So one critical issue in information retrieval is to develop a single measure for evaluating search results.

Section 2

Composite Measurement

We must however discuss this issue of developing a composite measure or "complex search characteristic" in more detail. To restate the problem, the issue is the following. We do indeed already have evaluation measures for information systems - namely, precision and recall. The problem is, however, that precision and recall are inversely related to each so that normally when one is high the other is low and vice versa. Under these circumstances, how can we tell which is better, the one with the higher precision, and lower recall, or the one with the higher recall and lower precision? In order to get around this dilemma and be able to compare systems, we have to have one measure, not two inversely related measures; hence the search for a "complex search characteristic". Not only that, but the ideal CSC would probably be some function or combination of the known measures of P and R, as Voikunskii correctly points out (and goes on to show how his suggestions are truly a function of P and R) (Voiskunskii, 1997).

In looking for a way of combining P and R, one possibility might be to discover some "ideal" rate at which we would be willing to sacrifice some degree of P for some degree of R and vice versa.

However, no such ideal tradeoff rate of exchange is readily apparent. Another possibility is combining P and R into some measure which would make some intuitive sense. For example, the length and width of an object are combined into one measure called area - because area makes some intuitive sense in that it measures the expanse covered. So, too, we might look to discover some combining measure that would make some intuitive sense. However, no one would attempt to combine the height and weight of a person into one measure, perhaps calling it bulk, because it makes no intuitive sense, and it would be a very rare case indeed in which such a measure would find application, while height and weight alone are important measures in their respective contexts. Our case seems more similar to the case of height and weight than to the case of length and width in that no combining measure that makes intuitive sense comes readily to mind.

Our problem is not unique; we have a similar problem in the field of finance. In finance, investments are evaluated in terms of risk and return. The best investments are those which have the lowest risk for the highest return. However, in the real world, high returns come with high risks - i.e., the higher the risk, the higher the return. That means that, for example, one might be comparing one investment with returns of 10 percent and risk of .07 with another investment with 20 percent returns and risk of .12. Which is better? How much extra risk is each

percent of extra return worth? We cannot elaborate on the entire theory of portfolio management here, but the short answer to the question is the following. In the marketplace, investors can find investments with various different risk and return characteristics. At each level of risk, there will be one or more investments or portfolios which offer the highest return for that level of risk. The locus of all these maximum returns constitutes the "efficient frontier" or the "capital market line" (CML). The investor should choose only from among these investments, and not from among any others which offer lower returns for the same level of risk. In other words, the marketplace, or the real world, sets the risk-return tradeoff - the existing realities in the investment marketplace are the only choices open to the investor, he has no others. He then chooses among what is offered to him based on his own personal risk preferences. Each investor will position himself at a different risk-return point, and no one can say that any risk-return combination is superior to any other. They are all equal except with regards to personal risk preferences. The only thing is, for his personal level of acceptable risk, the investor should choose only from the efficient frontier, as these investments constitute the highest returns for each level of risk.

If we were to apply a similar model to evaluation of information systems, we would be saying that all precision-recall pairs are equal, as

long as the user has the highest R for that P class and/or the highest P for that R class - and that it depends on user preferences. If, say, precision is most important to the user, then he should choose the highest precision, and choose the highest recall he can get for that level of precision; if recall is most important to him, then he should choose the highest level of recall, and choose the highest level of precision for that level of recall. If the user wants 100 percent recall, for example, one way to achieve would be to retrieve all the documents in the collection. This, however, would yield the worst level of precision. Indeed, the precision would be the percentage of relevant documents in the entire collection. The user could easily improve upon this by just having some way of eliminating one of the irrelevant documents. He could further improve by eliminating a second irrelevant document. He could continue in this way until the next document to be eliminated is a relevant one, and then he would stop. In other words, in order to get 100 percent recall, the user would find the highest level of precision for that level of recall - and he would not have to accept the worst level of precision yielded by retrieving all documents. If the user would be satisfied with less than 100 percent recall, then he could continue eliminating documents well into the area where relevant documents are also being eliminated and stop when recall degrades to the point beyond which he would no longer be satisfied with the recall level. In general,

as we say, the user would search for the highest level of precision for that level of recall that he has set as his goal. If he has precision as his main goal, he would perform similarly, setting the level of precision at, for example, 100 percent by retrieving one relevant document. He would then continue to retrieve until the precision degrades below 100 percent or below the level that he has set for his goal, just as by recall above.

However, in information retrieval system evaluation, this approach would not work, and indeed no expert in the field ever suggested such an approach. In the area of investments, all possible investments are available to the investor, any one of the risk-return combinations on the efficient frontier, and all he has to do is choose one. Different investors, depending on their risk preferences, will choose differently. In information retrieval systems, the user cannot choose in this way. The managers of the collection must choose one information retrieval system, one set of algorithms, one approach, to use to access the collection of documents. All users of the collection are limited to the one system chosen.

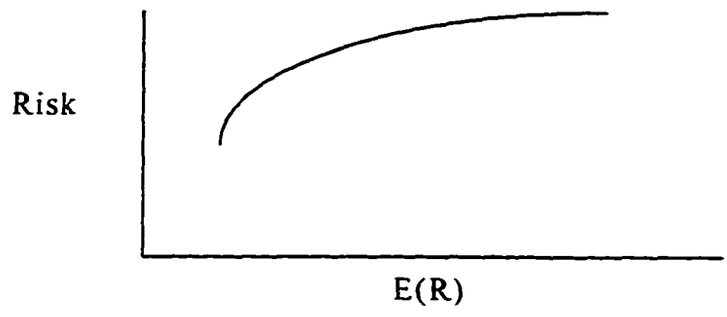
To make information retrieval systems analogous to financial markets, we would have to have a situation in which every document collection could be accessed by any one of many IR systems, any one of many algorithms, each with different P-R characteristics; then each user

could choose the P-R pair he is comfortable with. Perhaps sometime in the future we will have the resources and/or the technical ability to do that, but today each collection must choose one IR system, and therefore we need one measure to help managers evaluate systems and choose the best one.

Another difference between financial markets and information systems is the following. In investments we are trying to find the rate of exchange, the rate of trade-off, between risk and return - i.e., how much additional return should be given for each additional degree of risk. In fact, the efficient frontier and the capital market line tell us just that. The consensus of investors in the real world creates the tradeoff. There is a marketplace, the real world, where all investors "vote", and the consensus of these votes form the "efficient frontier" and the slope of the "capital market line" as the proper tradeoff conditions between risk and return. (We are simplifying a bit here. In fact, it is not investors that establish the efficient frontier and the CML, but the real world - i.e., the actual investments in the real world. The real world, in a sense, is the supplier of the risk-return choices, and the investors are the potential buyers. (This will be discussed in more detail in the next paragraph.) However, the real world and its characteristics, including its risk-return characteristics, is not formed completely independently, but is at least partially, and perhaps mostly,

formed by people. These people are the people involved in the business world - i.e., investors. That is to say, investors, in fact, are not only buyers of investments, but are also suppliers of those investments. The marketplace, or the community of people in the marketplace, consists of businessmen who are sometimes sellers and sometimes buyers of investments. That is, the same individuals are sometimes sellers and sometimes buyers, and they form the characteristics of investments. So we can in fact say that investors themselves establish the efficient frontier and the CML. Moreover, clearly, as a seller of an investment, the businessman is forced to form the investment in such a way as to be attractive to the potential buyer (whom he, in fact, knows intimately, because it is sometimes he himself), or else no one will buy. So it is the case that investors, and their opinions, form the efficient frontier and the CML.) The individual investor can choose any point along the CML as his personal risk-return preference. In fact, at a certain level of risk, an individual investor will stop from going to higher levels of risk. Why? Because as far as his personal preferences are concerned, he disagrees with the consensus of the market, and does not believe the extra returns beyond this point are worth the extra risk. In information retrieval systems, however, there is no marketplace where users vote, and therefore we cannot ascertain what the proper consensus tradeoff obtains between precision and recall - hence the search for a CSC.

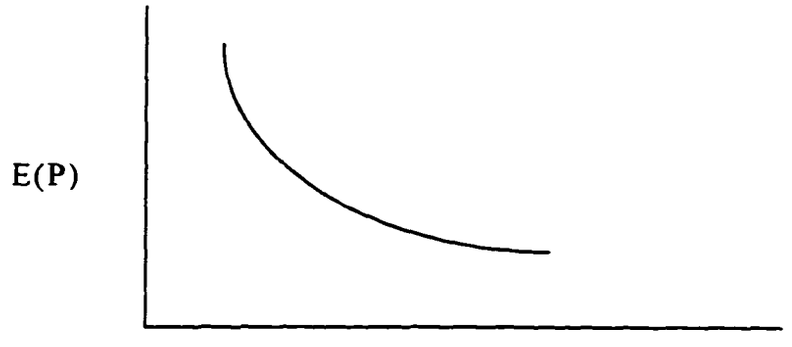
For those who wish to go deeper into the question, still another difference between financial markets and information retrieval systems is the following. Assume that investments in the real world are indeed formed, exclusively or chiefly, independently from the objective reality of the opportunities available along with their respective risk-return characteristics. To get the efficient frontier, we plot the investments on a risk-return graph and choose the highest returns for each level of risk, as in the diagram below



Note that the horizontal axis is E(R) - expected returns, as calculated by the summation of the probabilities times each return and the vertical axis is risk as established by some standard measure of risk. for example, the standard deviation of the returns around the expected returns. The highest returns will be along the curve called the "efficient frontier". (The CML (capital market line) is an addition to the analysis which introduces the concept of a "riskless asset" and how including the riskless asset in the portfolio can raise returns still higher for each level

of risk - higher than the pure efficient frontier. The underlying concept and analysis remain substantially the same, and further exposition of the concept of the CML is beyond the scope of this paper, and really not important for our purposes here.) Investors would be best served by choosing somewhere along the efficient frontier for the obvious reason that their returns will be maximized for each level of risk given the state of the real world as it exists.

We might wish to do a similar thing for information retrieval systems. We would calculate the recall-precision characteristics of many different algorithms and systems. We would plot them on a graph, establish an efficient frontier and advise the user to choose only from systems along the efficient frontier, as in the diagram below.



The horizontal axis represents expected recall (E(R)), calculated by averaging recall over many queries, and the vertical axis represents expected precision, calculated by averaging precision over many queries. This method will maximize the user's results, as he will be

getting the highest recall for each level of precision and the highest precision for each level of recall, depending on his main preference. That is the proposal. However, this is not what we are looking for in information retrieval systems.

`In economics and finance, the objective is to economize or maximize the results for the given reality - there is no objective to try to find out whether the reality itself is the best we can get. The objective is simply to maximize given the reality as it is. In general, economics and finance take this modest objective as their purpose; objective truth is beyond their scope. In our case, as information scientists, we are not economists, and do not take this modest objective as our objective - we aim higher. Were we to take the point of view of the economist, our proposal would be perfectly reasonable, if impractical in light of today's state of technology, as we said above. However, what we are looking for is some objective tradeoff value between recall and precision. That is, how much recall for how much precision should the user accept - how much recall is objectively worth how much precision. Hence the search for a composite measure - for an objective tradeoff between recall and precision, or some composite measure which reflects the objective tradeoff between recall and precision.

Section 3

Voiskunskii's Composite Measure / Complex Search Characteristic

There have been a number of attempts to develop a composite measure to evaluate information systems (Van Risjbergen, 1979, Voiskunskii, 1997). The most recent has been Voiskunskii, who suggested an approach that we will outline here. We concentrate on Voiskunskii's approach because it is his measure that we will start with to extend and modify in order to arrive at a composite measure that works for non-binary evaluation.

Here is Voiskunskii's approach. Let us assume that a search is performed on a collection of N_0 documents. The collection and results can be represented in the following conjugate table:

	pertinent	non-pertinent	
retrieved	r	l	N=r+l
non-retrieved	b	d	M=b+d
	C=r+b	L=l+d	N_0=r+l+b+d

Table 3.1 Conjugate Table

This table shows that the collection can be divided in two ways: retrieved documents vs. non-retrieved documents and pertinent documents vs. non-pertinent documents. (The table also gives convenient symbols for each of the parts of the collection. Since we are using Voiskunskii's model as our starting point, we will also use these symbols in our subsequent formulas throughout this paper.) Thus two sets of evaluations are produced, namely, a set of evaluations of retrieved/non-retrieved documents, and a set of evaluations of pertinent/non-pertinent documents. These evaluations are normally done either by the user or by independent experts. (Please note that we are using the term "pertinent" in place of the term "relevant" following Voiskunskii's terminology. They mean the same thing and will be used interchangeably throughout this paper.) These two evaluations can be represented by vectors in the following manner. Let k be the vector representing the collection in terms of pertinence and non-pertinence.

$$k=(k_1,k_2,\ldots k_{N0})$$

Each document is represented by a k_i, where $k_i=1$ if the document is evaluated to be pertinent, and $k_i=0$ if the document is evaluated to be non-pertinent. Let v be the output vector in terms of retrieved and non-retrieved.

$$v=(v_1,v_2,\ldots v_{N0})$$

Again, each document is represented by a v_i, where v_i=1 if the document has been retrieved, and v_i=0 if the document has not been retrieved. Now the closer these two vectors are to each other, the more closely the retrieval matches the actual pertinence. In the extreme, if the two vectors coincide, then the system has done a perfect job - i.e., it has retrieved all and only those documents that are actually pertinent. So the composite measure we can use (or in Voiskunskii's terminology, the complex search characteristic or CSC - we will be using composite measure and CSC interchangeable throughout this paper as well) will be some measure of vector closeness; i.e., some function which measures the closeness of (two) vectors. A standard measure of closeness of vectors is the cos measure, which measures the cos of the angle between (two) vectors. In our case, this becomes:

$$\cos\varphi_{KV}=\frac{\sum_{i=1}^{No} k_i\, v_i}{\sqrt{\sum_{i=1}^{No} (k_i)^2}\sqrt{\sum_{i=1}^{No} (v_i)^2}} \qquad (3.3.1)$$

The closer this value is to 1, the smaller the angle and the closer the two vectors are to each other; and the more distant it is from 1, the larger the angle, and the less close the two vectors are to each other.

Voiskunskii then goes on to show what the measure reduces to in terms of recall and precision. The number of ones in vector k is equal

to the number of pertinent documents in the collection; i.e., in the symbols of the conjugate table above, to C. Consequently,

$$\sum_{i=1}^{No}(k_i)^2 = C$$

The number of ones in vector v is equal to the total number of documents retrieved; i.e., again in the symbols of the conjugate table above, to N. Consequently,

$$\sum_{i=1}^{No}(v_i)^2 = N$$

The number of positions having a one in both the k vector and the v vector is equal the number of documents both pertinent and retrieved; i.e., in the symbols of the conjugate table above, to r. Consequently,

$$\sum_{i=1}^{No}(k_i\, v_i)^2 = r$$

Then

$$\cos\varphi_{KV} = \frac{\sum_{i=1}^{No} k_i\, v_i}{\sqrt{\sum_{i=1}^{No}(k_i)^2}\sqrt{\sum_{i=1}^{No}(v_i)^2}} =$$

$$\frac{r}{\sqrt{C}\sqrt{N}} =$$

$$\sqrt{\frac{r^2}{CN}} =$$

$$\sqrt{\frac{r}{C}\frac{r}{N}} =$$

$$\sqrt{R \cdot P} \qquad (3.3.2)$$

Thus says Voiskunskii, the cos measure reduces to a well-known measure already mentioned in the literature as a possible composite measure. This gives the cos measure further credence. This measure, just like standard recall and precision, is designed to be a binary measure. What we mean by that is that retrieval and pertinence are evaluated as only either a 0 or a 1 - there is no possibility of partial pertinence or partial retrieval relevance. It is clearly so in light of Voiskunskii's derivation of $\sqrt{R \cdot P}$. He is only able to reduce the cos measure to $\sqrt{R \cdot P}$ based on the fact that $\sum_{i=1}^{No} (k_i)^2 = C$, which will only be the case if the k_i's are either 0's or 1's but not any value in between. as we will explain in more detail in Section V.

Having a binary composite measure, Voiskunskii then attempts to use this measure for non-binary evaluation (i.e., evaluation where pertinence and retrieval can be given fractional values, as well as the values 0 and 1) as well, and fails. Indeed, he is unable to find any non-binary composite measure. This failure is the subject of this thesis. We suggest some non-binary composite measures. These suggestions will be discussed in detail in Chapter VI. First, however, we must explain the mathematical background of fuzzy logic and fuzzy sets in Chapter IV and the non-binary concept in general in Chapter V.

CHAPTER IV

Fuzzy Logic and Fuzzy Set Techniques

Section 1

The Fuzzy Logic Concept

The difference between fuzzy logic and standard crisp logic is in the treatment of the concept of class (McNeill and Freiberger, 1993). Classification and categorization are basic to human thinking. Indeed, theorist David Marr suggested that handling classes is the paramount role of the neocortex, the gray matter of the brain (Marr, 1970). If we human beings did not classify , we would be hopelessly lost in detail and would not be able to make sense of the world. Learning is classifying. We see a book - it is unique. We see another, and start to realize that both are particular instances of a general class of objects. From then on, whenever we see a similar object, we place it in the general class of books, and can, for many purposes, ignore the particular characteristics of each one. Thus we can make sense of the world - if we see a member of the class book, we know it belongs on a shelf - and we know its

general relationship to other objects in our world. If we did not classify the objects in our universe, then each one would be unique, and we would be lost in detail and have no idea of the relationships of objects to each other. Even animals classify. A cat sees a small gray object run across the floor, realizes it is a mouse, and begins to chase it, even though it has never seen that particular mouse before. By realizing that the object is a mouse, the cat can tap memory and call up information about it (McNeill and Freiberger, 1993). Without classification, the cat would be totally at loss, and would be unable to function in its world. Human beings, too, if not for classification, would be unable to make sense of the world in which they live, and would be unable to function.

Mathematicians and logicians have studied and formalized the concept of class. Georg Cantor was the first; he developed the concept of class into what is today known as set theory. Cantor defined a set as a collection of definite, distinguishable objects. Hence, a class is a set. Cantor's sets are crisp. Each object in the universe is either a member of a particular set or it is not - none straddle the line. For example, if we want to define the set of all tall men, we establish a cut-off point; e.g., anyone who is 6 feet tall or above will be considered tall, anyone under is not. Sets are often depicted using Venn diagrams - and this depiction mirrors the in or out nature of Cantor's sets; either the object is in the circle or outside of it.

This crisp delineation of an object being in or out did not start with Cantor. It began with Aristotle (McNeil and Freiberger, 1993). Aristotle developed and codified the rules of logic which are still generally accepted till today. He used the mathematics of Pythagoras as his model and extended the step-by-step method of geometry and geometrical proofs to reasoning in general (McNeill and Freiberger, 1993). Realizing that just as in geometry, you have to start somewhere - with axioms - just so in logic. Aristotle declares, "It is not everything that can be proved, otherwise the chain of proof would be endless. You must begin somewhere, and you begin with things admitted but undemonstrated." (Aristotle, 1966). Thus Aristotle begins with truths so obvious that we accept them without proof. These were the axioms of logic (McNeil and Freiberger, 1993).

The first axiom is the Law of Contradiction. In the Metaphysics, Aristotle expresses it as follows. "The same thing cannot at the same time both belong and not belong to the same object and in the same respect. This is the most certain of principles." (Aristotle, 1966). In other words, A cannot be both B and not-B. The second axiom is the Law of the Excluded Middle (or the Law of Bivalence). Again in Aristotle: "Of any subject, one thing must be either asserted or denied." (Aristotle, 1966). In other words, A must be either B or not-B. These two laws are similar, but not identical. The Law of Contradiction

disallows true and not-true at once. The Law of the excluded Middle disallows anything other than true or not-true. Hence, not only can't A be both B and not-B, it must be either B or not-B (McNeill and Freiberger, 1993). (logic, by the way, gains by these two axioms a method of proof that turns out to be very useful - namely, reductio ad absurdum. This is an important advantage (McNeil and Freiberger, 1993).) Cantor, in his set theory, accepted these laws of logic and posited that an object is and must be either a member of a particular set or not; there is no middle ground. So standard set theory based on standard (Aristotelian) logic is black and white, in or out, 1 or 0, bivalent.

Fuzzy logic and fuzzy set theory denies the yes or no attitude outlined above. The fuzzy principle states that everything is a matter of degree (Kosko, 1993). The classic case is the example of tall men mentioned above. Where standard set theory would consider a person either a member of the set (of tall men) or not all, fuzzy set theory would attribute graded membership to each person in the set. For example,

Person	Height
A	6 ft. 6 in.
B	6 ft. 1 in.
C	5 ft. 9 in.
D	4 ft. 11 in.

Table 4.1 - Tall Men

Where crisp sets might consider 6 feet as the cut-off point for tall men, putting A and B in the set, and C and D outside it, fuzzy sets would establish the following graded membership

Person	Height	Membership in Set
A	6 ft. 6 in.	.95
B	6 ft. 1 in.	.9
C	5 ft. 9 in.	.8
D	4 ft. 11 in.	.25

Table 4.2 - Fuzzy Membership in Set of Tall Men

The justification of considering things in this way is that this is the way it is in the real world. Nothing in the real world is absolute - everything is a matter of degree - everything exists on a continuum.

Modern fuzzy logic was developed and popularized by Dr. Lofti Zadeh, first in a seminal paper (Zadeh, 1965), and then in many more additional writings. However, it does have precedents. The first inkling can be found in the early Greek philosopher, Zeno. He developed a series of paradoxes, one of which became known as *sorites*, the paradox of the heap (McNeil and Freiberger, 1993). Imagine a heap of sand. Take away one grain of sand and you still have a heap. Take another from it, and it still remains a heap. Eventually, one grain is left. Is it still a heap? Remove it as well, and you have nothing; is it still a heap? If not, when exactly did the heap cease being a heap? At

which grain of sand? Other paradoxes along the same lines have subsequently been developed (McNeil and Freiberger, 1993). For example, Wang's Paradox. If x is a small number, x+1 is also small. Then so is x+1+1. Therefore, five trillion is a small number, and so to infinity. When does the number pass from being small to being not small? If we set an arbitrary point, say 30, where 30 and below is small, and anything larger is not small, then it becomes intuitively difficult to justify that 30 is small, and just a bit above 30 is no longer small. Fuzzy logic does away with these kinds of paradoxes by attributing to heapness and smallness degrees; and the heap and the small number pass gradually from 100 percent to 0 percent.

Plato also saw fuzziness and degrees (McNeil and Freiberger, 1993). He realized, for example, that no chair is perfect. It is only a chair to a certain degree. If it is only partly a chair, it must also be partly not as chair. But that is a contradiction. Can a contradiction exist? Then no chairs exist. Then what is real? So he developed his Theory of Ideals. There exists an Ideal Chair - as, indeed, there exists an Ideal Everything. It is 100 percent chair. It is perfect. It exists in our minds from birth, and can be accessed by thought alone. Experience in the "real" world is delusion, and only the Ideals are eternal and changeless, the only true knowledge available.

Aristotle himself admits to degrees in places other than in Metaphysics - namely in DeInterpretatione, where he uses such terms as truer and falser. Even in Metaphysics, a few pages before he formulates the Law of the excluded middle, he says, "The more and the less are still present in the nature of things", and adds that one who thinks four equals five is more correct than one who thinks four equals one thousand. In DeInterpretatione he presents arguments against his own axioms. The most famous is in Chapter 9, where he questions whether the statement, "There will be a sea battle tomorrow", is true or not true. He appears, by some interpretations, to conclude that it has a intermediate truth value, between true and false. This question was a matter of tremendous debate among logicians and mathematicians during the Middle Ages.

Ancient Eastern philosophy embraced fuzziness and espouses contradiction. (Therefore fuzziness is part of the culture of Orientals. This may be the reason why fuzzy techniques are much more accepted and developed in Japan than in the West. Furthermore, just as fuzziness is so much a part of eastern culture, Aristotle is so much a part of Western culture. Therefore we in the West find fuzzy logic strange.) Jainism developed a distinctive logic in which there is no certainty, no one sees everything, truth is many sided and thus every statement is partly true and partly false. This is best illustrated by their parable of

the blind men and the elephant. A group of blind men encounter an elephant and examine it by hand. One feels the ear, and declares it a fan, one the tail, and declares it a rope, one the leg, and declares it a pillar, etc. All are partly correct, none is fully correct. So in reality; no one sees everything, and each statement is partly true and partly false.

Buddhism resembles Jainism in many ways - some suggest they have a common base. Buddhism is more popular than Jainism and in fact has spread throughout the East. It has many fuzzy elements. Kosko (Kosko, 1993) says that Buddha was the first fuzzy theorist. The Buddhist notion of wisdom refers to a deeper entity than reason, that unites, rather than analyzes, as our reason does, and therefore embraces paradoxes. The flower is red and not-red, A is not-A and therefore A is A; these are some of the pronouncements of Buddhism. There are other Eastern philosophies that embrace paradox and vagueness. Therefore, this notion is pervasive in Eastern thought and has found expression in the partial contradiction of yin and yang.

Leaving the ancient world, the first modern thinker to deal seriously with vagueness was Charles Sanders Pierce (McNeil and Freiberger, 1993). He contributed to many disciplines, including Boolean logic, experimental psychology, map projection, and mathematical economics. He held that everything is continuous and that

"the sheep and the goat separators", who divide the world into true and false, are wrong. Size is continuous, as *sorites* shows; so are speed, weight, distance, all sorts of intensities, and consciousness itself. Vagueness is ubiquitous and can "no more be done away with in the world of logic than friction in mechanics". He faulted logicians for giving vagueness "the go-by", and claimed to have "worked out the logic of vagueness with something like completeness", but we cannot find where.

Jan Lukasiewicz, a Polish logician and mathematician, took the first step towards a formal model of vagueness (Lukasiewicz. 1970). He invented the first multivalued logic. In it, 1 stood for true, 0 for false, and $\frac{1}{2}$ for possible. A statement could have any of these values. In this logic, "It will snow tomorrow", has a truth value of $\frac{1}{2}$, "It will not snow tomorrow", has a truth value of $\frac{1}{2}$, and so statement = not-statement, or A=not-A. He went on to say that we could insert any number of extra values , with each representing the degree of truth attributed to the statement. This would be superior to the three-valued logic, as it would indicate greater precision. It could quantify degrees of truth. It also preserves binary logic - at the extremes there is still 1 and 0 - but it adds to it. So any statement can have any degree of truth. What is negation then? The negation of a part truth is the additional

amount necessary to make it an entire truth. So if A is .6 true, its negation, not-A, is .4 true. After Lukasiewicz multivalued logics proliferated, as other logicians and mathematicians developed their own multivalued logics.

Max Black was the next to advance towards a full-blown discipline of fuzzy logic. He outlined what one author (McNeil and Freiberger, 1993) calls "proto-fuzzy sets". He said vagueness stems from a continuum, and every continuum implies degrees. It can be continuous or discrete, but if the intervals are small enough, they will escape notice. for example, dialects of German change slightly from village to village. The difference is negligible from one to the next, but it slowly accumulates until speakers at opposite ends of the country cannot understand each other. Or imagine an exhibit in the Museum of Applied Logic. A line of chairs is exhibited. At one end is a Chippendale, next a near Chippendale very similar to the first, and succeeding chairs are less and less chairlike, finally ending in a lump of wood. The observer has great difficulty in drawing the line between chair and not-chair, in deciding where chair ends and not-chair begins. The better thing to do, and the thing which approximates reality more closely, is to pin a number to each item, indicating its degree - Black said, degree of usage (not, as Zadeh later argued, degree of graded membership or degree of truth, i.e., degree of chairness). What this

number would mean is the percentage of people who would call the item a chair. It is a probability; the percent of people who call it a chair and thus the probability that any random person will call it a chair. In an appendix, Black suggested that these vague terms could form sets and subsets - e.g., fewer people will call any object a Chippendale than a chair, so Chippendale is a subset of chair.

Section 2

The Mathematics of Fuzzy Sets

As we said, it was left to Lofti Zadeh (Zadeh, 1987) to develop fuzzy logic and fuzzy sets to their full scope. He wrote the original paper and continues to explore it, popularize it, and expand its horizons, and has single-handedly been successful in gathering others to its banner. The key insight was that of graded membership or degrees of membership. As we explained above in the example of tall men, an object need not be either a member of a set or not at all, rather each object is a member of a set to some degree, on continuum from 0 to 1. He then expanded this concept into a full analysis of these kinds of sets and the operations that can be performed on them.

Fuzzy sets actually include the crisp sets of Cantor (McNeil and Freiberger, 1993). A crisp set is simply a fuzzy set , but limited to membership values of 0 and 1. The operations that we can do with fuzzy sets are the same operations that we can do with crisp sets, but they result in different calculations and different values. The following are the basic operations we can do with fuzzy sets.

- The empty set: Which sets have no members? A fuzzy set is empty only if all of its candidates have zero membership.

- Complement: The complement of a fuzzy set is the amount the memberships need to reach 1 - the amount that completes or fills the set. Referring to Figure 4.2 above, the set of tall men, the complement of this set, or the set of not-tall men is

A	.15
B	.10
C	.20
D	.25

Table 4.3 - Set of Not-Tall Men

- Containment: How do we define the subset of a set? In crisp sets, the definition is simple - a subset is a set such that all of its members belong to a larger set. In fuzzy sets, we follow Black's intuition (Black, 1937) by defining a subset as follows. A set is a subset of another set if all of its members have graded membership no greater than the graded membership of the superset. So the following set would be a subset of the set of tall men, Figure 4.2.

E	.8
F	.75
G	.6
H	.4

Table 4.4 - Subset of Set of Tall Men

- Intersection: Membership in both of two sets. In crisp sets, an intersection is the members that two sets share. In fuzzy sets. it is the degree of membership that the two sets share. The most commonly used definition for fuzzy intersection (and the one we will use) is the minimum membership values in both sets (Yager, 1994). So

$$A \cap B = \min(A,B).$$

For example, given the following fuzzy sets, one of fat men

A	.8
B	.75
C	.6
D	.4

Table 4.5 - Set of Fat Men - Set Y

and one of tall men (from Figure 4.2)

A	.95
B	.9
C	.8
D	.25

Table 4.6 - Set of Tall Men - Set Y

the set of tall and fat men equals

$$X \cap Y = \min(X,Y) = (.8,.9,.7,.25)$$

In other words, A is both tall and fat to degree .8, B to degree .9, C

to degree .7, and D to degree .25.

Note that this rule satisfies crisp sets as well. In a crisp or standard sense, if John is not tall (0), but is fat (1), we take the minimum of the two (0) and give zero membership to the intersection. So fuzzy intersection includes crisp intersection, but says more (McNeil and Freiberger, 1993).

There are indeed other definitions of fuzzy intersection (Yager, '1994, Klir, 1995).

$$A \cap B = AB$$

$$A \cap B = \max(0, A+B-1)$$

However, these are not the standard definitions, and, for our purposes, when we considered them in our research into evaluation of information retrieval systems, they did not yield workable results.

- Union. Membership in either of two sets. In crisp sets, the union is the members that are in either set, so in fuzzy sets, it is the degree of membership that is in either set. The most commonly used value for union is the maximum value in both sets. So

  $$A \cup B = \max(A, B)$$

  For example, the set of tall and fat men (from Figures 3.2 and 3.5 above)

  $$X \cup Y = \max(X, Y) = (.95, .95, .8, .5).$$

  In other words, A is fat or tall to degree .95, B to degree .95, etc.

Again, this rule satisfies crisp sets as well. In a crisp sense, if John is not tall (0) but is fat (1), we take the maximum of the two (1) and give a membership of one to the union. So fuzzy union includes crisp union, but says more.

There are also other definitions of Fuzzy union (Yager, `1994, Klir, 1995).

$$A \cup B = A + B - AB$$

$$A \cup B = \min(1, A + B)$$

However, max is the most commonly used.

Section 3

The Concept of Subsethood

Zadeh and others went on to broaden and deepen the subject of fuzzy logic. However, only one development is of interest to us here, since we will use it later on, namely, the concept of subsethood (Kosko, 1993). Fuzzy sets are sets whose elements are partially contained in the set. In our example of tall men, each man in the set is only partially tall. Thus, in some sense, the set is not fuzzy, but its elements are. We can refer to this type of fuzziness as *elementhood.*

What about sets - that is, sets containing other sets? Can the containment be partial? Kosko said it can. Some sets contain other sets fully, others only partially. The degree to which one set contains another set would be fuzzy containment or *subsethood.* So subsethood holds between sets - just as standard fuzzy theory or elementhood holds within sets.

Kosko uses this concept of subsethood to explain probability. Traditionally, the intuitive explanation of probability is relative frequency of occurrence. Furthermore, it is not clear that fuzziness is a new concept - some people argued that fuzziness is simply probability and has nothing new about it. Kosko turned the entire question on its

head. He argued that just as a whole can and does contain a part, a part can be conceived of containing a whole partially, to some degree. The degree to which the part contains the whole is the probability. To see this, we must consider what it is that we mean by the probability of success. It is no more than the degree to which all trials are successful, or the degree to which the set of successful trials contains the set of all trials. So we have another intuitive explanation of probability, and more importantly, have an explanation that shows that fuzziness is not probability, but that probability is fuzziness. In any case, this further explains what we mean by subsethood.

Let us now see how subsethood is defined mathematically. First we must define the concept of scalar cardinality. For any fuzzy set A defined on the universal set X, we define its scalar cardinality, |A|, by the formula

$$|A| = \sum_{x \in X} A(x)$$

For example, the scalar cardinality (or sigma count) of set X, where

$$X = \{.3, .2, .9, .7\}$$

is

$$|X| = .3 + .2 + .9 + .7 = 2.1$$

For any pair of fuzzy subsets defined on the universal set X, the degree of subsethood, S(A,B), of A in B is defined by the formula

$$S(A,B)=\frac{1}{A}(|A|-\sum_{x\in X} max\,[0,A(x)-B(x)]$$

The $\sum$ term describes the degree to which the subset inequality $A(x)\leq B(x)$ is violated. The difference is the lack of such violations, and the cardinality in the denominator is a normalizing factor. It is easy to convert this to the more convenient form:

$$S(A,B)=\frac{|A\cap B|}{|A|}$$

where $\cap$ represents the standard fuzzy intersection. For example, given sets

$$X=\{.6,.3,.4,.8\}$$

$$Y=\{.4,.7,.2,.9\}$$

the degree of subsethood of X in Y is

$$S(X,Y)=\frac{|X\cap Y|}{|X|}=\frac{1.7}{2.1}=.809$$

It will be noticed that the degree of subsethood of A in B is a formula with |A| in the denominator - i.e., it seems that we are looking for A in B and we are calculating a percentage of A, not of B. It is indeed the case that A belongs in the denominator; fuzzy subsethood is unlike conventional percentages. Fuzzy subsethood should intuitively be thought of as the degree of the whole in the part, not the part in the whole, and therefore A is in the denominator.

CHAPTER V

Non-Binary Evaluation

Section 1

The Non-Binary Evaluation Concept

Traditionally, evaluation of information systems and determination of pertinence has been binary - i.e., a pertinent document is evaluated as 1 and a non-pertinent document is evaluated as a 0. Some authors (Voiskunskii, 1997, Miyamoto, 1990) have mentioned the possibility of evaluating pertinence in a fuzzy manner - that is by assigning a degree of relevance to a document, a percentage which can vary anywhere from 0 to 1. Some have suggested that in many cases the user would find it difficult to judge whether a document is pertinent or not, and would find it helpful to be given the latitude to evaluate the document on a fuzzy scale. To a fuzzy theorist, no such justification is necessary - many, if not all, things should be measured in a fuzzy manner, and information systems are no different. However, of all the researchers in the information retrieval field, very few have considered this possibility at all, and none have been able to produce complete and

satisfactory results. From among those who have mentioned this possibility, Miyamoto states explicitly that evaluation of effectiveness of fuzzy information retrieval systems is one of the areas that he specifically does not address. Voiskunkii does make an attempt to analyze evaluation of fuzzy information retrieval systems, but concludes that he is not sure that any of the measures he has used for evaluating binary retrieval would work. In fact, he shows that one of the main measures he has used for binary retrieval, namely the cos measure, leads to paradoxical results when applied to non-binary retrieval. It is this paradox, and Voiskunskii's inability to adapt his measures to non-binary retrieval, that this thesis will address itself to and offer a solution to the problem.

First let us consider exactly what we mean by evaluating pertinence in a fuzzy manner. Let us return to the simple, traditional evaluation measures of recall and precision, defined as follows:

$$R=\frac{\#\ retrieved\ and\ relevant}{total\ relevant}=\frac{r}{C} \qquad (3.1.1)$$

$$P=\frac{\#\ retrieved\ and\ relevant}{total\ retrieved}=\frac{r}{N} \qquad (3.1.2)$$

Which concept should be given a fuzzy interpretation? The concept of relevant, or the concept of retrieved, or both? Clearly the concept of relevant in total relevant (the denominator of R) should be given the ability to be evaluated in a fuzzy manner - i.e. when we consider in the

collection itself whether a document is relevant or not, we can assign to the document varying degrees of relevance from 0 to 1. But the question is, in the output, in total retrieved (the denominator of P), should the system also be limited to assigning a 1 to a document if it is retrieved, and a 0 if it is not, or should we allow the system the latitude to state that a particular document is somewhat less than perfectly relevant, say only .4 relevant. Voiskunskii (Voiskunskii, 1997) states explicitly that he considers the output to be only binary - we will take the point of view that the output vector can also be fuzzy. The best evidence that the output can also be fuzzy is the fact that in the real world today, on the Internet, for example, many search engines return, along with the documents retrieved, a percentage indicating how relevant the document is considered to be - and the documents are retrieved and ordered in terms of relevance, and listed from most to least relevant. (This indeed is the only practical thing for them to do, for if not, the user would be swamped with information overload of possibly tens of thousands of documents retrieved for each query.) Voiskunskii, on the other hand, when he discusses non-binary retrieval, defines v as the output vector and w as the true pertinence vector, and evaluates v in terms of 0's and 1's only, while the values in w can be non-binary. We, as we said, will assume that both of the vectors can be calculated in a fuzzy manner.

Section 2

Non-Binary Recall and Precision

Now let us consider the mechanics of calculating fuzzy or non-binary recall and precision. First let us review the mechanics of calculating standard or binary recall and precision. Let us assume that we have a collection of documents which we are querying and that we have independently established that five of these documents are relevant to the query. We will represent the collection as a vector, which we will call the pertinence vector, with each position representing a document; a 1 signifies that the document is relevant and a 0 signifies that the document is not relevant. Calling the pertinence vector w (following the notation of Voiskunskii, 1997; note that he changes the symbol for the non-binary pertinence vector to w, to clearly distinguish it from the binary pertinence vector, k, cited in Chapter 3, Section 3) we could have, for example:

$$w=(1,0,1,0,1,0,1,1)$$

meaning that the first document is relevant, the second not relevant, the third relevant, etc.

We queried the system and got results. Again, the results can be represented as a vector, with a 1 signifying that the system has decided

that the document is relevant, and a 0 signifying that the system has decided that the document is not relevant. Calling the system vector v, we could have for example

$$v=(0,1,1,0,1,1,0,0)$$

meaning that the system considers the first document to be not relevant, the second and third to be relevant, etc. Thus, in this case, the second, third, fifth, and sixth documents were retrieved by the system since they were considered relevant. Now, documents number 3 and 5 are both relevant and retrieved. The total number of relevant documents is five. Therefore, by 3.1.1,

$$R=\frac{2}{5}=.4.$$

As far as precision is concerned, again only two documents are both relevant and retrieved. The total number of retrieved documents is four. Therefore, by 3.1.2,

$$P=\frac{2}{4}=.5.$$

Let us now calculate recall and precision for the non-binary case. Consider the following example. Assume the pertinence vector, w, is

$$w=(.5,0,1,.8,.9,.3,.6)$$

and the system vector, v, is

$$v=(1,.3,.2,.4,.3,.5,0)$$

How should we do the calculating of the numerator and denominator of R? We will describe two possible methods.

Method #1: Looking at recall first,

$$R=\frac{\sum(w\cap v)}{\sum w}=\frac{\sum min(w,v)}{\sum w} \qquad (5.2.1)$$

This approach is simply to sum the percentages. For example, in our case above, the denominator, or the total relevant, will be .5+0+1+.8+.9+.3+.6=4.1 As for the numerator, retrieved and relevant, $w\cap v$, should be evaluated as every other fuzzy intersection, i.e.,

$$a\cap b=\min(a,b)$$

So for the first document, for example, where the pertinence is .5 and the system shows 1, we would evaluate relevant and retrieved as min(.5,1)=.5 For the second document, where the pertinence is 0 and the system shows .3, min(0,.3)=0 Evaluating $\sum w\cap v=\sum\min(w_i,v_i)$ we get:

$$.5+0+.2+.4+.3+.3+0=1.7$$

So for our case

$$R=\frac{1.7}{4.1}=.4146$$

For precision,

$$P=\frac{\sum(w\cap v)}{\sum v}=\frac{\sum min(w,v)}{\sum v} \qquad (5.2.2)$$

For the denominator we would sum the percentages of total retrieved. In this case

$$1+.3+.2+.4+.3+.5+0=2.7$$

The numerator, number retrieved and relevant, min(w,v), we just calculated as 1.7, so

$$P=\frac{1.7}{2.7}=.6296$$

Method #2: A case can be made for the following kind of calculation. For recall, first take each document and calculate what percentage of each document was recalled by the system. Then average over all documents. For example, in our case above, the third document was 1.0 pertinent, but the system showed it to be only .2 relevant, so the percentage recalled is .2 The fourth document was .8 pertinent, but the system showed it to be only .4 relevant, so the percentage recalled is $\frac{.4}{.8}=.5$ The first document (here it is a bit more complicated) was .5 pertinent, but the system showed a relevance of 1.0 What should we do? The answer is that as far as recall is concerned, the document was totally recalled; the fact that the system showed more relevance than the reality is the province of precision, and will show up in the precision calculation. But as far as recall is concerned, we calculate

$$\frac{\textit{relevant and retrieved}}{\textit{relevant}}=\frac{\textit{min}(w_i,\ v_i)}{w} \quad (5.2.3)$$

for each document separately and then sum in order to later calculate the average, which we discuss below. Thus:

$$\sum\left(\frac{\textit{relevant and retrieved}}{\textit{relevant}}\right)=\sum\left(\frac{\min(w_i, v_i)}{w}\right) \quad (5.2.4)$$

In our case

$$\frac{.5}{.5}+\frac{0}{0}+\frac{.2}{1}+\frac{.4}{.8}+\frac{.3}{.9}+\frac{.3}{.3}+\frac{0}{0}=1+0+.2+.5+.3+1+0=3.03$$

For precision we would make a similar calculation. First we take each document and calculate what percentage of each retrieved document is actually relevant. Then we average over all documents. For example, in our case above, the first document was retrieved at 1.0, but was really only .5 relevant, so the percentage precision, $\frac{\textit{relevant and retrieved}}{\textit{retrieved}}$, is $\frac{.5}{1}=.5$ The second document was retrieved at .3, but was actually not relevant at all, 0 relevant, so the percentage precision is $\frac{0}{.3}=0$. The third document was retrieved at .2, but the reality showed 1 relevance; in this case, as far as precision is concerned, the retrieval did not surpass the relevance, and was, in fact, less, so the percent of the retrieved document which is actually relevant is actually 1 (or 100 percent) - the fact that the relevance is more than the retrieval is the province of recall, not precision. Thus we would calculate

$$\frac{relevant\ and\ retrieved}{retrieved} = \frac{min(w_i, v_i)}{v} \quad (5.2.5)$$

for each document and then sum over all documents in order to calculate the average. Thus:

$$\sum\left(\frac{relevant\ and\ retrieved}{retrieved}\right) = \sum\left(\frac{min(w_i, v_i)}{v}\right) \quad (5.2.6)$$

In our case we would calculate:

$$\frac{.5}{.1}+\frac{0}{.3}+\frac{.2}{.2}+\frac{.4}{.4}+\frac{.3}{.3}+\frac{.3}{.5}+\frac{0}{0}==4.1$$

This brings us to the question of averaging for both recall and precision in this second method. One might be tempted to simply take an arithmetic average of the weights, i.e., the sum divided by the number of documents in the collection. Thus, in the case above we have seven documents, and since for recall we had a sum of percentages of 3.03, we would have an arithmetic average of $\frac{3.03}{7}$=.4328. This would, however, be incorrect. Calculation of non-binary recall must be consistent with calculation of binary recall. In binary recall, we do not in any way take into account the number of documents in the collection - just the number of relevant documents. So what about averaging by the sum of the relevance, i.e., $\sum w$. This would also be incorrect. Take the following example ;

w=(1,1,1,1)

v=(1,1,1,1)

Recall seems to be 100 percent, and by any calculation it will indeed be 100 percent.

Now take

$$w=(.5,.5,.5,.5)$$

$$v=(.5,.5,.5,.5)$$

It looks like the same 100 percent recall. However, if we calculate by method #2

$$\frac{.5}{.5}+\frac{.5}{.5}+\frac{.5}{.5}+\frac{.5}{.5}=4$$

and the average by $\sum w=2$ we get $\frac{4}{2}=2$ (which is 200 percent). How can recall be greater than 100 percent? Clearly what is wrong here is the averaging. Wherein lies the problem? It lies in the following. You see, even though in the first case above, the binary case, it looks like we might be averaging by $\sum w$ which is 4, in fact, we are not - we are averaging by the number of documents which are relevant. In the non-binary case, this becomes the number of documents which show any relevance at all. So in our case above, we would calculate $\frac{4}{4}=1$ (which is 100 percent) recall, just as we suspected. So we conclude that we should average not by the number of documents, nor by $\sum w$, but by the number of documents which show any relevance at all.

So the final formula for recall by method #2 is

$$R=\frac{\sum\left(\frac{\textit{relevant and retrieved}}{\textit{relevant}}\right)}{\textit{number of documents that are relevant at all}} \qquad (5.2.7)$$

$$=\frac{\sum\left(\frac{\min(w_i\ v_i)}{w}\right)}{\textit{number of documents that are relevant at all}} \qquad (5.2.8)$$

In our case this becomes

$$R=\frac{3.03}{4}=.7575$$

A similar argument can be presented for the case of precision. In the formula for precision we do not take into account the total number of documents here either - just the number of retrieved documents. We calculated the sum of the percentages for precision in the case above to be 4.1. Since we are not averaging by the total number of documents, we should not take the arithmetic average - $\frac{4.1}{7}$. Should we then average by the number of retrieved documents, in this case $\sum v$? The answer is no, just as for the case of recall above. Here it is clear from the example itself that we cannot do so. In this example, $\sum v = 2.7$. If we take $\frac{4.1}{2.7}$ we get a precision greater than 100%, which is impossible. What is wrong here again is the averaging process. Even though in the binary case looks it like we might be averaging by the $\sum v$, in fact we are not. We are averaging by the number of documents which were retrieved - which in the non-binary case becomes the number of

documents which are retrieved at all. So the final formula for precision by method 2 is

$$P=\frac{\sum\left(\frac{\text{relevant and retrieved}}{\text{retrieved}}\right)}{\text{number of documents that are retrieved at all}} \qquad (5.2.9)$$

$$=\frac{\sum\left(\frac{\min(w_i\ v_i)}{v}\right)}{\text{number of documents that are retrieved at all}} \qquad (5.2.10)$$

In our case the number of documents which are retrieved at all is 5, so our average for precision is $\frac{4.1}{5}$=.82.

So we have two methods for calculating non-binary recall and precision, namely, methods 1 and 2 described above. What is the difference between these two methods? The difference is whether, in non-binary evaluation of recall and precision, we give each document equal weight or not. If we give each document equal weight, that is, if each document counts as much as each other document in the total value of recall and precision, then we would use method 2. If, on the other hand, each document does not get equal weight, and what is important is the total relevance in the collection and the total retrieved in the retrieval, then we would use method 1.

Just to be a bit more clear on the issue, what does it mean to give each document equal weight? If we use as our example recall, it means the following (a similar explanation can be given for precision). If, for

example. a document is .2 relevant and we retrieved it a relevance of .1, by method 2 we would add into the average for recall a .5 - i.e., $\frac{.1}{.2}$ = .5. So by method 2, this document's contribution to recall would be 50 percent, and that is what we would average into the total recall. This 50 percent contribution would be equivalent to another document, which was 100 percent relevant, and which we succeed to retrieve at level .5 relevance. We consider ourselves 50 percent successful if we retrieved half the true relevance, no matter how much the relevance is and how much relevance we retrieved. Moreover, each document's performance contributes the same amount to the total. By method 1, however, in the first case we mentioned, the numerator of the recall formula, by formula 5.2.3, would only be increased by .1 and the denominator by .2, while in the second case, the numerator would be increased by .5 and the denominator by 1. All this would be added to the values of all the other document. The net result - i.e.: each of these document's contributions to the overall recall - is not clear. What is clear, however, is that method 1 considers only the total relevance and the total retrieved, and does not give each document equal weight.

Let us note that these two methods differ only insofar as non-binary retrieval and evaluation are concerned. For binary retrieval and evaluation these two methods reduce to the same calculation. Let us use the case we cited at the beginning of this section as an example

$$w=(1,0,1,0,1,0,1,1)$$

$$v=(0,1,1,0,1,1,0,0)$$

and also calculate only for recall (noting that a similar argument can be made for precision).

Using method 1, by 5.2.1, we would calculate the denominator quite simply as $\sum w$ because that is the total number of documents which are relevant. As for the numerator, retrieved and relevant means that the document was both in the relevant vector and in the retrieved vector - that there is a 1 in that position in both the w vector and the v vector. This can therefore be expressed in the same way we expressed the numerator for non-binary, i.e.,

$$\sum(w \cap v)=\sum\min(w_i,v_i)$$

because if there is a zero in either vector, the above expression will be zero, and it will only be 1 if there is a 1 in both vectors. In this case, $\sum\min(w_i,v_i)$ equals 2 - i.e., there is a one in both vectors in only two positions. So

$$R=\frac{2}{5}=.4$$

just as we calculated above - this calculation for method 1.

Turning to method 2, we would calculate formula 5.2.3 for each document, and then sum and average over all documents. Now, if a document is not relevant it will not be part of the calculation at all. So we limit ourselves to the documents that are relevant - i.e., the

documents that have a 1 in the w vector. If that particular document has a 1 in the v vector as well, then

$$\frac{r}{C}=\frac{1}{1} \qquad (5.2.3)$$

If not, then

$$\frac{0}{1}=0$$

We do this calculation for each document. In our case, this is only done for documents 1,3,5,7, and 8, as documents 2,4, and 6 have zeros in the w vector and hence are not relevant at all and not part of the discussion. Doing the calculation, we get

$$\frac{0}{1}+\frac{1}{1}+\frac{1}{1}+\frac{0}{1}+\frac{0}{1}=2$$

We then average over all relevant documents - just as in non-binary, over all documents that show any relevance at all - yielding, in our case

$$\frac{2}{5}=.4$$

which is the same value as we calculated for method 1.

What is going on here is this. In binary evaluation the averaging will be by the same value - i.e., Σw. This is because in binary the document is either fully relevant or not relevant at all, so it can only have a value of 0 or 1. Therefore the number retrieved and relevant will be averaged by the total relevant, which is Σw or the total number of relevant documents. So as far as the denominator is concerned, the

value will be the same in both method 1 and 2. As far as the numerator is concerned, method 1 uses the total number of documents where a 1 appears in both the w vector and the v vector; and for method 2 the numerator value reduces to the same thing. This is because in method 2 in the numerator we calculate for each document

$$\frac{r}{C}$$

Now since we are limiting ourselves only to relevant documents, the relevant in the denominator will always be equal to 1. So the fraction will be 1 only when the numerator is also 1, which will only be when there is a 1 in both the w vector and the v vector, because that is the only time the document is both relevant and retrieved,. So for method 2, the numerator will also be the sum of documents where there is a 1 in both the w vector and the v vector. For both methods then, both the numerator and denominator are equivalent, therefore they reduce to the same value.

Section 3

Non-Binary Composite Measurement

Voiskunskii's Attempt and Failure

Let us now turn to Voisunskii's attempt to use his composite measure in the non-binary situation . In Chapter 3, Section 3 we described Voiskunskii's composite binary measure. Basically, as a result of a search of a collection, two vectors are constructed, a pertinence vector, k, representing the collection,

and the true pertinence of the documents in it, and a retrieval vector, v, representing the documents retrieved by the system and the system's evaluation of their pertinence. Now if the two vectors coincide, then the search has been completely successful; if not , then the measure of how successful the search has been can be represented by the degree of correspondence between the two vectors as measured by the cos of the angle between the two them, as in the following formula.

$$\cos\varphi_{KV}=\frac{\sum_{i=1}^{N_0} k_i\, v_i}{\sqrt{\sum_{i=1}^{N_0} (k_i)^2}\sqrt{\sum_{i=1}^{N_0} (v_i)^2}} \qquad (3.3.1)$$

Voiskunskii then considers the case of evaluating the relevance of the documents in a non-binary manner. In order to avoid confusion between the binary and non-binary cases, he calls the non-binary pertinence vector w instead of k (as we did in Section 2 above). He finds at least one case in which two searches and results, one which is clearly better than the other, results in the cos measure showing that the better is worse and the worse is better. (Please note that henceforth we will be numbering cases in order to facilitate referring to them throughout the rest of the paper. In keeping with this, we will refer to this case as Case #1.) He gives a case of a collection of N documents being queried. In the w vector, a value represents that the document is pertinent and the value itself represents the degree of pertinence, and a 0 represents non-pertinence - that in the non-binary aspect of the evaluation. Two searches are performed using two different search requests, and the results are represented by the vector where a 1 represents that the document was retrieved, and a 0 represents that it was not. (Note that Voiskunskii does not consider the possibility of the retrieval vector being evaluated in a non-binary manner as well; we , in our discussion assume that it can.) Here are the results:

Case #1: $v_1 = v_2 = (1,1,1,1,0,0,0,...)$

$w_1 = (0.1,0.1,0.1,0.1,0,0,...)$

$w_2 = (0.1,1,1,1,0,0,.....)$

Using his cos measure, Voiskunskii calculates a CSC for each of the searches with the following results:

$$\cos w_1 v_1 = 1$$

$$\cos w_2 v_2 = .89$$

So, Voiskunskii says, it seems from the calculated cos measure that the first search is better than the second, when clearly the second is better than the first, because in the second more documents were successfully completely retrieved than in the first. What Voiskunskii is pointing out is that in this case at least, when an attempt is made to use the cos measure in the non-binary case, the results are paradoxical, even though the binary measure is quite successfully used the binary case. In the case that is cited, clearly, w_2v_2 in which three complete matches obtain (i.e. three ones in the output matching three ones in our evaluation of pertinence), is superior to w_1v_1 where there are no complete matches: yet the cos measure shows the opposite.

This thesis takes the position that the reason that Voiskunskii fails in his attempt to apply the cos measure to the non-binary case is that he fails to use the principles of fuzzy sets discovered and established by fuzzy theorists over the past few years. Indeed, it is clear that the non-binary case is the fuzzy case. In order to analyze it properly we will suggest the use of fuzzy techniques as the answer to this and similar paradoxes. We will use fuzzy principles to resolve

Voisunskii's case, find and resolve other paradoxical cases, and study in depth the entire subject of CSC's for fuzzy information retrieval.

Chapter VI

Non-Binary Evaluation Using Fuzzy Set Techniques

Section I

A Fuzzy Composite Measure Extending Voiskunskii's Cos Measure

In this section we are going to show how fuzzy techniques can be used to resolve Voiskunskii's paradox in the case that he cited. First. however. let us consider how the cos measure can be used in the non-binary case at all. As we showed above, Voiskunskii shows that in the binary case $\cos\varphi_{kv}=\sqrt{R\cdot P}$ (formula 3.3.3) (where R is recall and P is precision). He does this, it should be recalled, by saying $\sum(k_i)^2=C$ (the number of pertinent documents in the collection), $\sum(v_i)^2=N$ (the number of documents in the output), and $\sum(k.v)=r$ (the number of positions having a one in both vector k and vector v). As a result:

$$\cos\varphi_{kv}=\frac{\sum k.v}{\sqrt{\sum(k_i)^2}\sqrt{\sum(v_i)^2}}=\frac{r}{\sqrt{C}\sqrt{N}}=\sqrt{\frac{r^2}{C.N}}=\sqrt{\frac{r}{C}\frac{r}{N}}=\sqrt{R\cdot P}$$

Now it should be noted that clearly the above equalities, i.e.: $\sum(k_i)^2=C$, $\sum(v_i)^2=N$, and $\sum$(k.v)=r, do not apply to the non-binary case. The only reason they are true in the binary case is because if a document is pertinent, or outputted, it gets a 1 - and then 1^2 is also equal to 1, so $\sum(k_i)^2$ and $\sum(v_i)^2$ are simply the sum of all pertinent or outputted documents, respectively. In the non-binary case, however, w_i is a decimal number (as opposed to k which is the binary version of w), and squaring each decimal and adding them up will not be equal to C. which is the sum of the vector, as we defined it above. In other words, the summation of the w_i^2 will not be equal to the summation of the w_i's ($\sum w_i$), so the $\sum w_i^2$ will not be equal to C if C is the sum of all the pertinent decimal percentages, as we defined it above. Similarly for v, and certainly for k.v, where if k and v (or w) are both percentages. k.v will not be the number of positions having a one in both vector k and vector v, and hence $\sum$ k.v will not be equal to r. So for the non-binary case $\cos_{wv}$ cannot be shown to be equal to $\sqrt{R.P}$. This being the case , how do we justify using the cos measure in the non-binary case? We can take a number of approaches. First of all, the cos measure can stand on its own feet - it does in fact measure the similarity of two vectors, so why not use it in the non-binary case. Second, in the binary case it resolves to a well-known measure , so it

makes sense to say that it would be a good measure in the non-binary case as well. Third, instead of using the cos measure itself, which in the non-binary case does not resolve to $\sqrt{R.P}$, start with $\sqrt{R.P}$ and work backward to $\frac{r}{\sqrt{C}\sqrt{N}}$ and use this formula in lieu of the cos measure. It is just this direction that our thesis will take. We will use $\frac{r}{\sqrt{C}\sqrt{N}}$ as our cos measure - because, as noted the "pure" cos measure does not resolve to $\sqrt{R.P}$.

If we use as $\frac{r}{\sqrt{C}\sqrt{N}}$ our cos measure and apply it to Voiskunskii's paradoxical case, we get a result which conforms with our intuition concerning which of the two cases is better. However, for the concept "documents both pertinent and retrieved", which is the intersection of the set of pertinent documents and the set of retrieved documents, we will use the fuzzy intersection. Now, in our notation from Chapter V, Section 2 above, w is the pertinence vector, v is the retrieval vector, and r refers to the total number of documents both pertinent and retrieved. The value r is therefore the intersection of C and N ,or w and v. In fuzzy sets, recall that intersection is normally defined as the min operator. In this case $w \cap v = \min(w,v)$. Applying the formula

F_1: $$\frac{r}{\sqrt{C}\sqrt{N}} = \frac{min(w,v)}{\sqrt{\sum w}\sqrt{\sum v}} \qquad (6.1.1)$$

we get:

$$F_1(w_1v_1) = .31622$$

$$F_1(w_2v_2) = .8803408$$

So w_2v_2 , which looks better intuitively, actually calculates to a higher value. (Please note that we will be labeling our measures F_i in order to facilitate referring to them in the last section of this chapter.)

Let us now point out why Voiskunskii's measure did not work for the particular case which he cited. The reason that Voiskunskii's case results in w_1v_1 being equal to one is that the pure cos measure will reduce to 1 whenever the v_i's are equal to each other and the w_i's are equal to each other, and this value of 1 does not reflect the "true" correspondence between the two vectors in this case. The proof is as follows. If the w_i's are equal to each other, and the v_i's are equal to each other then

$$\frac{\sum w \cdot v}{\sqrt{\sum (w_i)^2}\sqrt{\sum (v_i)^2}} =$$

$$\frac{n \cdot w \cdot v}{\sqrt{nw^2}\sqrt{nv^2}} =$$

$$\frac{n \cdot w \cdot v}{\sqrt{n^2 w^2 v^2}} = 1$$

For example, in Voiskunskii's case, all the v_1's are 1 and all the w_1's are .1. Applying the formula above, we get

$$\frac{n \cdot w \cdot v}{\sqrt{n^2 w^2 v^2}} = \frac{4 \cdot .1 \cdot 1}{\sqrt{4^2 \cdot .1^2 1^2}} = 1$$

The reason for this that cos is a measure of the angle between two vectors. If the v_i's are equal to each other and the w_i's are equal to each other , then the vectors are parallel and the cos value of the angle between them will equal 1. The value will be 1 regardless of the relative magnitudes of the vectors, or whether they are equal to each other. The cos measure will therefore show perfect correspondence even when no perfect correspondence exists. For the binary case this is not a problem. If the v's are equal to each other (and non-zero). they can only be 1's. And if the w's are equal to each other (and non-zero) and are therefore also 1's, then perfect correspondence does indeed exist. But in the non-binary case, the two vectors can be very different than each other, but as long as the v's are equal and the w's are equal, the measure will show perfect correspondence, but none obtains. This is precisely the situation in Voiskunskii's case cited above. The vector w_1v_1 calculated to 1 because the v's and w's were equal to each other - and not because of any underlying correspondence between w and v. So for w_2v_2, where there is non-perfect correspondence and non-parallelism, we get .89, but for w_1v_1, where there is also non-perfect correspondence, but parallelism, we get 1.

Section 2

Characteristics and Limitations of the Fuzzy Composite Cos Measure

Let us extend Voiskunskii's objections to using the cos measure in the non-binary case by showing some additional paradoxical cases. In light of what we said in the last section, let us turn to another case:

Case #2: $v_1=(1,1,1,1)$

$w_1=(0.1,0.1,1,1)$

$v_2=(1,1,0,0)$

$w_2=(0.1,0.1,0,0)$

We submit that the above two results are equivalent - for the first two documents, v (the system) shows that they are perfectly relevant, and w (the reality) is that they are only 0.1 relevant in both cases, so insofar as the first two documents are concerned, the results match. For the next two documents, v_1 shows 1,1 - i.e. that they are perfectly relevant - while w_1 also shows perfect relevance, and v_2 shows 0 - no relevance - with w_2 agreeing at 0 - no relevance. In other words, for the second two documents the system matches with the reality in both cases. So with matching results for two documents and equivalent results for the

next two documents, the results should be considered to be equivalent. Yet, when we apply the cos measure to the two case we get:

$$\cos\varphi(w_1v_1) = .774$$

$$\cos\varphi(w_2v_2) = 1.000$$

Now the $w_2v_2 = 1.000$ is the result of the w's being equal to each other and the v's being equal to each other, as noted above. Nevertheless, here we have another case in which the cos measure gives us results other than our intuition. If we apply our measure, 6.1.1, we get:

$$F_1(w_1v_1) = .7432$$

$$F_1(w_2v_2) = .316$$

which still does not match, but our measure shows w_1v_1 to be better than w_2v_2, whereas Voiskunskii's measure shows w_2v_2 to be better than w_1v_1. We can therefore resolve the problem with our measure in the following way. The cos measure, whether in its pure form, or in our form, is biased towards recall - even though $\sqrt{R.P}$ doesn't look like it. The bias becomes apparent when one looks at the formula $\frac{r}{\sqrt{C}\sqrt{N}}$. If more documents are retrieved (successfully), then the numerator is larger (making the number larger); and even though the denominator is also larger in N (making the fraction smaller), it is only larger to the extent of taking the square root, which is less of an effect than that of the numerator. The net result is that as a result of more retrievals the numerator increases the fraction more than the denominator decreases

it. So the measure is biased towards recall - which therefore causes w_1v_1, where more documents are retrieved (successfully) to be greater than w_2v_2. However, Voiskunskii's measure shows w_1v_1 to be worse, and not better, and so Voiskunskii's measure cannot be resolved as above.

The following is a similar case in which both our measure and Voiskunskii's measure do not show equivalence when we think that they should, but both show a bias towards recall as we argued above.

Case #3:

$v_1=(1,1,1,1)$

$w_1=(1..5,1,.5)$

$v_2=(1,0,1,0)$

$w_2=(1,.5,1,.5)$

We would claim that these two results should be considered to be equivalent - for the same reason as the last case cited. In both results here, two documents are perfectly relevant (w=1) and they are retrieved as such (v=1). As far as the other two documents are concerned, in one result, $v_1=1$, i.e. they are retrieved as being perfectly relevant, but $w_1=.5$, in reality they are only partially (.5) relevant. For the other result, $v_2=0$, i.e. they are not retrieved, but $w_2=.5$, in reality they are partially (.5) relevant. These two results should be equivalent, as there are matching results for two documents, and for the other two documents, in one case the system overestimates their true relevance by

a half and in the other case it overestimates their true relevance by a half, so they should be equivalent. But they are not. Voiskunskii's measure gives

$$\cos\varphi(w_1v_1) = .9487$$

$$\cos\varphi(w_2v_2) = .894$$

and our measure gives

$$F_1(v_1w_1) = .866$$

$$F_1(v_2w_2) = .8165$$

So indeed any cos measure, Voiskunskii's or ours, is biased towards recall. What is the justification for this bias? We can argue one of two things. One, a system which does not find a somewhat relevant document (w_2v_2) cannot be considered equivalent to another one which finds documents , but just underestimates their relevance (w_1v_1): the second is clearly superior. Two, both measures are biased towards recall, and therefore show w_1v_1 to be better. In reality, argument number one is actually saying the same thing and more: that the measure is biased towards recall, and that indeed there is an intuitive and philosophical reason why a bias towards recall is reasonable.

Indeed, let us consider an argument to the effect that an aggregate CSC should be biased towards recall. Intuitively, it makes sense to consider the finding of a relevant document that is there to be more important, and count more, and be weighted more, than the non-

finding of a non-relevant document. After all, the main purpose of an information system is to find relevant documents, and when it does, it should count heavily in the evaluation of that system. For the user, too, finding something relevant is more important than rejecting something non-relevant. However, this second argument is not at all clear. Some users may consider rejecting and "weeding out" non-relevant documents to be more important than finding relevant ones. An example of such a user is one who is subjected to information overload, as in the internet, for example. If many such users exist, then we may not be able to argue that for this user a measure biased towards recall is better.

In any case, to summarize what we have said so far, we started with Voiskunskii's example, and used our own version of the cos measure to resolve his paradox. We also cited two other examples which intuitively appear to be equivalent, but both cos measures, ours and Voiskunskii's, give non-equivalent results. We resolved our own measure by saying that any version of the cos measure is biased towards recall, but Voiskunskii's measure shows the instance of greater recall to be worse in one of the cases. That case, the second one cited above, is very similar to Voiskunskii's own example. They both have one result equal to one - and we showed that this is the result of the two vectors being parallel and that this would always be the case whenever the v's are equal to each other and the w's are equal to each other. To be sure,

this is the main result of the discussion so far - that is, Voiskunskii's measure cannot work on fuzzy results because it will show perfect correspondence whenever the v's are equal and the w's are equal even though no perfect correspondence exists.

Let us further investigate the properties of Voiskunskii's measure and our measure by the use of some more examples. First let us clarify exactly what we mean by equivalence. Consider the following case.

Case #4:

w_1=(.9,.8,.9,.8)

v_1=(.8,.9,.8,.9)

w_2=(.5,.4,.5,.4)

v_2=(.4,.5,.4,.5)

Is w_1v_1 equivalent to w_2v_2 or is one better than the other? If one is better, which one? Consider that the two pairs show equivalent total distance - i.e. $\sum |w_i-v_i| = .4$ in both cases. If one uses a pure distance criteria then the two cases should be equivalent. (Indeed, perhaps this absolute distance itself would be a good composite measure; i.e., the smaller the absolute distance between the two vectors, the better the results. We will show here that in the non-binary case, at least, it is not.) However, we would argue that in spite of the equivalent absolute distance w_1v_1 should be considered to be better. This is because in w_1v_1 for each document the system error from the reality is a lower percentage than that of w_2v_2: i.e.., $\frac{.8}{.9}$ is a higher percentage than $\frac{.4}{.5}$

(.8). This example clarifies to us what we really mean by equivalence. It shows that two equidistant cases should not necessarily be considered equivalent - the true measure of equivalence is percentage retrieved. (Another reason for not using the distance measure pure and simple is the fact that, as Voiskunskii points out, it does not have the order preservation property as his measure does.) The fact that w_1v_1 is indeed better than w_2v_2 is borne out by Voiskunskii's measure:

$$\cos\varphi(w_1v_1) = .993103$$

$$\cos\varphi(w_2v_2) = .97561$$

and by our measure:

$$F_1(w_1v_1) = .941176$$

$$F_1(w_2v_2) = .888889$$

Here is another case:

Case #5: $w_1 = (.6,.2,.6,.2)$

$v_1 = (.2,.6,.2,.6)$

$w_2 = (.8,.4,.8,.4)$

$v_2 = (.4,.8,.4,.8)$

again the absolute distances, $\sum |w_i - v_i|$, are equivalent at 1.6, but the percentage of reality retrieved, $\sum \frac{w}{v} \cup \frac{v}{w}$, are different; for w_1v_1, .33*4, and for w_2v_2, .5*4. Therefore, w_2v_2 should be better, as indeed, for Voiskunskii's measure:

$$\cos\varphi(w_1v_1) = .6$$

$$\cos\varphi(w_2v_2) = .8$$

and by our measure:

$$F_1(w_1v_1) = .5$$

$$F_1(w_2v_2) = .667$$

Consider the following case, however,

Case #6: $w_1=(.9,.8,.9,.8)$

$v_1=(.8,.9,.8,.9)$

$w_2=(.9,.8,.9,.8)$

$v_2=(.8,.7,.8,.7)$

Again, the distance measure ($\sum |w_i-v_i|$) shows equivalence. However, the percentage measure shows w_1v_1 to be better as $\frac{.8}{.9} \geq \frac{.7}{.8}$ ($.888 \geq .875$). Applying our measure we get,

$$F_1(w_1v_1) = .941176$$

$$F_1(w_2v_2) = .939336$$

However, applying Voiskunskii's measure we get,

$$\cos\varphi(w_1v_1) = .993103$$

$$\cos\varphi(w_2v_2) = .999969$$

Why? (Can it be argued that the recall bias outweighs the percentage effect? Only if w and v are as above - what if w and v are reversed with

$w_2=.8,.7,.8,.7$

v_2=.9,.8,.9,.8

then the recall is actually better in w_2v_2.) So we have another problematic case for Voiskunskii's measure.

The final case (of this series,) is the following,

Case #7:

w_1= 1,.9,.8,.7

v_1= .2,.3,.4,.5

w_2= .7,.8,.9,1

v_2= .2,.3,.4,.5

Again, the pure distance measure ($\sum |w_i-v_i|$) shows equivalence (2.0 for both sets). As far as the percentage measure is concerned, taking individual percentages and summing them up, i.e., $\frac{v}{w} = \frac{.2}{1} = \frac{.3}{.9}$ .... and summing them up, $\sum \frac{v}{w}$, for w_1 and v_1 we get 1.747619, and for w_2 and v_2 we get 1.605159. Now we want to get the average percentage. There are two ways of doing this. One, we can give each document equal weight and calculate $\frac{\sum \frac{v}{w}}{4}$, or, two, we can take $\frac{\sum \frac{v}{w}}{\sum w}$. In this case $\sum w$=3.4 in both cases; if we reverse the v's and the w's $\sum w$=1.4 in both cases. In any case, since we will be dividing by the same number in all cases, whenever the total percentage ($\sum \frac{v}{w}$) is greater,

the average percentage will also be greater. For $\frac{\sum \frac{v}{w}}{4}$, for w_1 and v_1 we get .436905, and for w_2 and v_2 we get .40129. So the percentage measure shows w_1v_1 to be better. When applying Voiskunskii's measure we get

$$\cos\varphi(w_1v_1) = .904762$$

$$\cos\varphi(w_2v_2) = .984127$$

which indicates w_2v_2 to be better. Applying our measure we get,

$$F_1(w_1v_1) = .641689$$

$$F_1(w_2v_2) = .641689$$

indicating that they are equal. Another case problematic to Voiskunskii, but also questionable in terms of our measure. (We cannot appeal to recall bias because (a) neither has recall bias, and if one does, it is the one with the greater percentage recall, and (b) if v and w are reversed, the same result would obtain.) The justification of our measure lies in the fact that it does not give each document equal weight - to the degree that it is the sum of the fuzzy values that is important, not each one individually. The conceptual explanation agrees with the actual arithmetic of the calculation. In any case, we have a third problematic case for Voiskunskii's measure where our measure yields the correct intuitive result.

For completeness sake let us look at another case which seems to present some problems for us as well as for Voiskunskii.

Case #8: $w_1=(.5,.5,.5,.5)$

$v_1=(.4,.4,.4,.4)$

$w_2=(.5,.4,.5,.4)$

$v_2=(.4,.5,.4,.5)$

The distance measure ($\sum |w_i-v_i|$) clearly shows w_1v_1 to be equivalent to w_2v_2. The percentage measure also shows equivalence; it calculates to 4.0 in both cases. However, upon calculating Voiskunskii's measure we get:

$$\cos\varphi(w_1v_1) = 1.000$$

$$\cos\varphi(w_2v_2) = .97651$$

and upon calculating our cos measure we get:

$$F_1(w_1v_1) = .894427$$

$$F_1(w_2v_2) = .888889$$

Now clearly the value 1.000 that we get for w_1v_1 is the result of the vectors being parallel, as explained above. Nevertheless, both cos measures, even ours, evaluate w_1v_1 to be better than w_2v_2. It seems that the case in which w is a constant .5 and v is a constant .4 is better than the case where w is alternately .5 and alternately .4, and so is v. Let us prove that this is general and will result in all such cases - that is, in all cases where $w_i > v_i$ (or $v_i > w_i$) as opposed to the case where w_i and v_i

alternate in their magnitudes we would get the constant case to be better than the alternating one. Let a > b and

$$w_1=(\ a,a,a,a,\ldots)$$

$$v_1=(\ b,b,b,b,\ldots)$$

$$w_2=(a,b,a,b,\ldots)$$

$$v_2=(\ b,a,b,a,\ldots)$$

Using our cos measure $\frac{r}{\sqrt{C}\sqrt{N}}$:

$$F_1(w_1v_1) = \frac{bn}{\sqrt{an}\sqrt{bn}} = \frac{bn}{n\sqrt{ab}} = \frac{b}{\sqrt{ab}} = \sqrt{\frac{b^2}{ab}} = \sqrt{\frac{b}{a}}$$

$$F_1(w_2v_2) = \frac{bn}{\frac{n}{2}(a+b)} = \frac{2b}{a+b}$$

Now

$$\sqrt{\frac{b}{a}} > \frac{2b}{a+b}$$

$$\frac{b}{a} > \frac{4b^2}{(a+b)^2}$$

$$\frac{1}{a} > \frac{4b}{(a+b)^2}$$

$$(a+b)^2 > 4ab$$

$$a^2 - 2ab + b^2 > 0$$

$$(a-b)^2 > 0$$

Hence, w_1v_1 will always be greater than w_2v_2 whenever the two vectors are defined as above, with w_1v_1 having w consistently greater than v.

and w_2v_2 having the w's and v's alternate in magnitude. How can we justify this result? Why should w_1v_1 always exceed w_2v_2 ? We can suggest the following explanation. Consistency of results is given greater value than non-consistency - w_1v_1 is consistent in its error while w_2v_2 is not. Consistency is valuable in and of itself, and more so because if the results are consistent, even if they are in error, as long as they are consistent in their error, the error can be adjusted for. So it is fitting that consistent results be evaluated higher than non-consistent ones.

We have seen so far that by modifying Voiskunskii's cos measure and using fuzzy techniques, we have been able to develop a measure that works in the non-binary case, where the "pure" cos measure itself does not work. We have shown a number of cases where intuitively we would expect one result, but when we calculate Voiskunskii's measure we do not get the intuitively expected result. However, when we calculate our own modification, we do indeed get the intuitively expected result. We have based our intuition on a percentage measure - i.e.. the percentage of reality yielded by the system or the percentage of the system represented by the reality -

$$\sum \frac{min(w,v)}{max(w,v)}. \qquad (6.3.2)$$

Now the question arises, if our intuitive measure can tell us what would be considered equal and what would be considered better, why

not use the intuitive measure itself as our measure, where the higher the percentage, the better the system. The answer is as follows.

Any measure we use must work for both the totally binary case and the total non-binary case, and for partially binary and partially non-binary cases as well. Now there will be cases, and indeed many cases. where the above percentage measure will give erroneous results. For example

$$w_1 = (.2,0...)$$

$$v_1 = (\ 0,.2...)$$

$$w_2 = (\ .6,0...)$$

$$v_2 = (\ 0,.6...)$$

The percentage measure (6.3..2) will evaluate them both at 0, when clearly w_1v_1 is better than w_2v_2, as the error is less in w_1v_1 than in w_2v_2. In this case, the distance measure ($\sum |w_i-v_i|$) would have been better. but we saw above that the distance measure does not give the correct intuitive result in other cases. Indeed, whenever w or v contains zeros the percentage measure will always evaluate the percentage comparison as zero, regardless of the absolute distance. And the greater the absolute distance, the worse it is, but the percentage measure does not reflect this fact. This kind of case is by no means uncommon, but on the contrary, quite common, and in fact takes place quite frequently. Search results and evaluations, w and v, will always contain some zero

values, and for these cases the percentage measure fails. We can even say that in the vast majority of cases it will fail. In summary, the percentage measure will not work, in spite of the fact that it helps us see which results are intuitively better than others, or equivalent, as above.

So far we have shown how fuzzy techniques can be used to modify the "pure" cos measure and make it work. However, in a number of cases above, we posed some questions on the new measure and have had to justify it in a number of ways. In case #2 above, we had to argue that any cos measure is biased towards recall. In case #7 above, we had to explain that the measure does not give equal weight to each document. In case #8 above, we had to explain that consistency is evaluated higher than non-consistency. In sum, we have had a lot of explaining to do to justify our results. Perhaps there is some other measure that works and is less problematic.

Section 3

A Fuzzy Measure Using the Concept of Subsethood

In Chapter V we explained the concept of subsethood and defined it mathematically as

$$S(A,B)=\frac{|A\cap B|}{|A|}$$

In light of this definition, let us consider what we are looking for in our case. To measure the effectiveness of a system we should be looking for the degree of subsethood of the output vector (v), in the pertinence vector(w), and the degree of subsethood of the pertinence vector in the output vector. (In a sense, the output vector in the pertinence vector is recall and the pertinence vector in the output vector is precision and we are looking for both.) We should then define the degree of subsethood of v in w which is

$$\frac{|v\cap w|}{|v|}$$

and the degree of subsethood of w in v which is

$$\frac{|w\cap v|}{|w|}$$

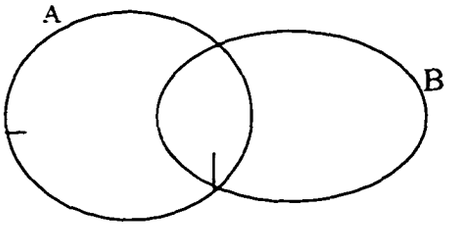
as

$$\frac{|v\cap w|}{|v\cup w|}$$

where $\cap$ denotes the standard fuzzy intersection and $\cup$ denotes the standard fuzzy union. The formula then becomes

$$F_2: \qquad \frac{\sum min(v,w)}{\sum max(v,w)} \qquad (6.4.1)$$

In fact, Van Rijsbergen develops just such a measure for binary evaluation. His development is along the following lines. We denote A as the set of documents found in the search, and B as the set of pertinent documents in the collection. The relationship between the sets is represented in the figure below.



The two sets partially overlap and the intersection of the two sets represents the found documents which are indeed pertinent. (The non-intersecting part of A represents the documents found and not pertinent and the non-intersecting part of B represents the documents pertinent but not found.) Clearly, the higher the degree of coincidence of sets A and B, the better the search results. If the two sets coincide completely we have perfect results, as all pertinent documents were found in the search. If we define $|A \cap B|$ as the number of elements in the intersection of A and B , and $|A \cup B|$ as the number of elements in the

union of A and B (where the || symbol is now Van Rijsbergen's symbol, not the fuzzy scalar cardinality above. even though, in fact, we will see both these concepts collapse to the same concept), then a natural measure to determine the degree of coincidence of the intersection to the union is

$$\frac{|A \cap B|}{|A \cup B|} \qquad (6.4.2)$$

or the percent of A∪B that A∩B represents. This ends the description of Van Rijsbergen's exposition.

Now Van Rijsbergen's |A∩B| is equivalent to the sum of documents in the intersection of A and B. which in the non-binary or fuzzy context is exactly equivalent to the scalar cardinality of the intersection of A and B (coincidentally. the symbol for the scalar cardinality is exactly the symbol Van Rijsbergen uses, namely ||). And so too for Van Rijsbergen's |A∪B|, which is the sum of the documents in the union of A and B, in the fuzzy context exactly equivalent to the scalar cardinality of the union of A and B. So Van Rijsbergen's binary measure, 6.4.2, turns out to be exactly the same measure we derived using the fuzzy technique of subsethood.

Now we showed above that the percentage measure itself would not be a good evaluation technique because of the "zero problem" which we described in detail. In truth, this problem can easily be addressed by replacing the "pure" percentage measure with the following function:

$$\min(v,w)\{\neq 0 : \sum (1-\frac{max(v,w)-min(v,w)}{max(v,w)})$$

$$= 0 : \sum (1-\frac{max(v,w)-min(v,w)}{1})\}$$

(Clearly this function reduces to the pure percentage measure except for the zero case - the reason we introduce it and put it in this form will become clear below.) This measure is a combination of the distance measure and the percentage measure in such a way that it reflects the distance of the system from the reality and the percentage of error of that distance as well. Subtracting from 1 has the effect of showing the degree to which there is no error, rather than measuring the percentage error, resulting in a measure which has a higher value the better the results are. When min(v,w) = 0, however, the upper measure does not work, and so lower measure above has to be substituted to obtain the same result. For example

$w_1=(.3,.6,.2)$

$v_1=(.2,.8,0)$

$w_2=(.3,.6,.6)$

$v_2=(.2,.8,0)$

Clearly, w_1v_1 is better than w_2v_2. Calculating for w_1v_1 we get 2.41, and for w_2v_2 we get 1.81.

Now looking at the formula for the case where min(v,w) is not 0:

$$\sum \left(1 - \frac{max(v,w) - min(v,w)}{max(v,w)}\right) = \sum \left(\frac{max(v,w)}{max(v,w)} - \frac{max(v,w) - min(v,w)}{max(v,w)}\right) =$$

$$\sum \frac{min(v,w)}{max(v,w)} \qquad (6.4.3)$$

So it reduces to $\sum \frac{min(v,w)}{max(v,w)}$. Why didn't we write this in the first place? The reason is because of the "zero" case. The only way we could devise an expression which would work correctly for the zero case is by including the 1 in the expression and subtracting. We then devised a congruent formula for the non-zero case. The formula for the zero case is a special case and special formula to correct the main formula for the zero case. So the formulas intuitively should look congruent. In any case, the main formula is the one for the non-zero case, 6.4.3, as we said. Now what is the relationship between this and the formula we derived above using the concept of subsethood, namely, 6.4.1? The two formulas are actually trying to measure the same thing. the difference is simply whether we treat each document equally and give each one equal weight or not. If we do, then we use 6.4.3. If we do not, and consider the sum to be the important thing, then we use 6.4.1. We made a similar distinction above, when we discussed the mechanics of calculating non-binary recall and precision.

Section 4

Other Fuzzy Measures Using the Concept of Subsethood

Let us consider another possibility. As we said ,we are looking for a measure which will find the degree of subsethood of v in w and of w in v . We might then define the degree of subsethood of v in w which is

$$\frac{|v \cap w|}{|v|}$$

and the degree of subsethood of w in v which is

$$\frac{|w \cap v|}{|w|}$$

as

$$\frac{|v \cap w|}{|v|} + \frac{|w \cap v|}{|w|}$$

This then becomes

$$F_3: \qquad \frac{\sum min(w,v)}{\sum v} + \frac{\sum min(w,v)}{\sum w} \qquad (6.5.1)$$

For example, in the case mentioned above, doing the appropriate calculations would result in

$$F_3(w_1 v_1) = \frac{.2+.6+0}{.2+.8+0} + \frac{.2+.6+0}{.3+.6+.2} = \frac{.8}{1} + \frac{.8}{1.1} = 1.527$$

$$F_3(w_1v_1) = \frac{.2+.6+0}{.2+.8+0} + \frac{.2+.6+0}{.3+.6+.6} = \frac{.8}{1} + \frac{.8}{1.5} = 1.33$$

Clearly, w_1v_1 is better than w_2v_2, as it should be.

Having arrived at this point - namely, the point where we are using the concept of subsethood in an additive fashion, as above, another possibility comes to mind. First, however, let us consider the recall-precision problem from another standpoint. Let us remember that the percentage of relevant documents retrieved is indeed a measure of effectiveness - indeed, it looks like complete measure of the effectiveness of the system. So why do we need another measure at all? The reason is that even though recall measures how effective the system is, how many "hits" the system has made, it does not measure how many errors the system has made. The number of errors should , in some sense, detract from the number of hits. The precision measure is meant to be a measure of errors.

However, there is a better measure of errors called "fallout". Salton (Salton, 1983) defines fallout as the number of documents retrieved and not relevant divided by the total number of non-relevant documents in the collection - i.e., the percentage of non-relevant documents retrieved by the system. Salton goes on to say that a recall-fallout pair as a measure is just as good as a recall-precision measure. The question then arises, why has precision become so widely used and accepted as the standard and not fallout? We would venture to say that

the reason is that fallout suffers from the same major disadvantage that recall does - namely, the fact that to measure recall we need to know the number of relevant documents in the collection, and to measure fallout we need to know the number of non-relevant documents in the collection, and these numbers are pragmatically difficult, if not impossible, to obtain. For calculating precision, however, we do not need to know anything about the collection, only about the results, and these values are readily obtainable. Therefore it has become generally accepted to use precision as a proxy for fallout, as it is practically easier to calculate.

However, precision is only a proxy - it does not give the same results as fallout. For example, consider a collection of 100 documents, where there are very few relevant ones, say 10, and all are retrieved. Recall will be 100 percent. If 10 more non-relevant documents are retrieved as well, precision will be 50 percent, but fallout will be 11 percent. Nevertheless, precision is the standard error measure because of its pragmatic calculatibility. Salton (Salton, 1983), in his discussion, says that the difference between precision and fallout is that precision is user oriented, while fallout is system oriented. It could be that what he means is that since recall is clearly system oriented, experts have chosen to use one system oriented measure and one user oriented measure, namely recall and precision, rather than two system oriented measures.

namely recall and fallout. However, in a sense, he is saying the same thing that we are. Fallout is system oriented, and its calculation therefore requires statistics about the system and collection, which are not easily obtainable. Precision is user oriented, and as such requires only statistics about the results given to the user, all easily obtainable. So Salton's distinction boils down to ours.

Whatever the case, precision is clearly an error measure. In fact, it is not an error measure, but a non-error measure. In other words, it actually shows the percentage of documents retrieved which <u>are</u> relevant. To get the error, or the percentage of documents retrieved which are <u>not</u> relevant, we would calculate 1-P. However, it is an error measure in the sense that it tells us how many errors the system has made through the calculation 1-P In light of this, a CSC which would combine recall and precision should be something that takes recall and reduces is by the degree of precision or non-precision in some way. The question is how? Should we use R-P? The problem with this is that if P is greater than R, we would get a negative number, and it is not clear exactly what this would mean. In light of the subsethood concept we introduced above, we would suggest the following.

We are looking for the degree of subsethood of the output vector(v), in the pertinence vector(w), which symbolizes the recall (or how many documents were relevant documents were retrieved), minus

the degree of non-subsethood of the pertinence vector in the output vector. which symbolizes the non-precision and is 100 percent minus the degree subsethood of the pertinence vector in the output vector. (As we said above, precision, or anything which is a proxy for or symbolizes precision, such as the degree of subsethood that we are using, is actually a non-error measure. So to get an error measure, we take 100% minus the non-error measure. That is why we said to subtract the degree of non-subsethood of the pertinence vector in the output vector.) Symbolically:

$$\frac{|v \cap w|}{|v|} - (1 - \frac{|w \cap v|}{|w|})$$

This then becomes:

$$F_4: \qquad \frac{\sum min(w,v)}{\sum v} - (1 - \frac{\sum min(w,v)}{\sum w}) \qquad (6.5.2)$$

For the example above, doing the appropriate calculations would give us

$$F_3(w_1 v_1) = \frac{.2+.6+0}{.2+.8+0} - (1 - \frac{.2+.6+0}{.3+.6+.2}) = \frac{.8}{1} - (1 - \frac{.8}{1.1}) = .5272$$

$$F_3(w_1 v_1) = \frac{.2+.6+0}{.2+.8+0} - (1 - \frac{.2+.6+0}{.3+.6+.6}) = \frac{.8}{1} - (1 - \frac{.8}{1.5}) = .333$$

Let us see what this resolves to in terms of recall and precision. It will be noticed that $\Sigma min(w,r)$ equals r, the sum of the fuzzy intersection between w and v. Σv equals N, or the number of relevant

documents in the output vector, and $\sum w$ equals C, or the number of relevant documents in the collection. Thus:

$$\frac{\sum min(w,v)}{\sum v}-(1-\frac{\sum min(w,v)}{\sum w})=$$

$$\frac{r}{N}-(1-\frac{r}{C}) =$$

$$\frac{r}{N}+\frac{r}{C}-1 =$$

$$\frac{rC}{CN}+\frac{rN}{CN}-1 =$$

$$\frac{r}{C}\cdot\frac{C}{N}+\frac{r}{N}\cdot\frac{N}{C}-1 =$$

$$R\cdot\frac{C}{N}+P\cdot\frac{N}{C}-1$$

Indeed, we would argue that this is a better measure than the additive measure above, 6.5.1,

$$\frac{\sum min(w,v)}{\sum v}+\frac{\sum min(w,v)}{\sum w}$$

because the additive measure can and does result in a value greater than 1, and what we are looking for is a percentage measure of results (meaning, for example, what percentage of perfect results our system exhibits). This subtractive measure will never result in a value greater than 1 by definition.

Using a similar derivation as above, we can show that 6.5.1 results in

$$\frac{r}{N}+\frac{r}{C}$$

We will use these derivations in the next section which discusses the "order preservation property".

Section 5

Order Preservation Property

A very important property of any composite evaluation measure, or indeed, any evaluation measure, of information retrieval systems, is what Voiskunskii calls the "order preservation property" (Voiskunskii, 1997). We can illustrate what is meant by the order preservation property by the use of the following example. Assume we have two arbitrary search methods. Assume also that both methods have been used in a search on the same query in the same search collection. We wish to evaluate the two search methods, and use some composite evaluation measure F_i to do so. We denote the value obtained for the first search method as F_1, and the value obtained for the second search method as F_2. Now if the sign of the difference $F_1 - F_2$ does not depend on the value of C (i.e., the number of relevant documents in the search collection). the composite measure F possesses the order preservation property. This will be the case if either C is not included in the expression of the difference, or if, with any admissible value of C, the sign of $F_1 - F_2$ will be the same. The reason this is called the order preservation property is that if a measure possesses this property, then, by using that measure, we can order the efficiency of different systems

from highest efficiency to lowest efficiency, without knowing or using the value of C (the number of relevant documents in the collection).

This not needing to know the value of C, is a very important property. A major problem with many evaluation measures, among them the standard recall measure, is the fact that in order to do the calculation, the actual number of relevant documents in the collection must be known. The problem is, how can we truly know this? Who perused the entire collection and evaluated which and how many documents were actually relevant to a particular query? (Salton, 1983, Voiskunskii, 1997, VonRijsbergen, 1979) (We actually alluded to this problem in passing above in Section 5.) It is true that the literature does offer methods to estimate the value of C (Salton, 1983, Van Rijsbergen, 1979), but these methods are just that - estimates.

Many methods suffer from the disadvantage of including C in the calculation. however, as long as we are comparing two different search strategies for the same search collection and the same query, our evaluation measure may still have the property that upon comparing the two values obtained, the difference between them does not depend on C, because the C's, being the same C in both values, is in some way canceled out of the difference, or otherwise eliminated from the difference. (We will illustrate clearly what this entails below.) If this is

the case, the measure would still possess the order preservation property.

Recall itself is an example of the above. As we mentioned above, the calculation of recall does indeed contain the value C ($R=\frac{r}{C}$). If two recall values evaluating two search strategies were obtained from different collections, then clearly the C's of each collection had to be known, or at least estimated. However, if the two search strategies were being tested on the same collection, then the C's being equal, can be factored out,

$$R_1-R_2=\frac{r_1}{C}-\frac{r_2}{C}=\frac{1}{C}(r_1-r_2)$$

and no matter what the value of C, we have a comparison between R_1 and R_2.

It is important to ascertain for each composite measure we propose whether or not it possesses this order preservation property. Those that possess it would be preferred to those that don't, since they would be substantially more useful in practice, as we would not have the problem of obtaining the value of C, as we explained.

Voiskunskii (Voiskunskii, 1997) shows that the cos measure does indeed have the order preservation property. He proves it as follows:

$$F_{11}-F_{12}=\sqrt{\frac{(r_1)^2}{CN_1}}-\sqrt{\frac{(r_2)^2}{CN_2}}=\frac{1}{\sqrt{C}}\left(\sqrt{\frac{(r_1)^2}{N_1}}-\sqrt{\frac{(r_2)^2}{N_2}}\right)$$

Now our first subsethood measure, F_2,which is somewhat equivalent to Van Rijsbergen's measure (Van Rijsbergen, 1979), is shown by Voiskunskii (Voiskunskii, 1997) not to have the order preservation property. First he shows what F_2 will be equal to in terms of C and N, and R and P.

$$\frac{|A \cap B|}{|A \cup B|} = \frac{r}{C+N-r} = \frac{1}{\frac{C}{r}+\frac{N}{r}-1} = \frac{1}{\frac{1}{R}+\frac{1}{P}-1}$$

So

$$F_{21} - F_{22} = \frac{r_1}{C+N_1-r_1} - \frac{r_2}{C+N_2-r_2}$$

Now, there is really no way to factor C out like we did with F_1 above. Voiskunskii further illustrates this by way of an example. assume we have two different search methods being used for the same search query on the same collection. Assume that the first method yielded an output of 20 documents of which 8 are evaluated as being truly relevant. The second method yielded an output of 12 documents, of which 6 are evaluated as being relevant. Consider two possible situations. Situation 1: The search collection contained 40 relevant documents (C=40). Then

$$\frac{r_1}{C+N_1-r_1} - \frac{r_2}{C+N_2-r_2} = \frac{8}{40+20-8} - \frac{6}{40+12-6} = \frac{8}{25} - \frac{6}{46} = \frac{368-312}{52 \cdot 46} > 0$$

Situation 2: the search collection contained 10 relevant documents. Then

$$\frac{r_1}{C+N_1-r_1}-\frac{r_2}{C+N_2-r_2}=\frac{8}{10+20-8}-\frac{6}{10+12-6}=\frac{8}{22}-\frac{6}{16}=\frac{128-132}{22\cdot 16}<0$$

So we see that in situation 1 and 2 the differences have opposite signs, and the only parameter distinguishing the two situations is the number of relevant documents in the search collection, or C. Hence evaluation measure F_2 does not possess the order preservation property.

As to the second and third subsethood methods, F_3 and F_4, we showed that F_3

$$\frac{|v\cap w|}{|v|}+\frac{|w\cap v|}{|w|}$$

reduces to

$$\frac{r}{N}+\frac{r}{C}$$

There is no way to factor out a C here, so it too does not possess the order preservation property. We can also prove this by using Voiskunskii's example above. Situation 1:

$$(\frac{r_1}{N_1}+\frac{r_1}{C})-(\frac{r_2}{N_2}+\frac{r_2}{C})=(\frac{8}{20}+\frac{8}{40})-(\frac{6}{12}+\frac{6}{40})=\frac{12}{20}-\frac{13}{20}<0$$

Situation 2:

$$(\frac{r_1}{N_1}+\frac{r_1}{C})-(\frac{r_2}{N_2}+\frac{r_2}{C})=(\frac{8}{20}+\frac{8}{10})-(\frac{6}{12}+\frac{6}{10})=\frac{6}{5}-\frac{5\frac{1}{2}}{5}>0$$

Again we see that in situation 1 and 2 the differences have opposite signs, and the only parameter distinguishing the two situations is the number of relevant documents in the search collection, or C. Hence evaluation measure F_3 also does not possess the order preservation property.

Now F_4 reduces to

$$\frac{r}{N}+\frac{r}{C}-1$$

Again there is no way to factor out C. Applying the numerical example.

Situation 1:

$$-\frac{1}{20}-1=-\frac{21}{20}$$

Situation 2:

$$\frac{1}{10}-1=-\frac{9}{10}$$

The value does not reverse from positive to negative, but the two values are not equal , so F_4 does not possess the order preservation property.

To conclude, then, of the four measures we suggest. only F_1 possesses the important property of order preservation, while the others, F_2, F_3, and F_4, do not.

Chapter VII

Conclusion and Summary

In this thesis we have explored an area that has only been touched upon until now - namely, the area of non-binary evaluation of information retrieval systems. We have first explored it in terms of the traditional measures of R and P. In this part of the analysis, we have described carefully how to calculate R and P in the non-binary context, something that has never been explicitly addressed before in the literature. We suggest two separate methods of calculating R and P, and explain the differences between the two methods.

We then take the analysis further into the area of *composite* non-binary evaluation of information systems. This part is indeed the major thrust of the paper, as it is an area that has stymied those few who have attempted it in the past. Those who have attempted (Voiskunskii, 1997) to develop such a measure, have concluded that their attempts were unsuccessful. We have used fuzzy set techniques to solve this problem and have developed four non-binary composite measures.

The first measure is an extension of a suggestion by Voiskunskii to use a cos measure in the binary case to calculate to what degree two

vectors coincide with each other. When Voiskunskii tries to extend this measure to the non-binary case he is unsuccessful. We use fuzzy set techniques and successfully make the transition to the non-binary case. We follow this with an extensive discussion of our measure, comparing it with Voiskunskii's unsuccessful attempt, showing the underlying reason why Voiskunskii's attempt didn't work, and in general delineating the characteristics and limitations of our measure and its superiority to Voiskunskii's measure through the use of actual examples and their results.

We then use the concept of subsethood, which is also a fuzzy set concept, to develop three additional measures. The first is a general measure of the degree to which the output vector is a subset of the pertinence vector and the pertinence vector is a subset of the output vector. The degree to which the two vectors are subsets of each other will be a measure of their coincidence. We proceed from here to two additional suggestions involving the concept of subsethood. One is an additive measure. That is, our composite measure will be the degree subsethood of the output vector in the pertinence vector plus the degree of subsethood of the pertinence vector in the output vector. The other is a subtractive measure. Noting that the degree subsethood is a recall concept, and the degree of subsethood of the pertinence vector in the output vector is a precision concept, and that precision is a kind of

error measure, our composite measure will be the degree of subsethood of the output vector in the pertinence vector minus the degree of subsethood of the pertinence vector in the output vector.

We succeeded in developing four composite non-binary measures, one cos measure and three subsethood measures. None is perfect, however. Each of these two categories has advantages and disadvantages. The disadvantages of the cos measure are the problematic cases we cite in the body of the discussion itself, which, in fact, we are able to resolve. However, the sum total of a number of problematic cases, albeit explained in various ways, leaves us a bit uneasy, as we say in the discussion itself. This is the motivation to develop the other measures, i.e., the subsethood measures. However, these measures themselves have a major disadvantage in that they do not possess the important "order preservation property" while the cos does. This is discussed at length in Section 4 of the previous chapter. So each of the measures we developed has its own advantages and disadvantages, and therefore they all have a role to play.

To conclude, then, we have extensively explored the area of non-binary composite measurement of information retrieval systems through the use of fuzzy set techniques, and have succeeded in developing some useful results, which is something that the few experts in the field who have addressed this question have not been able to do. As a result, we

obtained some original results which advance the state of the knowledge in this area. We hope that these results will be expanded upon in the future by ourselves and others.

Bibliography

Anderberg, M.R. (1973). *Cluster Analysis for Applications.* Academic Press, New York.

Bezdek, J.C. (1981). *Pattern Recognition With Fuzzy Objective Function Algorithms.* Plenum, New York.

Bezdek, J.C., Biswas, G., Huang, L. (1986). "Transitive Closures of Fuzzy Thesauri for Information Retrieval Systems." *International Journal of Man-Machine Studies.* Vol. 25, pp. 343-356.

Bezdek, J.C., Biswas, G., Huang, Li-Ya. (1986). "Transitive Closures of Fuzzy Thesauri for Information Retrieval Systems." *International Journal of Man-Machine Studies.* Vol. 25, pp. 343-356.

Bezdek, J.C., ed. (1987). *Analysis of Fuzzy Information.* CRC Press, Boca Raton, Florida.

Biswas, G., Bezdek, J.C., Marques, M., Subramanian, V. (1987). "Knowledge Assisted Document Retrieval: II. The Retrieval Process." *Journal of the American Society for Information Science.* Vol. 38, pp. 97-110.

Bollman, P. (1977). "A Comparison of Evaluation Measures for Document Retrieval Systems." *Journal of Informatics.* Vol. 1, pp. 97-116.

Bollman, P., Raghavan, V.V., Jung, G.S., Shu, L.C. (1992). "On Probabilistic Notions of Precision as a Function of Recall." *Information Processing and Management.* Vol. 28, pp. 291-315.

Boy, G.A., Kuss, P.M. (1986). "A Fuzzy Method for the Modeling of Human-Computer Interaction in Information Retrieval Tasks." In: Karwowski, W., Mital, A. eds. *Applications of Fuzzy Set Theory in Human Factors.* Elvesier, New York.

Buckles, B., Petry, F. (1982). "A Fuzzy Model for Relational databases." *Fuzzy Sets and Systems.* Vol. 7, pp. 213-226.

Buckles, B., Petry, F. (1983). "Information-Theoretical Characterization of Fuzzy relational Databases." *IEEE Transactions on Systems, Man, and Cybernetics.* Vol. 13, pp. 74-77.

Carpentere, M.P., Narin, F. (1973). "Clustering of Scientific Journals." *Journal of the American Society for Information Science.* Vol. 24, pp. 425-436.

Cawkell, A.E. (1975). "A Measure of 'Efficiency Factor' - Communication Theory Applied to Document Selection Systems." *Information Processing and Management.* Vol. 11, pp. 243-248.

Cherniavsky, V.S., Lakhuti, D.G. (1970). "The Problem of Evaluating Retrieval Systems." *Nauchno-Teknicheskaya Informatsiya (NTI).* Ser.2, No. 1, pp. 24-34.

Cleverdon, C.W. (1972). "On the Inverse Relationship of Recall and Precision." *Journal of Documentation.* Vol. 28, pp. 195-201.

Cleverdon, C.W., (1970). "Evaluation of Tests of Information Retrieval Systems." *Journal of Documentation.* Vol. 26, pp. 55-67.

Cleverdon, C.W., Mills, J., Keen, M. (1966). "Factors Determining the Performance of Indexing Systems." *ASLIB Cranfield Project.* Cranfield.

Comparative Systems Laboratory. (1968). *An Inquiry Into Testing of Information Retrieval Systems.* Case Western Reserve University.

Cooper, W.S. (1968). "Expected Search Length: A single Measure of Retrieval Effectiveness Based on Weak Ordering Action of Retrieval Systems." *Journal of the American Society for Information Science.* Vol. 19, pp. 30-41.

Cooper, W.S. (1971). "A Definition of Relevance for Information Retrieval." *Information Storage and Retrieval.* No. 7, pp. 19-37.

Cooper, W.S. (1973). "On Selecting A Measure of Retrieval Effectiveness." *Journal of the American Society for Information Science.* Vol. 24, pp. 87-100.

Cooper, W.S. (1973). "On Selecting A measure of Retrieval Effectiveness." *Journal of the American Society for Information Science.* Vol. 4, pp. 87-100. 413-424.

Dubois, D., Prade, H. (1980). *Fuzzy Sets and Systems, Theory and Application.* Academic Press, New York.

Frants, V.I., Brush, C.B. (1988). "The Need for Information and Some Aspects of Information Retrieval Systems Construction." *Journal of the American Society for Information Science.* Vol. 39, pp. 86-91.

Frants, V.I., Shapiro, J. (1991). "Algorithm for Automatic Construction of Query Formulations in Boolean Form." *Journal of the American Society for Information Science.* Vol. 42, pp. 16-26.

Frants, V.I., Shapiro, J. (1991). "Control and Feedback in a Documentary Information Retrieval System." *Journal of the American Society for Information Science.* Vol. 42, pp. 623-634.

Frants, V.I., Shapiro, J., Voiskunskii, V.G. (1993). "Multiversion Information Retrieval Systems and Feedback with Mechanism of Selection." *Journal of the American Society for Information Science.* Vol. 44, pp. 19-27.

Frants, V.I., Voiskunskii, V.G., Frants, Y.I. (1970). "Evaluation of the Magnitude of Possible Losses of Information During Indexing." *Nauchno-Teknicheskaya Informatsiya (NTI).* Ser.2, No. 5, pp. 14-15.

Goffman, W. (1964). "On Relevance As a Measure." *Information Storage and Retrieval.* No. 2, pp. 201-203.

Goffman, W., Newill, V.A. (1966). "A Methodology for Test and Evaluation of Information Retrieval Systems." *Information Storage and Retrieval.* No. 3. pp. 19-25.

Good, I.J. (1967). "The Decision Theory Approach to the Evaluation of Information Retrieval Systems." *Information Storage and Retrieval.* No. 3. pp. 31-34.

Guazzo, M. (1977). "Retrieval Performance and Information Theory." *Information Processing and Management.* Vol. 13, pp. 155-165.

Heine, M.H. (1973). "Distance Between Sets as an Objective measure of Retrieval Effectiveness." *Information Storage and Retrieval.* No. 9, pp. 181-198.

Heine, M.H. (1973). "The Inverse Relationship of Precision and recall In Terms of the Swets' Model." *Journal of Documentation.* Vol. 29, pp. 81-84.

Keen, E.M. (1967). "Evaluation Parameters." *Report ISR-13 to the National Science Foundation, Section II, Cornell University, Department of Computer Science.*

Keen, E.M., Digger, J.A. (1972). *Report of an Information Science Index Language Test.* Aberystwyth College of Librarianship, Wales.

King, D.W., Bryant, E.C. (1971). *The Evaluation of Information Services and Products.* Information Resources Press, Washington.

Klir, G.J. (1987). "Where Do we Stand on measures of Uncertainty, Ambiguity, Fuzziness, and the Like?" *Fuzzy Sets and Systems.* Vol. 38, pp. 141-160.

Klir, G.J. (1991) "Generalized Information Theory." *Fuzzy Sets and Systems.* Vol. 40, pp. 127-142.

Klir, G.J., Folger, T. (1988). *Fuzzy Sets, Uncertainty, and Information.* Prentice Hall, Englewood Cliffs, New Jersey.

Klir, G.J., Mariano, M. (1987). "On the Uniqueness of Possibilistic Measure of Uncertainty and Information." *Fuzzy Sets and Systems.* Vol. 24, pp. 197-219.

Klir, J.G., Yuan, Bo. (1995). *Fuzzy Sets and Fuzzy Logic, Theory and Applications.* Prentice Hall, New Jersey.

Kohout, L.J., Keravnou, E., Bandler, W. (1984). "Automatic Documentary Information Retrieval by Means of Fuzzy Relational Products." In: Zimmerman, H.J., Zadeh, L.A., Gaines, B.R., eds. *Fuzzy sets and Decision Analysis.* North-Holland, New York.

Kraft, D., Bookstein, A. (1978). "Evaluation of Information Retrieval Systems: A Decision Theoretic Approach." *Journal of the American Society for Information Science.* Vol. 29, pp. 31-40.

Lancaster, F.W. (1979). *Information Retrieval Systems: Characteristics, Testing Evaluation.* John Wiley and Sons, New York .

Larsen, L.H., Yager, R.R. (1990). "An approach to Customized End-User Views of Multi-User Information Retrieval systems." In: Kacprzyk, J., Fedrizzi, M., eds. *Multiperson Decision Making Models Using Fuzzy Sets and possibility Theory.* Kluwer, Boston.

Larsen, L.H., Yager, R.R. (1993). "The Use of Fuzzy Relational Thesauri for Classificatory Problem Solving in Information Retrieval and Expert Systems." *IEEE Transactions on Systems, Man and Cybernetics.* No. 23, pp. 31-41.

Li, C.J., Liu, D.B. (1990). *A Fuzzy Prolog Database System.* John Wiley, New York.

Lopez de Mantaras, R., Cortes U., Manero, J., Plaza, E. (1990). "Knowledge Engineering for a Document Retrieval System." *Fuzzy Sets and Systems.* Vol. 38, pp. 223-240.

Medina, J.M., Pons, O., Vila, M.A. (1994). "GEFRED: A General ModelOf Fuzzy Relational Databases." *Information Sciences: Informatics and Computer Sciences.* Vol. 76, pp. 87-109.

Miyamoto, S. (1990). *Fuzzy Sets In Information Retrieval and Cluster Analysis.* Kluwer Academic Publishers, Dordecht, Boston, London.

Murai, T., Miakoshi, M., Shimbo, M. (1988). "A Model of Search Oriented Thesaurus Use based on Miltivalued Logic Inference." *Information Sciences.* Vol. 45, pp. 185-215.

Murai, T., Miakoshi, M., Shimbo, M. (1989). "a Fuzzy Document Retrieval Method Based on Two-Valued Indexing." *Fuzzy Sets and Systems.* Vol. 30, pp. 103-120.

Negoita, C.V. (1973). "On the Notion of Relevance." *Kyberntes.* Vol 2, pp. 161-165.

Negoita, C.V. (1973). "On the Application of Fuzzy Sets Separation Theorem for Automatic Classification in Information Retrieval Systems." *Information Sciences.* Vol. 5, pp. 279-286.

Negoita, C.V., Flondor, P. (1976). "On Fuzziness in Information Retrieval." *International Journal of Man-Machine Studies.* Vol. 8, pp. 711-716.

Nomoto, K., Wakayama, S., Kirimoto, T., Ohashi, Y., Kondo, M. (1990). "A Document Retrieval System Based on Citations Using Fuzzy Graphs." *Fuzzy Sets and Systems.* Vol.38, pp. 207-222.

Ogawa, Y., Morita, T., Kobayashi, K. (19910. "A Fuzzy Document Retrieval System Using the Keyword Connection Matrix and A learning Method." *Fuzzy Sets and Systems.* Vol.39, pp. 163-179.

Prade, H., Testemale, C. (1984). "Generalizing Database Relational Algebra for the Treatment of Incomplete or Uncertain Information and Vague Queries." *Information Sciences.* Vol. 34, pp. 115-143.

Radecki, T. (1981). "Outline of a Fuzzy Logic Approach to Information Retrieval." *International Journal of Man-Machine Studies.* Vol. 14, pp. 169-178.

Radecki, T. (1983). "A Theoretical Background for Applying Fuzzy set Theory in Information Retrieval." *Fuzzy Sets and Systems.* Vol.10, pp. 169-183.

Raghavan, V.V., Bollman, P., Jung, G.S. (1989). "Retrieval System Evaluation Using recall and precision: Problems and Answers." *Proceedings of the Twelfth*

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Cambridge, MA. pp. 59-68.

Raju, K.V., Majumdar, A.K. (1988). "Fuzzy Functional Dependencies and Lossless Join Decomposition of Fuzzy Relational Database Systems." *ACM Transactions on Database Systems.* No. 13, pp. 129-166.

Robertson, S.E. (1969). "The Parametric Description of Retrieval Tests." *Journal of Documentation.* Vol. 25, pp. 1-27.

Robertson, S.E. (1977). "The Probabilistic Character of Relevance." *Information Processing and Management.* Vol. 13, pp. 247-251.

Robertson, S.E., Teather, D. (1974). "A Statistical Analysis of Retrieval Tests: A Bayesian Approach." *Journal of Documentation.* Vol. 30, pp. 273-282.

Salton, G. (1968). *Automatic Information Organization and Retrieval.* McGraw Hill, New York.

Salton, G., Mcgill, M.J. (1983). *Introduction to Modern Information Retrieval.* McGraw Hill, New York.

Saracevic, T. (1995). "Evaluation of Evaluation in Information Retrieval." *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA.* pp. 138-146

Saracevic, T. (1975). "Relevance: A review of and A Framework for the Thinking on the Notion in Information Science." *Journal of the American Society for Information Science.* Vol. 26, pp. 321-343.

Shaw, W.M., (1995). "Term-Relevance Computation and Perfect Retrieval Performance." *Information Processing and Management.* Vol. 31, pp.491-498.

Shenoi, S., Melton, A. (1989). "Proximity Relations in the Fuzzy Relational Database Model." *Fuzzy Sets and Systems.* Vol. 31, pp. 285-296.

Shenoi, S., Melton, A. (1990). "An Extended Version of the Fuzzy Relational Data Model." *Information Sciences.* Vol. 52, pp. 35-52.

Shenoi, S., Melton, A., Fan, L.T. (1990). "An Equivalence Class Model of Fuzzy Relational Databases." *Fuzzy Sets and Systems.* Vol. 38, pp. 153-170.

Shenoi, S., Melton, A., Fan, L.T. (1992). "Functional Dependencies and Normal Forms in the Fuzzy Relational Database Model." *Information Sciences.* Vol. 60, pp. 1-28.

Sparck Jones, K. (1978). "Performance Averaging for Recall and Precision." *Journal of Informatics.* Vol. 2, pp. 95-105.

Swets, J.A. (1963). "Information Retrieval Systems." *Science.* No. 141, pp. 245-250.

Swets, J.A. (1967). *Effectiveness of Information Retrieval Methods.* Bolt, Beranek, and Newman, Cambridge, Mass.

Tong, R.M. (1986). "The Representation of Uncertainty in an Expert System for Information Retrieval." In: Prade, H., Negoita, C.V., eds. *Fuzzy Logic in Knowledge Engineering.* Verlag TUV Rheinland, Koln.

Tripathy, R.C., Saxena, P.A. (1990). "Multivalued Dependencies in Fuzzy Relational Databases." *Fuzzy Sets and Systems.* Vol. 38, pp. 267-279.

Umano. M. (1982). "FREEDOM-O: A Fuzzy Database System." In: Gupta, M.M.. Sanchez, E., eds. *Fuzzy Information Knowledge Representation and Decision Analysis.* Pergamon Press, Oxford.

Van Rijsbergen, C.J. (1979). *Information Retrieval.* Buttersworth, London.

Voiskunskii, V.G. (1980). "Search Space and Documentary Search." *Nauchno-Teknicheskaya Informatsiya (NTI).* Ser.2, No. 9, pp. 17-22.

Voiskunskii, V.G. (1983). "One Class of Search Characteristics." *Nauchno-Teknicheskaya Informatsiya (NTI).* Ser.2, No. 8, pp. 12-15.

Voiskunskii, V.G. (1984). "The Distance in N-Dimensional Vector Space and Search Characteristics." *Nauchno-Teknicheskaya Informatsiya (NTI).* Ser.2, No. 1. pp. 18-20.

Voiskunskii, V.G. (1987). "Applicability of Search Characteristics." *Nauchno-Teknicheskaya Informatsiya (NTI).* Ser.2, No. 12, pp. 18-24.

Voiskunskii, V.G. (1992). "Construction of Search Characteristics." *Nauchno-Teknicheskaya Informatsiya (NTI).* Ser.2, No. 9, pp. 6-9.

Voiskunskii, V.G. (1997). "Evaluation of Search Results: A New Approach." *Journal of the American Society for Information Science.* Vol. 48, No. 2, pp. 133-142.

Weiler, G. (1962). "On Relevance." *Mind.* Vol. LXXI, pp. 487-493.

Zemankova, M. (1989). "FILIP: A Fuzzy Intelligent Information System With Learning capabilities." *Information Systems.* Vol 40, pp. 473-486.

Zemankova, M., Kandel, A. (1985). "Implementing Imprecision in Information Systems." *Information Sciences.* Vol. 37, pp. 107-141.

Zenner, R.B.R., De Caluwe, R.M.M., Kerre, E.M. (1985). "A New Approach to Information Retrieval Systems Using Fuzzy Expressions." *Fuzzy Sets and Systems.* Vol.17, pp. 9-22.

תם ונשלם שבח לקל בורא עולם