
# NAVAL POSTGRADUATE SCHOOL

**MONTEREY, CALIFORNIA**

# THESIS

**AUTOMATED PSYCHOLOGICAL CATEGORIZATION VIA LINGUISTIC PROCESSING SYSTEM**

by

Mark D. Eramo
and
Christopher M. Sutter

September 2004

| | |
|---|---|
| Thesis Advisor: | Raymond Buettner |
| Thesis Co-Advisor: | Magdi Kamel |

**Approved for public release; distribution is unlimited**

THIS PAGE INTENTIONALLY LEFT BLANK

| REPORT DOCUMENTATION PAGE | | | Form Approved OMB No. 0704-0188 |
|---|---|---|---|
| Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503. | | | |
| **1. AGENCY USE ONLY** *(Leave blank)* | **2. REPORT DATE** September 2004 | **3. REPORT TYPE AND DATES COVERED** Master's Thesis | |
| **4. TITLE AND SUBTITLE**: Automated Psychological Categorization via Linguistic Processing System | | | **5. FUNDING NUMBERS** |
| **6. AUTHOR(S)** Mark D. Eramo and Christopher M. Sutter | | | |
| **7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)** Naval Postgraduate School Monterey, CA 93943-5000 | | | **8. PERFORMING ORGANIZATION REPORT NUMBER** |
| **9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)** N/A | | | **10. SPONSORING/MONITORING AGENCY REPORT NUMBER** |
| **11. SUPPLEMENTARY NOTES** The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. | | | |
| **12a. DISTRIBUTION / AVAILABILITY STATEMENT** Approved for public release; distribution is unlimited | | | **12b. DISTRIBUTION CODE** |


**13. ABSTRACT (maximum 200 words)**

Influencing one's adversary has always been an objective in warfare. However, to date the majority of influence operations have been geared toward the masses or to very small numbers of individuals. Although marginally effective, this approach is inadequate with respect to larger numbers of high value targets and to specific subsets of the population. Limited human resources have prevented a more tailored approach, which would focus on segmentation, because individual targeting demands significant time from psychological analysts. This research examined whether or not Information Technology (IT) tools, specializing in text mining, are robust enough to automate the categorization/segmentation of individual profiles for the purpose of psychological operations (PSYOP). Research indicated that only a handful of software applications claimed to provide adequate functionality to perform these tasks. Text mining via neural networks was determined to be the best approach given the constraints of the profile data and the desired output. Five software applications were tested and evaluated for their ability to reproduce the results of a social psychologist. Through statistical analysis, it was concluded that the tested applications are not currently mature enough to produce accurate results that would enable automated segmentation of individual profiles based on supervised linguistic processing.


| **14. SUBJECT TERMS** Text Mining, Data Mining, Automated Psychological Categorization, Automated Psychological Segmentation, Linguistic Processing, Automated Linguistic Processing, Precision Influence, Automated Precision Influence, Information Operations, Influence Operations, Psychological Operations, PSYOP | | | **15. NUMBER OF PAGES** 139 |
|---|---|---|---|
| | | | **16. PRICE CODE** |
| **17. SECURITY CLASSIFICATION OF REPORT** Unclassified | **18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified | **19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified | **20. LIMITATION OF ABSTRACT** UL |

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

**Approved for public release; distribution is unlimited**

**AUTOMATED PSYCHOLOGICAL CATEGORIZATION VIA LINGUISTIC PROCESSING SYSTEM**

Mark D. Eramo
Captain, United States Marine Corps
B.S., University of Kansas, 1996

Christopher M. Sutter
Lieutenant, United States Navy
B.S., Hampden-Sydney College, 1996

Submitted in partial fulfillment of the
requirements for the degrees of

**MASTER OF SCIENCE IN INFORMATION TECHNOLOGY MANAGEMENT**

**and**

**MASTER OF SCIENCE IN INFORMATION SYSTEMS AND OPERATIONS**

from the

**NAVAL POSTGRADUATE SCHOOL**
**September 2004**

Authors: Mark Eramo

Christopher Sutter

Approved by: Raymond Buettner
Thesis Advisor

Magdi Kamel
Thesis Co-Advisor

Dan C. Boger
Chairman, Department of Information Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

# ABSTRACT

Influencing one's adversary has always been an objective in warfare. However, to date the majority of influence operations have been geared toward the masses or to very small numbers of individuals. Although marginally effective, this approach is inadequate with respect to larger numbers of high value targets and to specific subsets of the population. Limited human resources have prevented a more tailored approach, which would focus on segmentation, because individual targeting demands significant time from psychological analysts. This research examined whether or not Information Technology (IT) tools, specializing in text mining, are robust enough to automate the categorization/segmentation of individual profiles for the purpose of psychological operations (PSYOP). Research indicated that only a handful of software applications claimed to provide adequate functionality to perform these tasks. Text mining via neural networks was determined to be the best approach given the constraints of the profile data and the desired output. Five software applications were tested and evaluated for their ability to reproduce the results of a social psychologist. Through statistical analysis, it was concluded that the tested applications are not currently mature enough to produce accurate results that would enable automated segmentation of individual profiles based on supervised linguistic processing.

THIS PAGE INTENTIONALLY LEFT BLANK

# TABLE OF CONTENTS

# LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

# ACKNOWLEDGMENTS

This thesis is the compilation of many people's individual efforts. We would like to extend the utmost respect and appreciation to Anthony Pratkanis, PhD., for his unwavering support and extraordinary efforts. Without his support and expertise in the realm of social psychology, this project could not have been completed. Anthony's desire to give of himself in support of securing the American way of life is a model for all Americans to follow.

We would also like to acknowledge the efforts of LCDR Russell Gottfried, an instructor at the Naval Postgraduate School in Operations Research. His assistance and untiring dedication was instrumental in the ability to logically examine and concisely report the results of the research. He is truly dedicated to teaching students and embodies the characteristics needed by all academic instructors.

The testing and evaluation phases are a tribute to the outstanding vendor support that we received. Each vendor provided both evaluation copies of their software free of charge and the technical support needed to implement the tools. The vendors were:

- Nicko Goncharoff at ReelTwo Inc.
- Richie Kasprzycki at Megaputer Intelligence
- Bob Gray, Courtney Henson, and Shelby Ross at SAS
- David A. Luzier at SER Solutions Inc.
- Michael Falkowski at SPSS

For support in the realm of software acquisition we would have to thank Lillian Gassie, the senior systems librarian at Knox library. She was instrumental in the acquisition of one of the software tools; without her assistance our evaluation of the technologies would have been incomplete.

Lastly, we would be remiss if we did not acknowledge both our wives Danielle and Dana. Their ability to endure our long hours of research when we could have been spending time with the family, especially during the school breaks, is telling of their dedication to our efforts.

THIS PAGE INTENTIONALLY LEFT BLANK

# I. BACKGROUND

The instruments, or means, of national power/statecraft have been defined in four broad categories: political, economic, military, and psychological (now called informational).[1] With these tools, nation-states strive to achieve their strategic and tactical objectives driven by their national interests. It is commonplace to observe the United States applying political, economic, and military pressure on other nations for the purposes of national interests. The last instrument, informational, has also been in constant use for many years but it has been less visible. However, even though the informational instrument of statecraft is harder to interpret, it is becoming more common.[2]

The informational approach is a means to influence one's adversary. These operations have been around since the beginning of recorded history. The United States began using informational operations when leaflets were distributed to British troops during the Revolutionary War. The leaflets were designed to 'persuade' the enemy troops that they were not fighting for a reasonable cause. However, the new government of the United States was reluctant to use such tactics and the resulting campaign was a halfhearted attempt. After the war, such operations were viewed as "too dirty" and ineffective by the politicians in Washington. Thus, the informational instrument was not significantly employed again until World War II when the United States began using, in a covert manner, psychological operations (PSYOP) as a means to influence the enemy.

The informational instrument of statecraft is arguably the most important, but not widely accepted as such because the operations are not widely known and the benefits are not easily measured. It has been dubbed Information Operations by the United States military and includes: Psychological Warfare, Deception, Operational Security, Electronic Warfare, and Computer Network Operations.

Most people tend to focus on the first three instruments of statecraft because they are easy to identify. However the informational instrument, specifically the

[1] Association of the United States Army. *Army*. Vol. 13, No. 5. Washington, D.C. December 1962.

[2] Radvanyi, Janos. *Psychological Operations and Political Warfare in Long-term Strategic Planning*. Praeger Publishers, New York, New York. 1990.

psychological approach, can have tremendous effects on the first three instruments by changing an adversary's perception without always being obvious. Additionally, the use of PSYOP has the potential to achieve results without the use of other statecraft methods. For this reason, the informational arm of statecraft must be brought to the forefront regarding future operations and PSYOP must be given greater attention.

PSYOP is not something new; it has been a powerful tool throughout history. Even the ancient philosopher Sun Tzu showed an understanding of psychological operations when he wrote about "subduing the enemy's army without battle." Influencing one's adversary has always been an objective in both statecraft and warfare because it provides the ability to attain 'ends' without more costly kinetic 'means'. Although physical battle is needed in various conflicts, the use of force does not necessarily change the way people think. For example, if all terrorists were killed tomorrow, the existence of terrorism would not die along with them. The idea would remain. The best way to terminate an idea is to change the way those marketing the idea think. PSYOP is a tool that works to achieve that goal.

However, to date in both military and non-military actions, the majority of influence operations have targeted the mass audiences with little regard for the individual differences among the population. Examples of this are widespread: radio broadcasts, television airings, leaflet droppings, etc. Although somewhat effective, this approach is careless both with respect to high value targets and to specific subsets of the population. It is intuitive that high value targets must receive more personalized attention in order to conduct precision influence. For the remainder of the target population (several orders of magnitude larger) however, it is infeasible to provide individual attention to every target (person) since the group is too large; thus concessions are made. These concessions are in the area of accuracy; accuracy of delivery, precision of influence and exactness of desired results. When the target is an entire country's population, one message will not have the same effect on each person, nor will it solicit the same result.

When attempting to influence a large number of targets which can not be addressed individually, a different strategy must be formulated to tackle the problem. One strategy is to 'get the best result possible', by sending one message to the masses

while precisely targeting a very small number of high value targets. This type of strategy does not focus on specific groups but rather identifies a small number of individuals to target and ignores the idiosyncrasies of the remaining masses. Examples can be seen in the dropping of leaflets, or the random broadcasts via radio.

## A. MARKETING TECHNIQUES AND PSYOP

Trying to persuade people to think a specific way is not an uncommon task. Marketing has been employing similar techniques for years in an effort to influence potential customers to buy particular products or act in certain manner. The method in which a marketer undertakes the business of influencing customers is similar to psychological targeting, but the order of steps in the process is slightly different. Marketers first try to segment the market in which they wish to influence, then they target an identified segment, then they position themselves for maximum effectiveness. Marketers identify their target markets through market research and then refine developed communication methods to deliver a desired message to the target. The target market or target audience can be segmented, and traditionally is, in order to maximize the results of the desired actions. But in some cases segmentation is not conducted, but that is not germane for our purposes.

Segmentation is the development of subsets within a group/audience/market for which those subsets share similarities that can be influenced/targeted all together as a group.[3] The goal of segmentation, specifically within the context of marketing, is to match groups of purchasers with the same set of needs and buyer behavior. When that group/subset is established it is considered to be a segment.

When conducting segmentation in consumer markets certain criteria are used to ensure the segments that are created are useful. Again, from the perspective of business the criteria are:

- Segment identity: Each determined segment must be measurable for the purposes of specifically identifying the determined segment.
- Segment accessibility: Is the potential segment able to be tapped, can the business actually get into and influence the created segment?

[3] Wedel, Michel, And Wagner A. Kamakura. Market Segmentation: Conceptual and Methodological Foundations. International Series in Quantitative Marketing, Vol. 8. Kluwer Academic Publishers; 1st Edition, November 1999.

- Segment substantiality: Is the segment sufficient in size to justify the resources required to target it?
- Segment uniqueness: Is the segment unique enough to respond to differently to a myriad of marketing mixes?
- Segment durability: Is the segment stable? Segment stability minimizes the cost of frequent changes in marketing campaigns in order to successfully access the segment.

Segments may be determined or considered in many ways, it is similar to sorting by multiple variables and providing the output groupings. There are many bases for producing segments, some of the more general bases are: geography, psychographics, socio-cultural factors, and demographics. Geography is related to physical location on the globe; it may be where the product is bought, where the product is used, where the consumers of the product live, etc. Psychographics refers to the consumer's personality and emotionally-based behavior that is linked to purchase choices. Psychographic variables include: risk thresholds, impulsiveness, lifestyle, beliefs, morals, and general attitudes of the individual. Socio-cultural factors include the consumer's socio economic class, education level, job status, and several other cultural issues. Demographic segments may try to break the groups down by variables such as: age, gender, income, ethnicity, nationality, religion, martial status, and size of household. Using these segmentation methods marketers produce subcategories and proceed to examine the identified segment.

The next step in marketing is to identify and mold the persona of the target or targets that were segmented by any other defined methods. For business there are three main types of targeting.

- Single segment with a single product
- Single product with all segments
- Multi-segment approach

The single segment approach targets one segment with one product. This would be the equivalent of what PSYOPS calls precision influence. One message is generated and delivered to one specific segment or even an individual within the population that is

known to work. Marketing campaigns have different goals, but if it is assumed that action is the goal, this method could be the most profitable; but it disregards the remaining segments.

The single product with all segments approach is traditionally called "mass marketing." Regardless of differences among segments the same message/product is produced for delivery. This method is the 'shot-gun' approach, it is designed to be an inexpensive widespread delivery technique, but traditionally fails to be the most successful method of influence if measured in consumer action. This would be the equivalent of dropping the same PSYOP leaflet on an entire country regardless of differences among the population.

The last method is the multi-segment approach. In this method marketers target the identified segments, conscious of the segment differences. Different products/messages will be crafted that will provide the best results (actions) among the segments. Each segment is treated separately and identified as unique/discriminate customers. This is the approach explained in this research.

The final part of marketing is the process of 'positioning'. Positioning is where the product/message perception is created for the customers. Businesses try to position their products in a unique way from their competitors to attain competitive advantage (positioning strategy). A simplistic example of positioning would be to compare products by price versus quality. The positioning of a product is what creates the product's perception in the mind of the target (consumer), if the target's perception of the product is ideal for their situation then the positioning was correct.

Marketing has refined methods and tactics that have been developed over time to meet its objective. It is highly successful and is the baseline for all businesses that compete for consumer attention and ultimately the consumer's purchasing power. How a product is marketed directly affects sales. That is the critical link between marketing and psychological operations.

Although marketing and PSYOP may not seem to be intuitively linked, they certainly have similarities. Both are trying to influence people to view an issue in a specific way in order to direct a desired action. In marketing terms that concept is called positioning and in psychology it is called framing[4].

## B. CURRENT STATE OF PSYOP SEGMENTATION

However, neither marketing nor PSYOP have perfected the art of influence. The process currently used to determine what influence strategy to employ on an individual is lengthy and requires vast human effort.[5] Figure one shows the current scope of PSYOP tools. The "face-to-face" method is concurrent with specific, individual targeting because it involves a personal analysis of what the target will respond most favorably to.

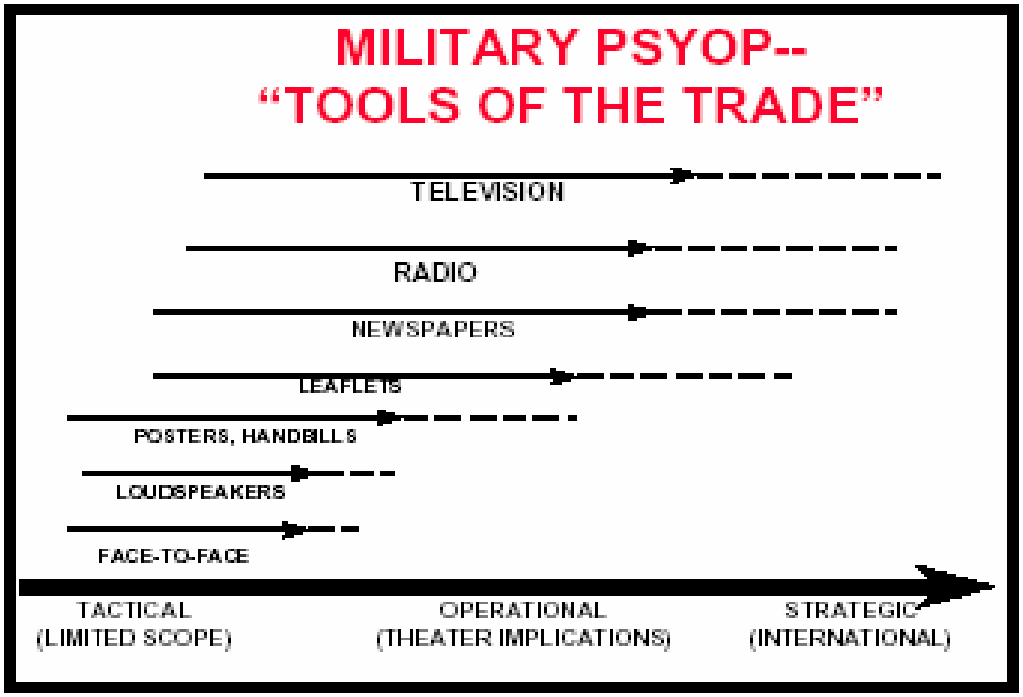


Figure 1. Military PsyOp "Tools of the Trade" (From: http://www.fi.uib.no/~antonych/deza/deza.html. August 2004.)

The resources needed to identify the most influential attributes for each target, especially when targeting a large group, are not readily available. Therefore, the current

[4] Rhoad, Kelton. "Working Psychology." Los Angeles. www.workingpsychology.com. August 2004.

[5] Office of the Under Secretary of Defense For Acquisition, Technology and Logistics. Report of the Defense Science Board Task Force on *The Creation and Dissemination of All Forms of Information in Support of Psychological Operations (PSYOP) in Time of Military Conflict*. Washington, D.C. 20301-3140. May 2000.

process may only have quality effects on the individuals that received personalized analysis whereas the remaining masses are essentially spammed with a bland, bulk message. For this reason, a more efficient method of analyzing targets is needed.

Although bulk messages (potentially a leaflet or a spammed email) can provide results, their effectiveness can be improved. This bulk effort ignores the fact that some targets will be unaffected by the message. Worse yet, some targets may actually be offended by a particular message and their resolve hardened against the PSYOP effort. For example, if a target is power hungry and the message appeals to that desire by offering tangible benefits, the outcome will have a high probability of success. However, if the target is an honor-driven individual with personal standards, the message will likely insult him or her and thus destroy a potentially fruitful relationship. In most cases, only one chance exists to influence a target and the bulk method is currently where efforts are focused. Thus the current mythology for influencing large groups of targets has significant room for improvement.

The proposed strategy to deal with this problem is to break down the target population into smaller groups based on vulnerable attributes. This would allow clusters to be targeted with or without overlap, depending on requirements. Although people could be categorized into several different targetable groups, there is a 'percentage fit' that can be used to differentiate among clusters. Through the process of clustering, segregations would be created, but not categories. Targetable categories are the true desired result, so further analysis must be done to determine the meaning of the clusters. This is essentially the same process that humans (social psychologists) go through in order to construct groups of categorized results. This task is labor intensive.

This thesis addresses the ability of information technology to conduct categorization of people based on their personal profiles. A key assumption is that profiles exist for each target in the desired population, but the task of categorizing all the personnel cannot be accomplished in a timely manner. Through the process of text mining and natural linguistic processing, computers can extract the attributes of a person from their profile and place them into a structured format of predefined categories for follow-on targeting.

The accuracy of the categorization depends on the accuracy of the profiles. For example, if a greedy, social person is described as having personal standards, the automated linguistic processor will categorize him into the personal standards group. Thus, if the description of the person does not accurately describe him or her, the targeting will be faulty. Thus, without accurate profiles, the system would be pointless. Thus, neither the manner in which the profile is attained nor the truthfulness of the profile will be questioned. However, to be representative and provide suitable data, the profiles must contain attributes and concepts about a person and cannot simply be a chronology of their lives.

Psychological profiling is a time consuming and interpretive task that requires human effort; this is the primary reason so few people can be precisely targeted. In order to precisely target more individuals, the ability of machines (more accurately: software) to categorize profiles based on limited personal data will be tested. The goal is not to replace human involvement but rather to increase the number of successfully influenced targets by using the human output in multiple ways. The high value targets will still require human analysis, but the remaining masses would be sorted into predefined categories based on the human categorization of the high value targets. In essence, the knowledge inherent in the output of the psychologists could be used unlimited times vice only once.

Since precision targeting will be more likely to effectively influence one's adversary, the conclusion can be drawn that a more tailored message to the remaining masses will also produce a greater percentage of people being successfully influenced. This is evidenced in the ability of companies to market various products to different segmentations of consumers. This research will determine the capability of the leading software solutions to accomplish this task.

## C. DESIRED APPROACH FOR PSYOP SEGMENTATION

Currently, the ability to determine how people can be influenced is a manual task that has demanded significant time from skilled personnel. The ability to wade through many profiles in a timely manner is a function of dedicated human resources and the quality of the analysis. This is why only a minimal amount of individuals can be precisely targeted based on their specific characteristics and vulnerabilities.

Although in a perfect world every target would be analyzed by experienced psychologists in order to determine the best PSYOP tactic to employ, it is not practical. A better approach would be to use the human effort of the psychologists to analyze the most valuable targets and categorize the remaining multitudes into several broad groups. These groups could then be targeted on their most vulnerable trait which would in turn yield a higher rate of success to the PSYOP campaign. The downside is that sifting through and sorting thousands of profiles would require significant time or manpower. Time is the critical driver since it would not be practical to postpone a PSYOP campaign. PSYOP messages must be delivered when they are most likely to be beneficial.

Even though such sorting seems like a simple process, the resources to perform that task would greatly decrease the number of high value targets that could be individually analyzed. There are two options: 1) hire more psychologists to enable the categorization of every profile in the allotted time, or 2) use the high value categorizations (performed by humans) as a template for IT to perform on a large scale. The cost of option one would greatly exceed the cost of the two. Thus, Information Technology (IT) may provide the means to increase the number of targets successfully influenced. If an automated process could be employed to categorize the remaining targets, this would eliminate the time and manpower constraints. Simply put, this equates to a more personalized message for each member of a larger.

The problem becomes how to implement an IT solution that can perform the same function that a psychologist provides via the manual process. To attack this problem from ground zero would entail sophisticated software engineering and robust testing. This development process would be both time-consuming and expensive; however, corporate America may already have the solution.

Businesses have been sifting through endless piles of data (records, messages, emails, etc…) for long periods of time in order to find new and lucrative ways to improve their efficiencies and profitability. With such a heavy reliance on existing Commercial Off-the-Shelf (COTS) technology, it is reasonable to evaluate what currently exists in

corporate America. The tools that perform these tasks for corporate America should, in theory, be able to automate the PSYOP targeting process to provide the desired increase in successful influence operations.

Because this type of research is dealing with written profiles of people that are considered unstructured data, the use of text mining and data mining are employed. This automates the process of analyzing text documents for influence campaigns. Specifically, the project employs the use of natural language processing to extract concepts and find meaning in a 'bag of words' or profile. In essence the profiles are transformed, or parsed, into pieces so that a machine can understand them and use them in further analysis. Data mining techniques for categorization and clustering are then applied to the extracted concepts to build predictive models based on patterns determined from known results.

The project focuses on the feasibility of commercial products to perform automated categorization of targets based on the existing profile data. Text and data mining are studied and evaluated to determine applicability, and the tools that provide the most promise are recommended for future optimization. Additionally, recommendations are made on how to improve the ability of IT systems to more effectively produce accurate results in target categorization.

This thesis begins by researching the methods of flat text categorization and specifically addresses the methods that will be most beneficial to the application of IT to influence targeting. Once a particular software package has been identified as suitable, a testing methodology will be developed to evaluate the advertised capabilities.

The final phase of the project will be to objectively evaluate whether or not the IT solution produces an accurate categorization of the target profiles. This will be done by processing new data with the proposed system and then comparing the results to the output produced by a psychologist.

Effectiveness/accuracy will be measured by how well the software tool can place targets into the same category a psychologist would. This will test the hypothesis that technology can accurately characterize a person with only a limited personal profile and place them into a predefined category based on their predominant attributes.

The benefit of this research is to know whether or not a commercial software solution can automate the task of sorting targets into predetermined categories so that the speed and quantity of processing will be greatly improved over current methods. The project is named *Automated Psychological Categorization via Linguistic Processing System,* or PsyCLPS.

## D. DESCRIPTIVE SCENARIO

Here is a hypothetical scenario of how PsyCLPS could serve the Intelligence Community.

### 1. Without PsyCLPS

A military member has been given the task of conducting PSYOP on a large group of people; the total number of people is approximately two thousand. Each person in the target group has a two-to-three page description about him or her (a profile). Psychologists are then asked to review the targets and describe the best PSYOP approach to deploy against each individual. The psychologists read several profiles before they realize they cannot read them all. Each profile requires at least ten minutes to read, interpret and describe in the terms required by the PSYOP community. This implies that over 330 hours would be required to accomplish the task. Even with a team of several psychologists, it would take weeks to process all the data so that it could be used in the theater of interest. As a default, the psychologists ask for a list of the most critical targets and the process is conducted on those few people (on the order of tens rather than thousands).

The individuals that were analyzed are then sent specific PSYOP messages geared to influence their behavior. The remaining masses are sent one bulk, generic message that has traditionally yielded unimpressive results. Worse yet, the bulk message actually did the opposite of its intended operation. Because it was trying to appeal to thousands of people, it consisted of ideas that actually infuriated one third of the group. This segment of the group felt angry after viewing the message that insulted their personal standards. Although the bulk message had a positive influence on many in the bulk list, it also hardened others and strengthened their resolve against the endeavors of the PSYOP effort.

### 2. With PsyCLPS

Imagine the same scenario, only this time, the PSYOP community is using an automated linguistic tool in an attempt to process more profiles into specific categories. Again the psychologists are given the profiles of high value targets from a repository of thousands of potential targets. After reading several documents, they begin to form three specific groups. The groups consist of 1) power, 2) personal standards, and 3) social. Targets placed in the power group are individuals that are Machiavellian in nature. They will do whatever it takes to gain power and control others. They are often ruthless and exhibit very little loyalty unless it serves their own ambitions. The second group is comprised of people who have strong internal beliefs and will risk being ostracized in order to defend them. This group tends to be loyal and often exhibits moral courage in the face of adversity. The third group is satisfied if they are surrounded by others who are satisfied. The social group cares more about pleasing others than about achieving power or defending their own personal beliefs. Although they may seek fame and fortune, they will not likely create conflicts to attain it.

After the social psychologists have categorized this more manageable amount of profiles, these profiles are then used to develop specific campaigns against these high value targets. Thus, the output in this scenario yields everything the previous scenario yielded. However, the classification that took place serves another purpose. The groups are used as "training documents" for a software suite to build a linguistic taxonomy. The concepts from each group of profiles are used to build a virtual category for each of the three groups. A technician then places the uncategorized profiles into the software application and allows a computer to sift and sort the profiles into the correct categories.

In a matter of minutes, thousands of profiles have been categorized based on the expertise of the social psychologists. Profiles that did not meet the criteria of any category (based on the software settings) are left out while the other profiles are now in a specific grouping that can be sent a PSYOP message that is likely to appeal more personally to each target.

PsyCLPS is a project that is attempting to make this discussion a reality. Increasing the number of successful influences for individuals is a primary means for achieving victory.

## E. ORGANIZATION OF THE STUDY

The remainder of this thesis is organized as follows. Chapter II discusses data mining and text mining along with the algorithms and models that enable the technology. Chapter III describes the software applications of relevance to this research. Chapter IV covers the limitations of the project, the testing methodology, and discusses the results of the experiments that were conducted. Chapter V draws conclusions from the results described in Chapter IV in order to prove or disprove the ability for commercial software applications to solve the research question posed in this thesis.

THIS PAGE INTENTIONALLY LEFT BLANK

# II. DATA MINING PROCESS

The ability to organize and manage information has always been important to organizations. With the advent of computer technologies over the past two decades, the need for such information management has become a mandatory ingredient to success. Computers have enabled the automatic collection of information, which has lead to enormous depositories of data so voluminous and multifaceted that manually searching through them would be futile. Thus, the need for automatic data classification was recognized and the advent of specialized programs designed to solve this critical problem began to emerge in the market place.

Using legacy methods of data analysis and extraction would be infeasible given the enormity of modern data sets. Data mining is a way of extracting valuable information from large volumes of data by using Information Technology to automate the process. The mined data can reveal meaningful patterns of interest within a data set or it can serve as a means to sort material into a manageable form. Figure 2 shows the data mining process:

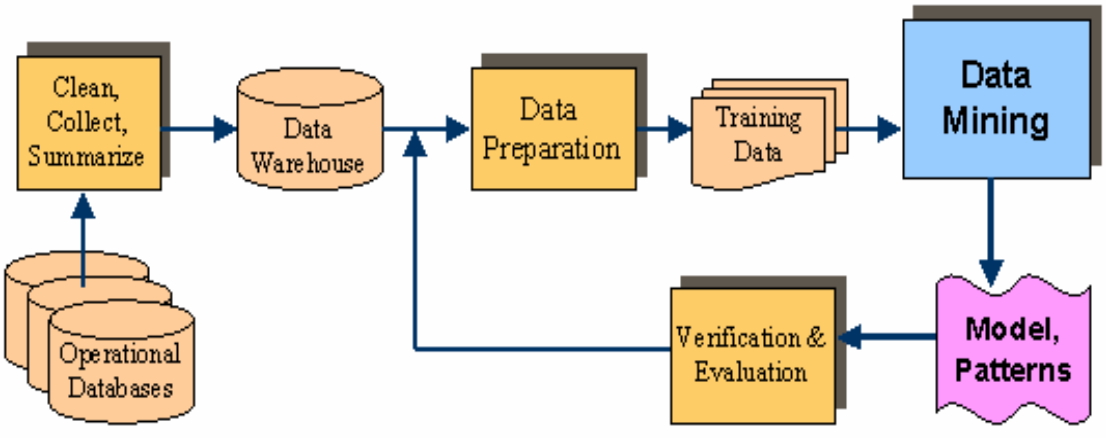


Figure 2. The Data Mining Process (From: http://www-users.cs.umn.edu/~mjoshi/hpdmtut/sld001.htm. October 2003)

The true purpose of the data mining process is to extract knowledge through discovery. This extraction consists of six steps; each step builds on its predecessor. These steps are:

- Selection and sampling
- Preprocessing and cleaning
- Transformation and reduction
- Mining
- Visualization
- Evaluation.

Selection and sampling can be further broken down into two steps, understanding the problem and obtaining the data set. To understand the problem there must be a clear comprehension of the purpose of the data-mining project or application. This is essential in the next part, obtaining the data set. This will often involve random sampling from large data sources to capture records to be used in the analysis.

Preprocessing and cleaning involves the verification of the data for completeness and provides a 'sanity check' of the data to ensure it is within bounds and does not possess obvious outliers. This can be accomplished by generating graphical depictions of the data and looking for 'holes' where no data are available and finding extreme outliers that do not fit due to inconsistent scales or units of measure.

For transformation and reduction, which may not always be needed, the data must first be separated for the purposes of training, validation, and testing. This depends on the data mining algorithm being used (the supervised mining tool). However, the data may need to be massaged so unnecessary variables are removed or variables that don't 'make sense' in their current state are transformed into units that can be measured and manipulated.

Mining involves the determination of the task and associating that task with a specific or general category of data mining techniques. If the task is categorization then the technique of neural networks may be selected for the actual mining process. In addition, mining algorithms need to be selected that will perform this task. This is normally an iterative process in which several different algorithms are attempted in an effort to extract the most data, which can be turned into knowledge/information.

Visualization is the process of interpreting the results of the algorithms for the best algorithm to employ. This goes hand-in-hand with the mining process and again is an iterative process. Defining the 'best' tool is hard to describe and difficult to determine. Depending on the sensitivity of the mining techniques infinite relationships may be discovered and may lead to confusion vice knowledge.

Finally, evaluation involves the employment of the data-mining tool within systems for the purpose of extracting valuable information that can assist in the production of decisions and actions.

## A. DATA MINING TASKS

Data mining breaks down into two basic forms: predictive and descriptive. Predictive mining attempts to forecast future variables while descriptive mining attempts to find patterns that will explain the current environment from which the data was collected. Within these two mining approaches, there are several techniques. (See Table 1.)

| Predictive | Descriptive |
|---|---|
| Classification | Association |
| Deviation Detection | Clustering |
| Regression | Time Series |

Table 1. Mining Techniques

- Classification/Categorization: Uses training documents to 'train' a model for future use. The training documents must be representative of the overall model.
- Clustering: Extracts data points from documents and matches analogous attributes so that the documents are 'clustered' into similar groups while at the same time separating clusters that are less similar to one another.
- Associations: Produce dependency rules which will predict occurrence of an item based on occurrences of other items. For example: knowing that beer is often bought along with diapers and milk is valuable to know when developing a shelf-stocking strategy.
- Time Series / Sequential: Given is a set of objects, each object having its own timeline of events, Time Series find rules that predict strong sequential dependencies among different events. Rules are formed by first

discovering patterns. Event occurrences in the patterns are governed by timing constraints.

- Regression: Predicts a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Deviation Detection (outlier analysis): Involves discovering the most significant changes in data from previously measured or normative values.

Each of these methodologies involves a myriad of complex algorithms and unique techniques that attempt to find meaning in large data sets. Given the problem that PsyCLPS is attempting to address, Classification/Categorization is the method of greatest interest. In particular PsyCLPS needs to classify flat text documents into predefined categories; this is a predictive task best performed by Categorization.

## B. MODELS AND ALGORITHMS INVOLVING CATEGORIZATION TECHNIQUES

Automatic text classification is currently in high demand because there is an urgent need to sift through the tremendous volume of text documents that are of potential value to an organization. How this classification is accomplished is both diverse and complex. Text classification tools have two fundamental approaches to organizing text documents. The first is extraction; after an appropriate taxonomy of the documents, categories are created that describe and summarize the meaning in the entire set of documents. The second approach is assignment; the documents are placed into a fixed number of predefined categories that are specified by the user (either up front or on the fly). This thesis is concerned with the second approach: assignment based on the description in Chapter I.

To accurately categorize a document, a text classification tool must perform the laborious tasks of performing a taxonomy of the document, extracting concepts, interpreting the collective meaning of the concepts, and then assigning it to the most appropriate category. The following methodologies are associated with Classification/Categorization:

- Decision Trees: Decision trees use tree structures to define true/false queries. Branches represent nodes and the leaves are the categories. When new documents are analyzed they flow down the tree answering each true/false query at each branch (node) until they reach the appropriate category.
- Decision Rules: Here each category has a defined rule set. As new documents are classified, they are run through the various rule sets to see which categories they fit into.
- k-Nearest Neighbor: This method compares documents via vector analysis to see how close each document is to another in the data set.
- Bayesian Approaches: Two approaches: naïve and non-naïve. In naïve, each word in a document is analyzed independently of other words around it; they are efficient, but sometimes ignore relevant data such as the meaning in the grouping of words. Non-naïve uses the same methods but also looks at word order; they are only slightly more accurate but take more time to process.
- Neural Networks: Neural networks use back propagation or counter propagation to learn the meaning of documents. They are expensive and complex.
- Regression Based Methods: Regression based methods use multivariate regression techniques to compare input and output matrices in order to find a best fit for a given data set.
- Vector Based Methods: Vector based methods use training documents (which are usually provided by the user) to define vectors for comparison between text files.

Many software packages combine multiple methods in an attempt to optimize the product output. The ways in which the algorithms are used is proprietary information that is hard to obtain. However, regardless of the approach (or the proprietary name given) there are eight basic methods (some of which are subsets of those described above) that provide the ability to classify flat text documents (documents without a fixed format or structure) into know categories.

Here these methods are discussed in detail to include their advantages and disadvantages. The following section provides varied levels of detail due to the level of complexity and maturity of the methodology.

### 1. Neural Networks

Neural Networks are analytic techniques modeled after (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain. Artificial

Neural Networks (how we refer to machines) use a similar method to learn patterns and relationships in data and are a substantial departure from traditional approaches of machine learning. They resemble the human brain in the manner that they acquire knowledge through learning and that the "knowledge" is stored within inter-neuron connection strengths known as synaptic weights.

Normally, machines are programmed with every facet of a problem so it can logically solve potential issues. Neural networks do not require explicit coding of every instance of a problem, they only require raw data related to a problem so it can sort through the information and produce an understanding of the factors impacting the problem.[6] This is done by the creation of neural network learning rules. These rules are the algorithms used to "learn" the relationships in the data. The rules enable the network to "gain knowledge" from available data and apply that knowledge in the form of meaningful output.

Neural networks are powerful tools for data mining. They can extract trends in large amounts of what is seemingly unrelated data. They are designed to work well where you wish to develop functional, classification, or time series models, or places where nonlinear relationships exist. Some typical applications of neural networks include: process modeling and control, machine diagnostics, portfolio management, target recognition, medical diagnosis, credit rating, targeting marketing, voice recognition, financial forecasting, quality control, intelligent searching, and fraud detection.[7]

Neural network techniques can also be used as a component of analysis designed to build explanatory models. Neural networks can help explore data sets in search of relevant variables or groups of variables.

Neural networks have many advantages as well as disadvantages. Advantages of neural networks are:

---

[6] Anderson, Dave and McNeil, George. "Artificial Neural Networks Technology: A DACS State-of-the-Art Report." Kamon Sciences Corporation, Utica, New York. August 1992. http://www.gaianxaos.com/PdfChaosLibrary_files/ArtificialNeuralNetworksTechnology.pdf. August 2004.

[7] Neural Networks Toolbox. www.mathworks.com/access/helpdesk/help/toolbox/nnet/nnet.shtml. August 2004.

- Capable of approximating any continuous function.
- Nonlinear, can interpret nonlinear relationships and signals.
- Can accommodate supervised learning.
- Adaptability can assign weights to factors and retrain quickly.
- Provides confidence levels, good for classification.
- Tolerant to problems and will gracefully degrade if damaged.
- Can be implemented widely.
- Disadvantages of neural networks include:
- Final solution depends on initial conditions.
- Virtually impossible to "interpret" the solution in traditional terms; difficult to build theories that explain certain things.
- Small networks take a long time to train.
- Larger networks may not generalize correctly.
- It is not easy to establish internal connection structure.

Neural networks use different types of learning strategies. The first method is the supervised learning strategy (Figure 3)[8] in which the learning algorithms generate an error function that requires to be minimized. The second method, unsupervised learning strategy (see Figure 4) is where the learning algorithm is based on the history of the system.

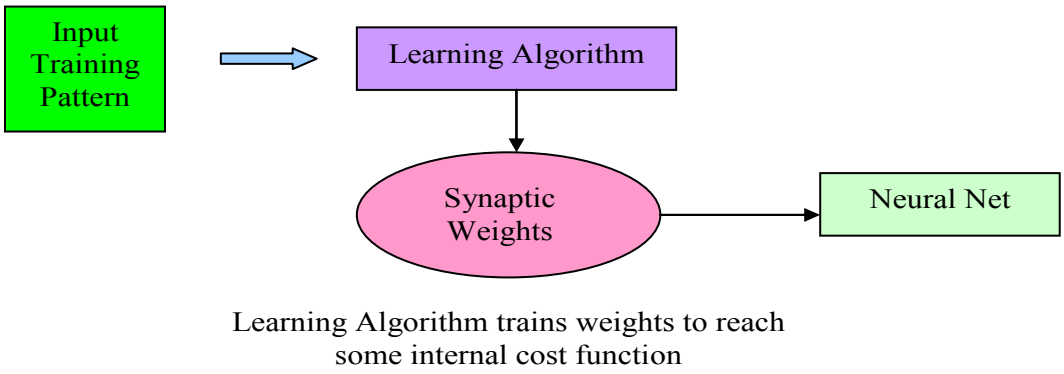


Figure 3. Supervised Learning Strategy

[8] Ibid.

Neural networks use many different kinds of learning functions.[9] The supervised learning algorithms that neural networks use to "train" themselves using a minimization of teaching error method are: back propagation, least mean square (LMS), potential, and correlation.

The back propagation algorithm is the most famous learning algorithm in neural networks. It works by computing the error (difference between the output and the teaching input) and then using that error with the output of the original source to determine the necessary changes. These changes are fed back into the system for computation. The other algorithms also work well in computing a derived weight for each node via different methods. Each method can be used in different applications, but the actual difference in results is relatively small.

The algorithms that are associated with Neural Networks are[10]:

- Back Propagation
- Lease Mean Square
- Potential
- Correlation
- LVQ (Learning Vector Quantization)
- RBF (Radial Basis Function)

By employing the algorithms, the system 'trains' itself to find relationships. How each algorithm operates is similar in mission, but the math in the algorithm is different. The following figure shows a basic flow of data through a Neural Network:

---

[9] Johansen, M. M. *Topics of Evolutionary Computation 2002 – Collection of Student Reports*. "Evolving Neural Networks for Classification." Department of Computer Science, University of Aarhus, Denmark. Fall 2002. http://www.evalife.dk/bbase/show_bibitem.php?bib_id=22588&idx=24. August 2004.

[10] Anderson, Dave and George McNeil. "Artificial Neural Networks Technology: A DACS State-of-the-Art Report." Kamon Sciences Corporation, Utica New, York. August 1992. http://www.gaianxaos.com/PdfChaosLibrary_files/ArtificialNeuralNetworksTechnology.pdf August 2004.

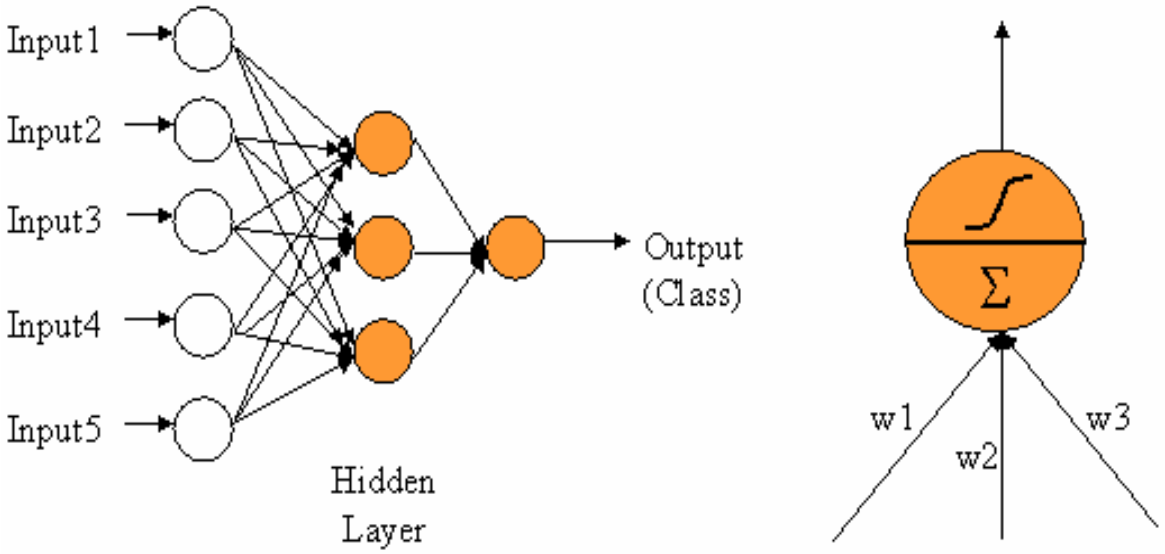


Figure 4. Unsupervised Learning Strategy (From: http://www-users.cs.umn.edu/~mjoshi/hpdmtut/sld001.htm. October 2003)

## 2. Decision Trees

A decision tree, or more properly referred to as a classification tree, is used to learn a classification function which predicts the value of a dependent attribute (variable) given the values of the independent (input) attributes (variables). Each branch of a decision tree represents a choice between a number of alternatives, and each leaf node represents a classification or decision.

Decision trees are constructed in two separate phases: tree-building phase (also referred to as tree-induction phase) and tree pruning phase or more specifically, a method of making the decision tree smaller. In the tree building phase the data that you want to classify is repeatedly partitioned until all the data in that partition belongs to one class or the partition is small enough to manipulate. In the second phase you try to either remove unrelated data or change the criteria for node splitting.

In tree-induction each node of a tree implements a decision rule that splits the examples into two or more partitions. New nodes are created to handle each of the partitions and a node is considered to be terminal or it is considered to be a leaf node based on predetermined stopping criteria. The stopping criteria must be identified prior to beginning the tree induction process.

In the second phase (commonly referred to as the pruning phase) the effort is to refine the results to the desired objective, basically to make the tree smaller to meet its intended goal. Three different methods can be used in this phase. The first method is to change the criteria of how data is split at each node. This is done through different statistical tests (chi-square, G statistic, and the GINI index of diversity)[11]. It is inconclusive which test works best in all cases, but they are better than random splitting.[12] The second method of pruning is exactly what the name implies; it is the removal of branches, during or after construction, that have little statistical significance to the category based on identified criteria. The final method is to produce smaller decision trees. This can be accomplished by a method called, look-ahead, which attempts to establish a decision at a node by analyzing the classifiability of the potential split data.

Advantages of decision trees:[13]

- Relatively fast compared to other classification models.
- Obtain similar and sometimes better accuracy compared to other models.
- Simple and easy to understand.
- Can be converted into simple and easy to understand classification rules.
- Can be used for both categorical and numerical data.
- Inexpensive to construct.
- Easy to integrate with database systems.

Decision trees do face specific challenges. One is the inability to develop algorithms that produce decision trees of small size and depth. This is critical for simplicity and ease of understanding the process that was generated. A second challenge is to maintain good performance of the generalization. It has been demonstrated that larger decision trees lead to poor generalization performance. Decision trees are not advantageous in all circumstances.[14]. Difficulties with decision trees include: the

---

[11] Kothari, Ravi and Ming Dong. "Decision Trees for Classification: A Review and Some New Results." in *Lecture Notes in Pattern Recognition*, S. R. Pal and N. R. Pal, (Eds.), Singapore, 2001, World Scientific Publishing Company. http://www.cs.wayne.edu/~mdong/papers/paper_review.pdf. August 2004.

[12] Ibid.

[13] Kamber, Micheline, Winstone, Lara, Gong, Wan, Cheng, Shan, Han, Jiawei. "Generalization and Decision Tree Induction Efficient Classification in Data Mining." http://www-faculty.cs.uiuc.edu/~hanj/pdf/ride97.pdf. August 2004.

[14] Ibid.

construction of rules, they are based on similar data, they may lead to very different results, and some models are not flexible at modeling data that has complex distributions.

Several different algorithms may be used in decision trees, they include[15]:

- Hunt's Algorithm
- CHAID
- CART
- ID3, C4.5 -5.0
- SLIQ, SPRINT
- FACT
- QUEST
- LMDT
- T1

**3. Rule-Based (Decision Rules/Rule Induction) Approaches**

Rule-based approaches to classification are designed for complex document sets and classification structures. They are beneficial in many ways but also possess drawbacks. These approaches are commonly used in concert with other methodologies to maximize their potential.

Rule-based classification systems learn and apply rules that it determines through the exploration of the data. The manner in which it learns and applies rules is relatively simple. Every set of data is broken down into record sets. Each record set follows the scheme of $(A_1, A_2,\ldots,A_k)$, where $A_1, A_2,\ldots,A_k$ are attributes.[16] If $A_k$ are not categorical attributes then all continuous attributes are further broken down into categorical attributes, until no continuous attributes remain.

All attributes, A, are given a value, ν, and combined into a value pair, $p$. A record set (also called a tuple), $t$, takes on the value, $v$ if and only if $t_i = \nu$, where $t_i$ is the value of the $i^{th}$ attribute of $t$. Rules, $r$, are generated when value pairs, $p$, are all similar and are

---

[15] Lim Tjen-Sien, Wei-Yin Loh, Yu-Shan Shih, "An Empirical Comparison of Decision Trees and Other Classification Methods." University of Wisconsin, Madison. January 1998.

[16] Yin, Xiaoxin and Han, Jiawei, "CPAR: Classification based on Predictive Association Rules." University of Illinois, Urbana-Champaign. http://www.ncassr.org/projects/sift/papers/xiaoxinCPAR_siam2003.pdf. August 2004.

associated with a class label *c*. A tuple *t* satisfies rule *r*'s body if and only if it satisfies every literal in the rule. If *t* satisfies *r*'s body, *r* predicts that *t* is of class *c*. If a rule contains zero literal, its body is satisfied by any tuple.[17]

The weighting that individual words carry in deciding how a document should be classified is controlled by their frequency, position, and context. The higher the frequency of a particular word in a document, the more likely that document is to be about that word. Also, the higher (sooner) the word appears in the document, the more likely that the document is to be about that word. Another consideration is that the relative position of words to other words affects their meaning.

There are two methods of generating association rules. One method is to generate association rules by beginning with certain support and confidence thresholds as candidate or initial rules. The system that is applying this rule based method then selects a small set of rules that have been pre-determined to form a classifier. When the system predicts a class label for the category it uses the best rule, the one with the highest confidence, whose body is satisfied by the example chosen for prediction. A second method generates and evaluates rules in a similar way but uses a more efficient frequency pattern tree[18],[19] structure. It uses multiple rules in prediction, using the chi-squared statistic for weighting. It has been demonstrated that the second method outperforms the first in accuracy.

For both of these methods, rule generation and selection is difficult when the datasets contain a large number of attributes. Thus, it may take a significant amount of time to generate and select rules.

Rule-based classifiers have several benefits. It supports greater manual intervention and is more transparent to the user. The user has a number of options for building the classifier:

---

[17] Ibid.

[18] Frequency Pattern Tree structure: consists of one root labeled as "null"; a set item prefix sub trees as the children of the root, and a frequent-item header table. Each node item prefix sub tree consists of an item-name, count, and node-link.

[19] Han, Jiawei, Pei, Jian and Yin, Yiwen. "Mining Frequent Patterns without Candidate Generation." School of Computing Science, Simon Fraser University, Vurnaby, British Columbia, Canada. http://citeseer.ist.psu.edu/cache/papers/cs/14568/ftp:zSzzSzftp.fas.sfu.cazSzpubzSzcszSzhanzSzpdfzSzsigmod00.pdf/han99mining.pdf. August 2004.

- Structure can automatically generate the rules base from the document training set
- The user can accept the rules or manually refine them to improve performance
- If no training set exists the user can author a new rule base

Additionally, rule-based classifiers are highly discriminatory. They can deliver greater precision and are suited to applications where it is mandatory that categories contain only highly relevant information that fits correctly in a given category. Conversely, the rigidity of the rules could reject closely related information (as opposed to precisely matching) from being included in a given category.

The difference between Decision Trees and Decision Rules is that the rule sets for decision rules are independent and are unlikely to form a tree like structure. Decision rules are only as good as the rules that it generates, there may be instances where the rules that are generated conflict and even produce results that have many outcomes, and in some cases there may be observations that are not covered by the rule sets. In these cases confidence levels are developed and used.

Algorithms associated with Rule Based Methods are[20]:

- Separate and Conquer Rule Learning (Family of Algorithms)
- Precision
- SQUEEZE
- IT Rule
- Incremental
- Fuzzy Logic

**4. K-Nearest Neighbor (Memory Based Reasoning)**

K-Nearest Neighbor is a computationally, burdensome method for computers. It computes predictor variables for specific cases and then compares the inputs of each case on the values of the preceding cases. The algorithm does not simply place a case (or assign a case) to a category based on its proximity to only one other case, but rather on all the preceding cases. If a new case is computed to be closer to a series of similar points then it is placed accordingly, otherwise it is places with a different series that more

[20] Higgins Jr., Charles M. "Classification and Approximations with Rule-Based Networks." Pasadena, California. 1993. http://neuromorph.ece.arizona.edu/~higgins/pubs/oldpubs/thesis.pdf. August 2004.

closely resembles its characteristics. K-Nearest Neighbor is sometimes referred to as memory-based reasoning. This is because in order to speed up K-Nearest Neighbor, the data is kept in memory for quicker access. The algorithm employed by k-Nearest Neighbor is called *Nearest Neighbor*; it is a statistical method of computing the closest related data point.

**5. Bayesian Networks**

These models work heavily on the calculation of probabilities. These types of models require 'complete' data to work properly, but due to the nature of statistics being based on random samples this problem can be overcome by several iterations. Bayesian models predict classifications based on the random sample of attributes. Whichever probability that is maximized by fitting a random category to the random samples of the attributes is the associated category. Depending on the type of Bayesian model selected determines the importance of word order and potential context. The underlying algorithms at work in Bayesian Networks all derive from the Bayes Theorem. They include[21]:

- Polytree
- Pearl's
- Local conditioning
- Associated Tree
- DFS (Depth First-Search)

Bayesian networks have been used for years to produce visualizations of predicted relationships learned from known data and they are becoming more popular as a method of reasoning using probabilities. They have been applied in such instances as medical diagnosis and language understanding. In general, Bayesian networks are directed acyclic graphs where the nodes are independent random variables. The conditional probabilities of particular nodes can be calculated given that other nodes (parents) have known values. Simply explained, Bayesian networks have a graph component and a probability component.

[21] Diez F. J. "Local Conditioning in Bayesian Networks." Technical Report R-181, Cognitive Systems Lab., Dept. of Computer Science, UCLA, July 1992. http://citeseer.ist.psu.edu/context/149372/130346. August 2004.

Dependency networks are similar to Bayesian networks in that they are also graphical models that can be used to encode, learn, and reason with respect to probabilistic relationships. One example is DnetViewer, developed by Microsoft Research, which employs dependency networks for data visualization. Dependency networks can be thought of as a collection of regressions or classification among variables in a domain that can be combined using the machinery of Gibbs sampling to define a joint distribution for that domain.[22] Essentially, dependency networks do not need to be acyclic. The graph component for a dependency network is a cyclic directed graph such that a node's parents render that node independent of all other nodes in the network.[23] Each node has its own conditional probability and joint probabilities are drawn from Gibbs sampling.

The advantage associated with dependency networks is that they provide more accurate learning than Bayesian networks. However, the data sets must be complete. Without complete data, inconsistencies are possible. The larger the data set, the less likely the inconsistencies will be. Therefore, Bayesian networks are better for encoding casual relationships and using knowledge based approaches. Another advantage to Bayesian networks is that they have relatively simple network structure and tend to run faster because their algorithms for exact inference are quicker than the Gibbs sampling technique used in dependency networks. The algorithms for dependency networks essentially consist of independently performing a probabilistic classification or regression for each variable in the domain.[24] Another advantage to dependency networks is in their ability to predict relationships and preferences; these predictions are essential for probabilistic queries. Both SQL Server 2000 and Commerce Server 2000 are examples of products that include dependency networks.

The question becomes which method to use since both provide unique advantages as well as disadvantages. The answer is to employ both methods sequentially. New software packages are being developed that will use algorithms capable of using both

[22] Heckerman D., D. M. Chickering, C. Meek, R. Rounthwaite and C. Kadie. *Dependency Networks for Inference, Collaborative Filtering, and Data Visualization*. Journal of Machine Learning Research, 1:49-75, October 2000.

[23] Ibid.

[24] Ibid.

Bayesian and dependency networks together in order to classify data. By learning Bayesian networks from dependency networks, the advantages of both representations can be achieved without significant disadvantage. Thus, learning a dependency network through a scalable algorithm and then using it as an "oracle" for the statistics needed for the Bayesian network is possibly the most beneficial approach.

However, although the combination of these two techniques looks to be beneficial, currently there are no robust software applications that employ this combination of technologies to obtain an end-to-end solution for the classification of data. At present Bayesian network models are predominant on the market.

### 6. Support Vector Machines

Support Vector Machines are a combination of a regression based methodology and 'training sets' to optimize the efficiency and accuracy of categorical placement of records. Record attributes are dimensioned to help compute functions. The functions combined with 'training sets' are optimized to produce predictions for follow on records. Support Vector Machines are a family of algorithms that are based on Quadratic Programming. An algorithm that is working in the background in this system is the SMO (Sequential Minimal Optimization) algorithm. SMO quickly trains the classifiers and avoids numerical difficulties associated with other optimization methods. This method is valuable to text classification because it is highly dimensional and requires flexible tools for generalizations.

Support Vector Machines are based on the Structural Risk Minimization principle from computational learning theory; the theory is essentially a method that guarantees the lowest true error.[25] With Support Vector Machines (SVMs), the task of classification involves both training and testing data that are made up of various data instances. These instances have within them a target value and multiple attributes. The goal of SVM is to produce a model that predicts target value of data instances in the testing set when given only the attributes.[26]

---

[25] Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[26] Cortes, C. and V. Vapnik. Support -Vector Networks. Machine Learning, 20:273-297. http://citeseer.ist.psu.edu/cache/papers/cs/23317/http:zSzzSzwww.research.att.comzSz~corinnazSzpaperszSzsupport.vector.pdf/cortes95supportvector.pdf. August 2004.

Support Vector Machines work by mapping data into high dimensional attribute space and then using those attributes (sometimes referred to as features) to compute linear functions. A criticism of SVM's is that this process can be time consuming compared to other categorization methods. However, SVM's can run faster if they are combined with learning theory and efficient training algorithms. This process produces data that is then easily accessible to optimization and analysis. Thus, standard quadratic programming tools can be used to solve the fundamental problem of optimization that is often associated with the SVM classification technique. One way in which SVMs can speed up the optimization process involved in classification is by using the sequential minimal optimization (SMO) algorithm. The SMO algorithm can quickly train the classifiers as well as avoid numerical difficulties associated with other optimization methods. There are two primary methods or algorithms[27],[28] that are employed in SVM, Chunking and Sequential Minimal Optimization (SMO). SMO was developed to overcome the shortfalls of Chunking.

Another feature of SVMs is that they use kernels. Kernels provide both advantages and disadvantages. They provide an inference between the algorithms and the data, which is beneficial, but the process of kernel selection can prove to be difficult since many kernels may need to be tested to ensure optimal results. Examples of various kernels include: linear, Gaussian, polynomial, and sigmoid. Since these decisions are generally made at run time, and since SVMs are extremely complex, SVMs may not always be the preferred method of classification. Another drawback to SVMs is that the number of support vectors in SVMs is potentially very large which could make the technique impractical for small matrix routines.

However, SVMs are particularly valuable for text classification. Text classification is a high-dimensional task and therefore requires a method such as SVMs because the ability of SVMs to generalize does not depend on the number of attributes obtained during the mapping process. Thus high dimensional input space can be handled by SVMs by using over fitting protection (preventing the combining, or overlaying, of

[27] Balcazar, Jose, Yang Dai, Junichi Tanaka and Osamu Watanabe. "Provably Fast Training Algorithms for Support Vector Machines."

[28] Platt. J. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization.* In Advances in Kernel Methods -- Support Vector Learning, pp. 42-65. MIT Press, January 1999.

attributes) and specifying which attributes are irrelevant. This allows SVMs to accept the remaining attributes as relevant which is advantageous in text classification. When comparing SVMs with other existing methods of text categorization, SVMs have been shown to achieve better results. In fact, SVMs were shown to substantially and significantly outperform four of the major industry accepted methods which are: density estimation using a naïve Bayes classifier, the Rocchio algorithm, an instance based k-nearest neighbor classifier, and the C4.5 decision tree/rule learner.[29] However, the complexity of SVM's and the slow speed associated with them is still a formidable barrier.

**7. Rough Sets**

Rough sets are essentially a method of matching based on similar information or meaning. Attributes are extracted and considered rough or elementary until joined with another rough set that has similar meaning. At which point the sets are no longer 'rough' and relationships are formed based on the similar nature of their meaning. If the rough sets can not be paired, they are replaced with more precise attributes until matches occur. Rough sets are advantageous from the standpoint of limited preliminary information requirements and systems that employ them can run on parallel computers for quicker processing. The algorithms employed in rough sets are:[30]

- Indiscernibility relation
- The lower and upper approximation
- The B-positive region
- Semi-Minimal reduct
- Random reduct
- Elimination of "irrelevant" attributes from the superreduct

Zdzislaw Pawlak's introduction of *Rough Sets* in the early 1980's was the first look at this new mathematical tool that deals specifically with vagueness and uncertainty. Rough sets are based on a philosophy that with every object of the universe we associate

[29] Joachims, T. *Text Categorization With Support Vector Training.* In Proceedings of the 1997 NIPS Workshop on Support Vector Machines, 1998.

[30] Nguyen Sinh Hoa and Nguyen Hung Son. "Some Efficient Algorithms for Rough Set Methods." To appear in Proc. of the IPMU-96, Granada, Espana. 6.

some information (data, knowledge).[31] Rough set theory has become fundamental to the study of Artificial Intelligence (AI) and other areas of cognitive science. In particular, rough sets are useful in areas such as machine learning, knowledge discovery, and pattern recognition.

Rough sets are essentially elementary sets of indiscernible attributes. The attributes are considered similar if they are characterized as having the same information (or meaning). An elementary set, comprised of indiscernible attributes, forms a basic piece of knowledge about the source from which it was drawn. If elementary sets are joined, the joined set becomes a crisp set; otherwise it remains rough, imprecise and vague. In rough set theory, any vague concept is replaced by a pair of precise concepts; that pair is comprised of the lower and upper approximation of the vague concept.[32] Indiscernible relations can be classified as redundant if the elementary sets are identical and thus can be considered dispensable.

The concept of rough sets employs and complements other mathematical algorithms and tools that try to make sense of vague and uncertain data. Rough sets have linkage to Boolean methods, decision analysis, and discriminate analysis while complementing fuzzy set theory. Using rough sets for classification will likely involve a combination of other methods of classification. Therefore, rough sets are essentially a combination of classification methodologies that are brought under an umbrella of rough set theory.

The primary advantage to using rough sets is that they do not need any preliminary or additional information about data such as probability distribution in statistics, basic probability assignments, or the value of possibility in fuzzy set theory.[33] Another advantage is that programs that implement rough set methodology are able to run on parallel computers. However, some disadvantages include theoretical problems that have not been solved such as the classification of rough logic. There is also a tedious

---

[31] Pawlak Z. Rough Sets*: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishers, Dordrecht, 1991.

[32] Polkowski, A. S. L. *Rough Sets in Knowledge Discovery*. Physica-Verlag, 1998.

[33] Grzymala-Busse, J. W. *Knowledge Acquisition Under uncertainty – A Rough Set Approach*. Journal of Intelligent and Robot Systems. 1:3-16, 1988.

process involved with rough set data collection and the creation of new software products. This has hampered the development of rough set applications and is the reason that less than ten products are currently available on the market.

Although rough set classification methods are not predominant in today's software packages, they present a promising tool for the future. From banking to medicine, rough sets will eventually support flexible and personalized information management. In fact, rough sets may potentially have their largest impact on text classification because they will, theoretically, be efficient and effective at extracting and assigning categories with regards to text documents.

### 8. Genetic Classification

Genetic classification is a method of classifying data by using genetic programming and algorithms. It is a systematic method that enables computers to automatically solve a problem by creating new programs.[34] Generating new programs (also know as automatic programming, program synthesis, or program induction) is accomplished by "breeding" the new programs through Darwinian natural selection and biological operations such as reproduction, crossover (sexual recombination), mutation, and architecture-altering operations patterned after gene duplication and gene deletion in nature.[35] These newly transformed program populations are often then optimized using traditional statistical tools.

Genetic programming is unique in its approach to Artificial Intelligence (AI) and is different from Machine Learning (ML). Genetic programming requires a user to input the parameters of the problem. Random programs (breed from a population of programs already in existence) are then created and screened by evolutionary filters (survival of the fittest, mutation, crossover, etc…) to create a new program.

Genetic programming is best suited for problems that do not have an ideal solution. For example: flying a plane. There is no ideal solution for flying an aircraft. However, genetic programming could find a best-fit solution for the inputted variables. The primary advantage to genetic programming is its flexibility with constantly changing

---

[34] *The Page of Genetic Programming Inc*. http://www.genetic-programming.com. August 2004.

[35] Ibid.

variables. However, genetic programming is a new technology. Although it will inevitably become more applicable to classification problems, it is currently not mature enough to deploy in major networks. The complexity of this type of classification is also very intense. It requires not only a proficient background in computing skills but also in biological processes.

The following chart from genetic-programming.com shows the basic process associated with genetic programming.

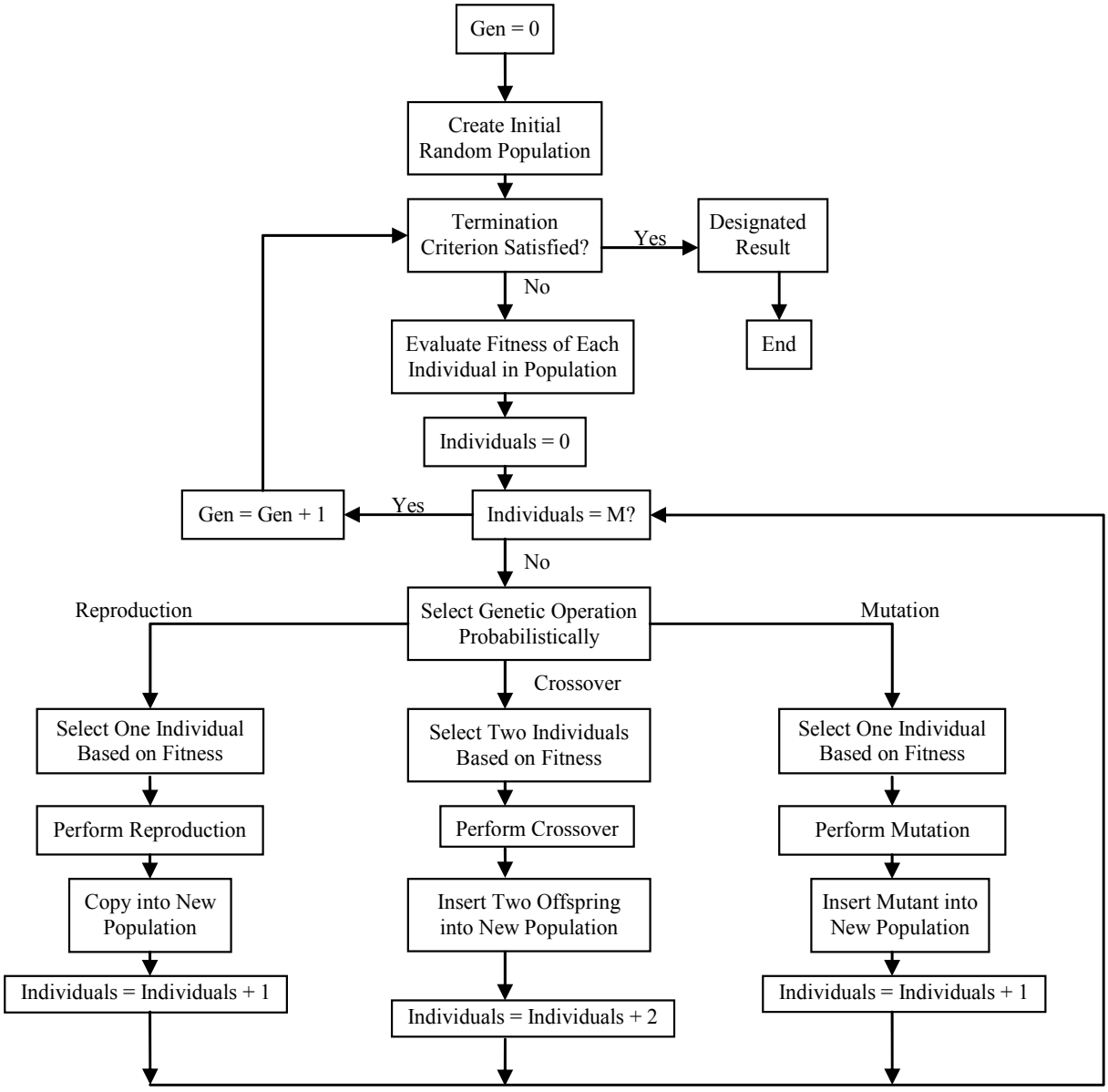


Figure 5. Flowchart of Genetic Programming (From: www.geneticprogramming.com. October 2003.)

## C. DATA MINING CONCLUSION

For each of the classification methods discussed, there are several algorithms that can be applied to this research. Although the algorithms were listed with respect to a particular method, they are not limited to such. The exact nature of how and when various algorithms are applied is convoluted since no one method can claim sole use of any one algorithm. Thus, determining which algorithms should be used or which methods should be employed is dependent on the type of problem being solved. Therefore, no one method is superior on all data sets and several methods must be attempted to find optimal results. The only reasonable conclusion that can be drawn is that applications that use several methodologies collectively will generally outperform individual methods.

Research has therefore shown that tools that incorporate multiple methodologies, using multiple algorithms, are best suited for this project. However such tools are proprietary in nature and therefore do not reveal the exact combination of methods and algorithms. Therefore it is difficult to specify which combination is most effective.

A comparison of the listed methodologies was conducted in several studies. In each case the comparison was extremely demanding. It has been noted that some of the results were heavily dependent on the test data set. In the one particular survey the test data was generated from pre-categorized documents. The reasoning was to determine the quality of the results from a pre-determined base line.

The results determined by several different experiments are summed up in the Table 2.[36] This table represents the comparison of the different categorization methods based on different types of testing data and the results that each produced. From reading though the results it can be stated that Support Vector Machines is a superior method over decision trees, K-nearest neighbor, Bayesian, and the Centoid method. But one draw back is that when comparing multiple methods no one method demonstrated a distinct advantage over another when categorizing large amounts of documents.

---

[36] Brucher, Heide, Knolmayer, Gerhard, Mittermayer and Marc-Andre. "Document Classification Methods for Organizing Explicit Knowledge." http://www.alba.edu.gr/OKLC2002/Proceedings/pdf_files/ID237.pdf. August 2004.

Aside from the conclusions in the table, what is of value is that no one method was the best in all cases. It was stated in a couple of experiments that a combination of methods produced better results. The final conclusion that can be derived is that multiple categorization methods' working in concert is the best methodology to undertake when producing standards for complex categorization problems. Some of the predominant classification products that use multiple approaches to classify data are listed in Table 3.

In conclusion, classification of data is not a trivial task whether being undertaken by humans or machines. There are methodologies that are utilized in order to take raw data and extrapolate it into a specific category. The reasoning of why certain items fit in certain categories and how well they fit is always a consideration. Computers assist humans with multiple tasks, but in order to help they must be "coded" with logic. Through this 'logic', machines must know what tasks are to be done and how to do them. Classification is very difficult and requires several layers of logic in order to make data points fit together. By this reasoning, the method that a machine uses to classify is important.

There are a plethora of software packages that can be used for categorization of multiple types of data. We have described the predominant methods of categorizations and the advantages and disadvantages of each. Each technique can be used in many different applications, but some methods work better than others due to the initial requirements. Overall, there is not one specific method that is best for all kinds of data. No matter what type of data is being classified, a combination of methods working together is better than any one specific method working alone. For our research we will be looking only at software packages that use multiple approaches for categorization. This will provide us with the most robust capability available in today's market.

| Author | Classification Method | | | | | | | | Test Data | | | Results |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Decision Tree | Rules Based | K-nearest neighbor | Bayesian | Neural Networks | Regression based | Centoid | Support Vector Machine | Newswires | Medical | Other | |
| 37 | X | X | | X | | | | | X | | | Rules based performed best, Bayesian & Decision trees performed similar, but worse |
| 38 | X | | | X | | | | | X | | X | Decision trees and Bayesian performed similar |
| 39 | | | X | X | | | X | | | | X | Independently each algorithm performed the same, but when combined they performed better |
| 40 | X | | | X | | | X | X | X | | | Support Vector Machines was the best methodology |
| 41 | X | | X | X | | | X | X | X | | | Support Vector Machines was the best methodology |
| 42 | | | X | | | | X | | X | X | | Combinations of both methods worked better then individual methods |
| 43 | | | | | X | | X | | | X | | Neural networks perform better then centoid |
| 44 | | | X | X | X | X | | X | X | X | | With fewer than 10 documents per category, Support Vector Machines, K-nearest neighbor, and Regression based methods |

[37] Apte, C., Damerau, F. and Weiss, S. M. (1994): "Towards Language Independent Automated Learning of Text Categorization Models." http://researchweb.watson.ibm.com/dar/papers/pdf/sigir94_with_cover.pdf. August 2004.

[38] Lewis, D. D. and Ringuette, M. (1994). "A Comparison of Two Learning Algorithms for Text Categorization." http://citeseer.ist.psu.edu/cache/papers/cs/508/http:zSzzSzwww.cs.cmu.eduzSzafszSzcs.cmu.eduzSzuserzSzmnrzSzwwwzSzpaperszSzcateg.pdf/lewis94comparison.pdf. August 2004.

[39] Larkey, L. S. and Croft, W. B. (1996). "Combining Classifiers in Text Categorization." http://citeseer.ist.psu.edu/cache/papers/cs/97/http:zSzzSzciir.cs.umass.eduzSzinfozSzpsfileszSzirpubszSzcombo.pdf/larkey96combining.pdf. August 2004.

[40] Dumais, S. Platt, J., Heckermann, D. and Sahami, M. (1998). "Inductive Learning Algorithms and Representations for Text." http://robotics.stanford.edu/users/sahami/papers-dir/cikm98.pdf. August 2004.

[41] Joachims, T. (1998). "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." http://citeseer.ist.psu.edu/cache/papers/cs/26885/http:zSzzSzranger.uta.eduSz~alpzSzixzSzreadingszSzSVMsforTextCategorization.pdf/joachims97text.pdf. August 2004.

[42] Lam, W., Ho, C. Y. (1998). "Using Bayesian Network Induction Approach for Text Categorization." http://www-ai.ijs.si/DunjaMladenic/papers/PWW/pwwAAAI98.ps. August 2004.

[43] Ruiz, M. E. and Srinivasan, P. (1998). "Automatic Text Categorization Using Neural Network." http://informatics.buffalo.edu/faculty/ruiz/publications/sigcr97/sigcrfinal2.html. August 2004.

| Author | Classification Method | | | | | | | | Test Data | | | Results |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Decision Tree | Rules Based | K-nearest neighbor | Bayesian | Neural Networks | Regression based | Centoid | Support Vector Machine | Newswires | Medical | Other | |
| | | | | | | | | | | | | perform significantly better than other methods, but with more than 300 documents per category all the methods performed about the same |
| 45 | | | X | | | | | X | | | X | Support Vector Machines perform better than K-nearest neighbor |

Table 2. Comparison of Categorization Methods

| Product | Company | Classification Methods Used |
|---|---|---|
| Affinium Model Suite | Unica | • Linear and Logical regression<br>• CHAID (Chi-Squared Automated Interaction Detection)/Statistical method<br>• Neural networks<br>• Genetic algorithms |
| ClassifyIt | Cim Vision Solutions | • K-nearest neighbor<br>• Neural networks<br>• SOM (Self-organizing map) |
| Clementine | SPSS | • Undisclosed multiple approaches |
| Enterprise Miner | SAS | • Undisclosed multiple approaches |
| KINOsuite PR | Toshiba | • Rules based<br>• Neural networks |
| Lexiquest Categorization System | SPSS | • Undisclosed multiple approaches |
| MarketMiner | MarketMiner Inc. | • Statistical Networks<br>• Logistic and linear regression<br>• K-nearest neighbors<br>• Decision trees |

44 Yang, Y. and Liu, X. (1999). "A Re-Examination of Text Categorization Methods." http://citeseer.ist.psu.edu/cache/papers/cs/26885/http:zSzzSzranger.uta.eduzSz~alpzSzixzSzreadingszSzYangSigir99CategorizationBenchmark.pdf/yang99reexamination.pdf. August 2004.

45 Siolas, G. and d'Alche-Buc, F. (2000). "Support Vector Machines Based on a Semantic Kernel for Text Categorization". In IEEE-IJCNN 2000.

| Product | Company | Classification Methods Used |
|---|---|---|
| Polyanalyst | Megaputer | • Decision Trees<br>• Fuzzy Logic<br>• Memory based reasoning |
| PredictionWorks | PredictionWorks | • Decision Trees<br>• Logistic regression<br>• K-nearest neighbor<br>• Naïve Bayes<br>• Linear regression |
| Predictive Dynamix Data Mining Suite | Predictive Dynamix | • Statistical<br>• Neural networks<br>• Fuzzy models |
| Previa Classpad | ElseWare | • Neural networks<br>• Decision trees<br>• Bayesian networks |
| Prudsys DISCOVERER | Prudsys | • Non-linear decision tree<br>• Sparse grid method |
| Reel Two | Reel Two Inc. | • Naïve Bayes<br>• Weighted learning; skewed data |
| SERglobalBrain | SER Solutions | • Neural networks<br>• K-nearest neighbor |

Table 3. Categorization Software That Uses Multiple Approaches

# III. SOFTWARE TOOLS

## A. CATEGORIZATION WITH SOFTWARE

The categorization of text documents is a supervised approach to grouping textual objects. More simply, categorizing flat text documents can be thought of as assigning a set of documents to predetermined categories that describe the overall meanings of the individual documents. For example, if hundreds of news articles were to be assigned to known categories such as political, business, and health, we would not necessarily know the specifics about each article but we would have some insight as to what type of information was addressed in each one. If we were only interested in business related news, we could save time by only looking in the appropriate category. In theory, a human could likely look at the title of each article and place it into the proper category and have a potentially high percentage of correct assignments. However, imagine what would happen if there were no titles associated with the articles. The human would now need to read, or at least skim, through the hundreds of text files to determine proper placement. Another human limitation is in the number of articles which he or she can process; even if each article had a title, if there were millions of articles to sort, the effort required to correctly assign the documents would be unacceptable due to time constraints. As an example, it would take in excess of three years to categorize 1.44 million documents even if a person was working twenty-four hours a day, seven days a week, and only spending one minute per document. Assuming that a computer could assign the documents with an acceptable success rate, the automated information technology (IT) method would then be much more beneficial in this type of situation.

The process of sorting documents into known categories for machines may seem simple at first. After all, with relative ease, a human could do the actual sorting without any significant training. However, the actual processes that take place inside a human's mind for this type of task are extremely complex. Therefore complex algorithms are required in order to perform an automated process such as text categorization without significant human input.

Chapter II defined the most common algorithmic methods used in text categorization. It was determined that each approach provided both advantages and disadvantages in attempting to solve the categorization problem. It was however most advantageous to use a combination of methods to perform the desired task of categorizing text documents into predefined categories. By using multiple approaches, either in series or parallel, the arduous task of categorization could be made more effective and more efficient.

There are hundreds, if not thousands, of software packages that address the concept of classification. Classification, however, is broken into two types of analysis: clustering and categorization. This research was only concerned with categorization. This significantly limited the number of software solutions available to address the need. The categorization of text documents into known categories has always been a much more complex process than clustering. Categorization tools were often an order of magnitude more expensive than clustering tools. Some of the best clustering software could examine documents, extract concepts, and then group the documents based upon those concepts. The clustering method would generate categories that were not specified and would either create too many, too few, or undesired groupings.

Categorization goes further than clustering by theoretically learning the predefined categories and then, by algorithmic manipulation, assigning uncategorized documents into them. Thus categorization tools often required a “learning process”. Computers, only have the knowledge that is provided to them and that knowledge must come in the form of 1's and 0's. Therefore, categorization is a more complex process because it adds another limitation to the system. Categorization must not only cluster documents, but it must do it based on predefined categories that have been set by a user.

The learning process that must take place can be inherent in the system (preprogrammed or ‘taught’ by the software engineers) or it can come from the user in the form of training documents or rule sets. The training documents are used to teach the system what a category ‘looks like’, while rule sets shape and weight categories based on attributes. A simplistic way of viewing training documents is to look at them as examples of what should be placed into a particular category.

After an extensive review of various software packages, there were five systems that stood out. Although no one system will likely provide the perfect end-to-end solution, these five systems represent the most likely candidates to perform the task of automatic flat text document classification given our specific needs. They are:

- Reel Two (Reel Two, Inc.)
- SERglobalBrain (SER Solutions, Inc.)
- PolyAnalyst 4.6 (Megaputer Intelligence)
- LexiQuest Categorize (SPSS)
- Enterprise Miner with Text Miner (SAS)

Each of these systems is highly valuable to its parent company; thus, the fine details of how each system works and how each uses various algorithms is propriety. In essence, each system is a "black box" regarding the intricate details concerning the engine that drives the process. We will discuss each of the five systems in as much detail as possible.

## B. EVALUATION SOFTWARE PRODUCTS

### 1. ReelTwo

The heart of the Reel Two solution for automatic text categorization was its patented automatic document classification system, "CS". The "CS" is a supervised learning system. For Reel Two, the supervision comes in the form of training documents. By taking documents that have known categories, the system learned what the category consisted of and what type of documents may be assigned there in the future. In general, a learning system is a computer program that infers the procedures and goals of human tasks and applies that knowledge to perform those tasks more efficiently, consistently, and accurately.[46] The scheme that the system comes up with did not mimic the expert that taught it (i.e. a human), but rather represented a method that produced equally accurate results.

An example of a supervised system is a computerized chess player. The computer does not try to think like a person, but rather calculates all possible moves and countermoves so that it will win the match. Reel Two acts in a similar fashion. It looks

[46] Reel Two. Classification System White Paper; The Reel Two Solution for Automatic Text Categorization. Reel Two, Inc. June 2003.

at how the training documents might be related and then applies those rules to new documents that need to be categorized. Whichever category provides the best fit winds up being the final solution for the document.

Reel Two CS has two phases of operation. The first is the learning mode. Documents with a known category assignment are fed into the system. Theoretically, the greater the number of documents inserted that accurately represent the category, the better the system would perform at categorizing new files. However, bad data certainly decreases accuracy in future categorization attempts. Once the files are selected, a format translator then breaks the documents into a common format that will allow the system to learn. This ensures important details about the text are made salient to the system. Since the training documents represent a known category, the system creates a "classifier" for the now known category. This process takes place for all categories that are used for a set of documents. The classifiers are then used to determine where new uncategorized documents are to be assigned. The user does not need to know any details about the documents or the categories being learned. The determination of which documents are used in the learning process is accomplished prior to run time and does not require hands on manipulation of the software. The user must only ensure that the proper files are being used for the learning of the specific categories.

After a classifier has been created (learned) for each category, the uncategorized documents are sent through the system. This is the second phase of the process, classification. The text documents are compared against each of the defined classifiers to determine which category yields the best fit. The following diagram (Figure 6) shows the flow of documents through the Reel Two CS.

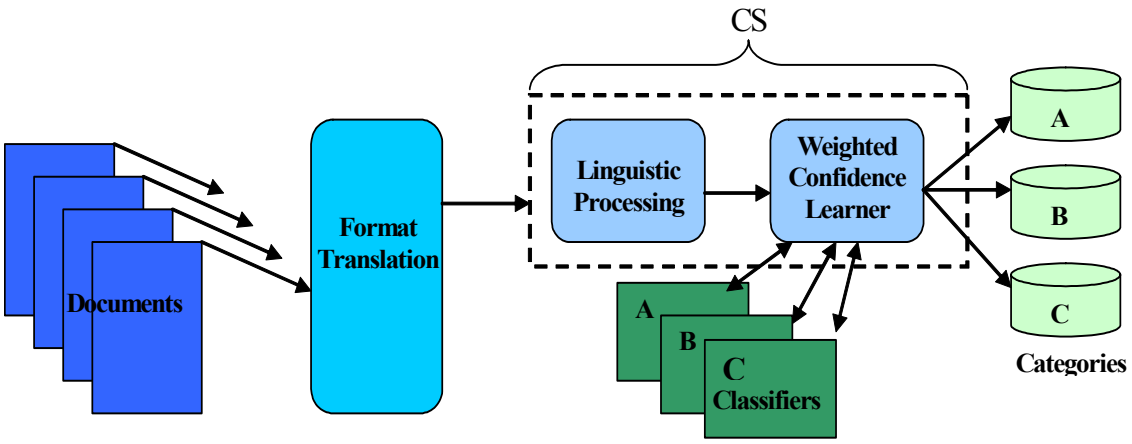


Figure 6. Reel Two System Flow

The Reel Two CS is comprised of both conventional techniques and heuristics. CS uses such conventional tools as stemming and parsing. However, CS only uses stemming when it will statistically improve performance. Stemming is helpful in small documents that contain insufficient vocabulary but is harmful in large documents because it introduces vagueness.[47,48] Thus, the CS technique is beneficial because it only uses stemming if it will enhance performance. CS also puts a spin on parsing by introducing 'shallow parsing'. This technique allows phrase finding, which is fast and able to recognize syntactic structures. Examples include: "an island in the South Pacific" and "the President of the United States." Heuristics are used to improve the system's performance when dealing with uncertainty such as numbers. The following table (Table 4) shows the CS linguistic processing techniques.

[47] Hull, D. A. *Stemming Algorithms: A Case Study For Detailed Evaluation*. Journal of the American Society for Information Science, 47(1): 70-84, 1996.

[48] Xu, J. and W. B. Croft. Corpus-Based Stemming Using Co-occurrence of Word Varients. ACM TOIS, 16(1):61-81, January 1998.

| Method | Description |
|---|---|
| Language Identification | Determining the language of the text (necessary for most other linguistic processes). |
| Tokenization | Isolating words; punctuation analysis; document format analysis |
| Shallow Parsing | Chunking noun phrases, complement structures, periphrastic structures, relative clauses, etc. |
| Case Analysis | Analyzing significance of cased letters for: proper noun extraction, acronym expansion, place name qualification, case-folding. |
| Lemmatization | Stemming of regularly inflected nouns and verbs; normalization of adjectives and adverbs. |
| Numeric Analysis | Determining functions of numeric tokens: ordinal, cardinal, ratios, currencies, phone numbers, zip codes, etc.; estimating relative magnitude and significance. |

Table 4. CS Linguistic Processing Techniques (From: Reel Two: Classification System White Paper)

The Weighted Confidence Learner (WCL) algorithm is the key to the CS process. The WCL is a patented algorithm that is rooted in the naïve Bayes family but has been proven to be more accurate than a standard naïve Bayes algorithm. This is due in part to the fact that unlike other naïve Bayes algorithms, WCL performance does not decline as the number of features becomes very large. The WCL carefully handles the highly skewed training data, differing amounts of noise (error) in different features, and the calibration of the final predictions to estimated probabilities.[49] The WCL uses weighted confidence rather than inferred probabilities from the sum of observations. Thus, the importance of each bit of evidence for predicting the category of a document is translated into a numerical weight and combined into a mathematical function that is then used to predict which category a new document will be assigned too.[50]

The WCL allows additional training documents to be added at anytime to increase the effectiveness of the classifier. This can be done without submitting all the training documents again. Another feature of the WCL is that it uses Leave One Out (LOO) evaluation to determine the fit of each training document. For example, if ten documents are used to train the category 'A' classifier, each document in the set of ten is then

[49] Reel Two. *Weighted Confidence Learner Algorithm*. Reel Two Technical Brief; Reel Two, Inc.

[50] Ibid.

compared against the other nine. This process reveals if any of the documents do not belong in the training set and also helps to minimize the number to training data needed to produce an accurate classifier.

Each classifier is adjustable. The confidence level of a classifier can be set by the user to increase or lower the degree of certainty required for a document to be assigned to a particular category. As the defined threshold rises, so does the precision. The trade off is that a document that should be assigned to a category might be left out. A low threshold could result in an incorrect assignment. Thus, thresholds must be set at optimum levels. Optimum levels are not normally known prior and must be determined through modification of settings over the course of several iterations. The default setting in CS is to set the threshold at the point where the number of false positives and false negatives reach a break-even point. The user may then adjust the system they see fit. The importance of false positives and false negatives can also be weighted. Thus, the user can skew the break-even point in the event that one error is more acceptable than another.

The final output of the CS process shows which documents were assigned to the given categories. Statistics can then be viewed as to how close a match the document had with its particular category. If a document fits into more than one category equally, it will be placed into both. Overall, the output allows the user to easily rank the relevancy of a document into a particular category. The test results that Reel Two reports will not be discussed, but rather the application will be tested to determine its ability to categorize flat text profiles in to defined categories.

The following is a list of the applicable settings that could be adjusted by the user[51]:

- Taxonomy Depth (Full or One Level): Taxonomies can either be imported so that categories are created for every directory within the imported hierarchy or for the top level directories only.
- Exclude Empty Directories (Yes or No): Prevents the import of empty folders/directories.

[51] Reel Two. Reel Two Help Guide within the Reel Two Classification System Version 2.4.6.

- Subsumption (Yes or No): A document is placed into a category based upon the entire directory that it came from rather than its individual content.
- Check for Duplicates (Yes or No): Tests file names or file content to prevent duplicate entries. This process slows the system down considerably.
- Split Documents (Yes or No): Splits documents by page boundaries during the import process. Useful when classifying large documents.

**2. SERglobalBrain**

The desire of SER to transform information into knowledge is attempted through their SERbrainware software package. SERbrainware is the core software engine for several applications that range from knowledge enabled solutions to integrated document management. The four core functions of SERbrainware are[52]:

- Association: SERbrainware does not rely on keyword or phrase comparisons, but is able to associate a network of elements in a given context. SERbrainware can access knowledge that is context relevant, based on an associated network of elements, not just a single keyword or phrase.
- Classification: After being trained on a given example of elements (the "learned set"), SERbrainware is able to recognize and categorize information into classes based on content, not just format.
- Extraction: SERbrainware can identify specific data, extract it and pass it along to a line of business application or individual for subsequent processing. SERbrainware is able to analyze both structured and unstructured documents to identify and capture the required data.
- Memorization: The key to optimally reusing knowledge is to store it so that it can be easily accessed at the appropriate time. SERbrainware stores the extracted information, the learned categories, associated content, and extraction patterns.

Of particular interest was SERglobalBrain which enabled the classification aspect of SERbrainware.

SERglobalBrain used neural network technology along with mathematical algorithms to simulate how the human mind analyzed concepts and categorized data. The system supposedly did not only use keywords or phrases to categorize documents, rather

52 SERbrainware. SERbrainware: The Full Perspective. Version 2.1 White Paper; SER Solutions, Inc. October 15, 2001.

it learned by example; all of the words within the document were used for analysis in order to extract conceptual elements which were then analyzed to find the meaning of the document.

SERglobalBrain initially appeared to be a typical search engine. However, it performed the task of categorization by using natural language formats to identify the elements (concepts) of documents vice searching for specific text and/or keywords. It had the capability to process two-hundred-twenty-five common formats which proved it preprocessing robustness.

A learning set was provided that embodied the overall concepts of a particular category because the software 'learned by example'. Documentation mentioned that typically, five to twenty-five representative documents had to be provided per category for the purposes of learning. Thus, the user was essentially teaching the system the definition of a category by providing the sample documents in the form of a learning set. Through the graphical user interface (GUI), the user created 'interest profiles' which were the categories that future uncategorized documents would be compared against. The profiles could have consisted of multiple subcategories or they have been flat in structure. Once the categories were defined and populated, the system compared the documents within each learning set and determined the overall match among the cluster. The essential concepts for the group were stored as that category's 'meaning'. If a document was determined to be an outlier, i.e. the system found unique concepts for a document, the system would display that information to the user via its 'Learnset Viewer' GUI. At that point, the user inputted the repository of uncategorized documents, had them preprocessed, and the system classified them into the most appropriate categories. The output was then displayed to the user through statistical and graphical means.

SERglobalBrain defined this process as a 'search'. It was searching the uncategorized documents to find its concepts which were then compared to the concepts associated with the learning sets. Each document was then placed into any category to which it 'fit'. If a document was placed in multiple categories, statistics showed the best fit. However, if the document was placed in many categories without a statistical significant difference, it was an indication that the training sets were not far apart enough

in meaning to distinguish between future unclassified documents. The research did support the indication of insufficient distance between categories. In order to deal with the lack of significance the minimum confidence level was reduced to its smallest value in order to produce a definitive category. If a document was split between categories with the same value then it was considered as a result of 'other' (see Chapter IV for a full explanation).

The software had the capability to employ the following 'search' methods[53]:

- Fuzzy searches: SERglobalBrain uses a fuzzy search to provide fault tolerant results on search queries. There are many circumstances where the data input into the data repository may have misspellings—due to OCR/ICR errors, etc. There may also be different spellings of a person's name or words that are of foreign origin. Because SERglobalBrain searches for content instead of exact word matches, it is extremely fault tolerant.
- Natural Language – Phrase Search: The input query can be entered as a word, combination or words, phrase, sentence, paragraph, or even the content [or concept] of an entire document. Unlike traditional natural language engines that simply parse the query into a Boolean string, SERglobalBrain searches for the content of the entire search query.
- Exact: SERglobalBrain provides confidence ratings for the returned results of a query. Only those with an exact match of the word or phrase will be returned with a relevance of 100%. The exact search can also be used to find documents that contain a set of words, not necessarily in the same order.

In addition to these search styles, each method had additional functionality such as 'positive' or 'negative' words / concepts. These represented the must-contain or must-not-contain words, phrases, or concepts for a designated category. However, using this functionality required a comprehensive working knowledge of the system and the concepts contained within the data set.

SERglobalBrain was designed to replace the manual effort required to classify and index documents. The solution claimed to mimic the human learning process by classifying documents based on similar content.

53 SERglobalBrain. SERglobalBrain. Technical White Paper; SER Solutions, Inc. April 2004.

The following is a list of applicable settings that were modifiable[54]:

- Minimum Pattern Size (2-10): Only words with at least the minimum pattern size (characters) are taken into account.
- Maximum Dictionary Size (100-100,000): The maximum number of words to be inserted into the dictionary from the documents being processed.
- Include Numbers (Yes or No): Includes numbers within the learn set (training data set).
- Absolute Relevance Value (1-100): A document must have at least the specified relevance to be placed into a particular category.
- Relative Distance of Relevance Values (1-100): Sets the relevance threshold between categories in order for a document to be distinguished between them (two competing categories).

### 3. PolyAnalyst

PolyAnalyst 4.6 developed by Megaputer is a data/text mining system that has several machine learning algorithms that can manipulate data in several ways. (Figure 7) Each algorithm is designed for a specific purpose but many provide overlap on their capabilities. Some of the exploration engines deal with the problem of categorizing text data. PolyAnalyst allows a user to select a single method or to combine methods to produce more accurate results. Although PolyAnalyst consists of over fifteen mining techniques, the most applicable to our research are Nearest Neighbor (Memory-Based Reasoning), Decision Tree, Decision Forest, Text Categorization, and Taxonomies.

54 SERglobalBrain SERglobalBrain Personal Edition User Guide Version 1.7.0.

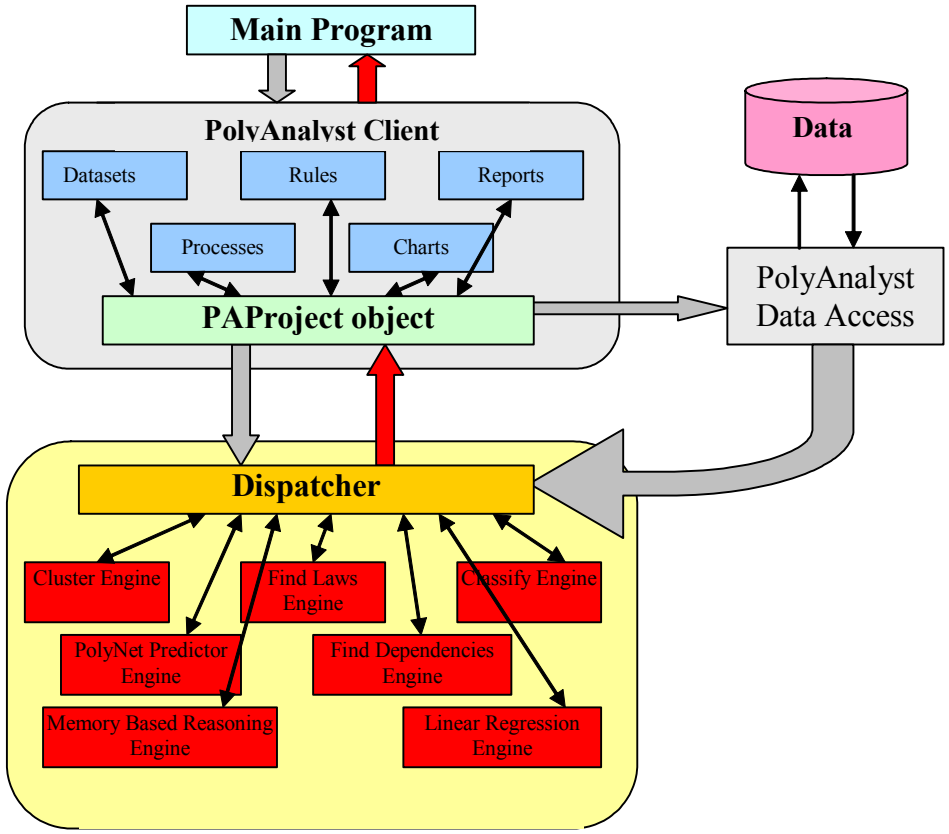


Figure 7. PolyAnalyst 4.6 Machine Learning Algorithms (From: Megaputer. PolyAnalyst 4 User Manual. August 2003.)

### *a. Nearest Neighbor*

The Nearest Neighbor exploration engine that PolyAnalyst uses was a memory based classification system that assigned values to data points (attributes) based on the proximity of other attributes. It was designed for prediction of attributes, whether discrete or continuous, and to classify data into multiple categories. This method requires a large amount of data (discussed below in scalability) to make accurate predictions. The memory-based reasoning algorithm compares new attributes to the training set attributes that were used to create the rules used in classification. Therefore the larger and more accurate the training data set, the better the results of the system. A down side of this algorithm is that it is time intensive; the computational time is proportional to the square of the number of records.[55]

[55] Megaputer. PolyAnalyst 4 User Manual. Megaputer Intelligence, Inc. August 2003.

The underlying technology of the nearest neighbor process is the comparison of the rule set that the system generated compared to the findings in the extracts of the document. Rules contain references to the dataset that the system was trained on or learned. The training process can be the pre-programmed system from the software designers or user specified. In comparing attributes found in new documents, each new attribute from the new document is compared to all the known attributes from the training process. A function is created that minimizes the distances between all attributes found and all attributes known. Through this minimization process, a best fit or nearest neighbor is found and a category is populated.

The scalability of this memory-based reasoning is limited. The optimal number of records is between 2,000 and 50,000, but the minimum and maximum are 100 and 100,000 respectively. Another limiting factor is the maximum number of attributes, which is only three hundred. This may lead to problems when dealing with a large amount of documents that are not closely related. The algorithm outputs are classification error, classification efficiency, and other statistical data that describes the fit of the documents based on the training set. The results are not particularly intuitive, but the engine will display a structured list of the categories and documents.

***b. Decision Tree***

The Decision Tree engine is also designed to solve the problem of classifying items into multiple (broad) categories. This process is the quickest algorithm that the system supports for classification of a large number of records. It also produces a simple output for interpretation.

The algorithm is designed to handle between 100 and 5,000,000 records at a time, but in order to work efficiently on larger amounts of records it was recommended to pre-process the data and remove excess attributes that are not useful to the exploration engine. The algorithm scales well linearly with the introduction of more data columns, but it grows more than linearly with the introduction of more records, N*log(N)[56] (N=number of records).

[56] Ibid.

It is recommended by the vendor to conduct pre-processing of the data prior to beginning the decision tree exploration engine. By calculating 'summary statistics' (another engine within the suite), users can deselect attributes that may not provide any insight for the exploration engine. This pre-processing can be labor intensive but may have significant impact on the Decision Tree engine's performance.

The underlying algorithms for this exploration engine are 'Information Gain' splitting criteria, Shannon information theory, and statistical significant test. Information Gain is a measure of quality for the splitting of attributes at a node. The purpose is to find meaningful splits that eventually break into unique criteria.[57] Shannon information theory says that a random variable with a specific probability density function can, through mathematical methods, define specific entropies.[58]

The report produced by this engine provides measures of non-terminal nodes (nodes where splits occur), number of leaves (terminal nodes), and the overall depth of the tree. In addition, it provides statistics on the decision tree such as the total classification error, percentage of undefined prediction cases, classification probability, classification efficiency, and classification error for each node. Other outputs include charts for interpretation, which include a final output of all records and the nodes that have ownership.

### *c. Decision Forest*

Decision Forests are used to categorize an attribute into several categories. Decision Forests differ from decision trees in that decision trees may lack the efficiency and accuracy needed in more precise applications. They also do not limit the data to only one decision tree.

The Decision Forest engine requires that the data be preprocessed with another algorithm, Text Analysis. Text Analysis moves through an entire document and summarizes the important concepts so the output is simplified to 'key' attributes. Once

---

[57] Aixploratorium. "Information Gain Seeking Small consistent Decision Trees." http://www.cs.ualberta.ca/~aixplore/learning/DecisionTrees/InterArticle/4-DecisionTree.html. October 2003.

[58] Page, D. D., A. F. Koschan, S. R. Sukumar, B. Roui-Abidi, and M. A. Abidi. "Shape Analysis Algorithm Based on Information Theory." University of Tennessee, Knoxville, TN. http://imaging.utk.edu/publications/papers/2003/page_icip03.pdf. August 2004.

the preprocessing is completed, the decision forest algorithm takes the summarized data and processes it through its algorithm to obtain the category in which the document fits 'best'. The Decision Forest engine works most effectively when trying to categorize records containing natural language text into known categories.

The underlying algorithm of Decision Forests is a decision tree. Several decision trees are constructed (several trees make up a forest) such that each one defines one classifier. The combined classifiers are then called a decision forest.

The Decision Forest engine requires education. This means that the engine needs to go through a learning process in order to obtain statistically verifiable results. For each target class or category, it is recommended that at least thirty records be used to represent that class.[59] In order for this engine to function it also requires a defined maximum and minimum number of levels for each tree (the depth of the tree). The last input required is the 'node splitting threshold'. This is essentially the level of type I error. In order to minimize the probability of making a false positive (Type I error), the node splitting threshold should be set to 'pessimistic'.

The outputs of the decision forests are similar to that of decision trees. Since one classification tree is constructed for every target category, the output is a series of decision trees. However, accuracy and precision may be sacrificed because documents can be assigned to multiple categories.

A limitation of Decision Forests is the amount of records required in order to process. The actual bounds of the system are a minimum of 1,000 records and a maximum of 10,000,000 records. More records result in better classification rules. Decision Forests provide confidence to users when a large volume of records exists and the system is well 'educated'.

***d. Text Categorization***

The Text Categorization exploration engine was specifically designed for the automated categorization of unstructured text. The text Categorization engine automatically builds a hierarchical tree-like taxonomy of topics and subtopics that are extracted from the unstructured text.

[59] Megaputer. PolyAnalyst 4 User Manual. Megaputer Intelligence, Inc. August 2003.

Text categorization of unstructured text is one of the most difficult tasks to accomplish and therefore requires layers of input to the system. Required settings are 'thematic context key sense type', 'text distribution mode', 'subcategory setting', and 'minimum node support'.[60] The thematic context is the environment in which a certain word exists. The English language, for example, contains words that sound alike (homonyms) and are spelled the same but have different meanings. For these words the system requires the context for which the homonyms are being used. To accomplish this, the system provides broad categories for the text (business, science, slang, art, technology, etc.). The key sense type is the determined significance of the words in the respective text field. Essentially, this is how often the word exists in the text field. However, product specifications are unclear regarding rule sets and must be treated as a 'black box'. The text distribution mode determines how terms are distributed among child nodes of a given classification node. This can vary from placing terms in the first available child node to more accurately placing them into the most significant node. The subcategorical settings are variations to the tree structure; they control the tree branching criterion by allowing the user to specify few to many subcategories (a specific number can not be specified). The last specification is the minimum node support setting. A user can set the minimum number of records that must occur in a given node for a node to exist.

The underlying algorithm in the Text Categorization engine is the Text Analysis engine that was described within Decision Forests. By finding 'true meaning' and using multiple subdivision splitting techniques (PolyAnalyist does not define these methodologies), a taxonomy is performed.

The data produced by this report displays the categories in tree-like structure along with the count of records pertaining to each category. Drill down functionality can be used to locate the actual records that belong to a particular category. Additionally, through the reporting features, categorization rules can be generated.

[60] Ibid.

The optimal number of records the algorithm can process is 100-50,000 at one time. Minimum and maximum values are not specified. The scalability of the system is also good; it grows linearly in regard to the number and length of the records. A helpful, but potential time consuming feature that is built into the system, is the preprocessing requirement of foreign word definition. This includes abbreviation, synonyms, stop words or stop phrases (user defined stopping points for the exploration). A bulk import can occur if pre-existing dictionaries are present, otherwise time-consuming customization must take place.

***e. Taxonomies***

An additional feature of PolyAnalysist is the ability to specify a user-defined taxonomy. These taxonomies are custom build categories that have all the specified criteria the system will look for in order to classify a record into a custom category.

The primary feature of the exploration engine is its ability to build a tree-like structure with the categories as nodes; words or phases are then used to specify each category. Additionally the algorithm allows the ability for the nodes to have subsets or child nodes with their own defined list of discriminators. A specific classification mode must therefore be determined; this will specify how matching of terms will occur. Depending on input specifications, the algorithm will place the record in the first matching category, the most significant category, or all the categories that apply.

The output from this engine is similar to the Text Categorization output, but it only provides the tree structure, the record count for each node, and the locations of the specified discriminates.

This engine provides predefined categories along with discriminates for each. The drawback is that the accuracy of the algorithm is strictly based on the words used to define each category. Therefore this methodology requires a well-defined taxonomy for the algorithm to produce accurate results. With the exception of foreign words, this is the only exploration engine that allows users to modify the defining lists of identifiers for nodes.

The following settings are adjustable by the user[61]:

- Key Sense type (All, More Frequent than Default, or Most Significant): Determined by significance of the word in the text field.
- All: The word is present in the respective text field for the record at least once.
- More frequent than default: The word is present multiple times within the record; more frequent than in average English text (specific value is not defined).
- Most significant: The word is significantly more frequent in the respective text field of the record than in average English text (specific value is not defined).
- Find Collocations (Yes or No): Collocations are sequences of words which occur in text frequently but are not phrases registered in its internal dictionary.
- Key Operators: (Used with Taxonomies).
- **" "** Used to specify more than one term such as a phrase. Quotation marks are not used for single words.
- **!** The explanation mark is used if the user desires to add the term's synonyms dictionary. The Specify synonym dialog will appear once the word expression box looses focus allowing the user to select the proper synonym from the list.
- **[ ]** Brackets around a term will include all synonyms for the given term rather than prompt the user for individual synonyms.
- **( )** Parentheses are used to set precedence of the users expressions. Words with parentheses have a higher priority then all words not in parentheses.

**4. Lexiquest Categorization System[62]**

LexiQuest Categorize is a linguistic and probabilistic powered application that attempts to automatically categorize records into predefined categories based on the content of the record. It is advertised as an end-to-end solution with the ability to text mine records, develop statistics on the concepts, and place the records into known/specified categories.

The system is mainly composed of two items. The first is a term (concept) extractor that parses the incoming unstructured records. The second is a categorization

61 Ibid.

62 SPSS. "LexiQuest Categorization System Algorithms." SPSS. Chicago, IL. October 2003.

engine that uses the extracted concepts as an input into its categorization algorithm. The extractor (text mining piece) is a linguistic processor that searches for 'key' words, pulls them from the document, and puts them into a structured environment along with given characteristics about that word (or concept). The categorization engine controls the extractor and ensures that all the concepts that have the same weight are extracted. Along with the concepts, the number of occurrences of a particular term is pulled from the document and recorded.

The categorization algorithm works as follows. Based on a training set, whether it is a manual description of each category and the attributes of that category or a set of learning records, an index is built within the system that comprises the categories and defining attributes. Based on this index, which can be thought of a tree structure or table, a search is conducted comparing the extracted concepts and the indexed table. The search compares the concept to the category and attribute: if a match occurs, a weight is placed on that concept to specify that it belongs to the matching category. As concept matches occur (which can occur over different categories), a summation is calculated and the predominant category will take 'ownership' of the record that has the largest number of matching concepts. It is noted that certain attributes can have higher weights than others. Therefore a higher weighted attribute that matches an extracted concept will have a greater overall value as it pertains to the determinant category.

The scalability of the system is not specified, but the actual amount of time the system requires to process records in not linear in all aspects. As the number of terms extracted from each record grows, the categorization time is linear. As the number of terms in the indexed table grows, the time is linear (if the search time is linear). However, as the number of target categories increases, the amount of time to process grows more than linear because each additional category will not have the same amount of defining attributes.

This categorization process claims to be more than a simple word comparison due to the ability to assign weight to defining attributes. In addition, rules can be established to assign records into specific categories based on specific concepts found. The limitations of this system are similar to that of the Taxonomies in PolyAnalyst. The

system is only as good as the category discriminates and weights that are assigned to them; therefore detailed knowledge is needed to generate the categories and defining attributes of those categories.

The following settings are adjustable by the user[63]:

- Minimum Number of Categories per Document (Any numerical value): The minimum number of categories that should be returned for a single document. If the document does not fit into any of the predefined categories, no category is assigned regardless of the value of this parameter.
- Maximum Number of Categories per Document (Any numerical value): The maximum number of categories a document can be assigned too. This value cannot be less than the minimum number of categories value. This setting allows documents to belong to more than the specified minimum number of categories if it is higher in value than the minimum.
- Maximum confidence level for single response (0-100): Each document is given a series of confidence scores (one for each of the predefined categories). The score is compared to the user input confidence level to determine if it will be assigned to that particular category. Example: if the value is set to 40, only those categories with a confidence value greater than or equal to 40 will be returned.

**5. Enterprise Miner[64]**

Enterprise Miner is designed to be an end-to-end Data and Text Mining solution developed by SAS. The Application provides a flexible framework designed to conduct various Data Mining tasks within structured and unstructured formats. The specific applications or subcomponents of interest to PsyCLPS include the Text Miner component and several of shaping, filtering, and visualization tools that are resident within Enterprise Miner.

The Text Miner has the capability of accepting text from several applications in multiple formats. These range from simple ASCII text files to vendor specific applications such as IBM DisplayWrite, without additional preprocessing or post import processing from the user.

---

63 Ibid..

64 SAS. "Getting Started with SAS Text Miner Software Release 8.2." SAS Publishing. SAS Institute Inc. Cary, North Carolina. 2002.

The text parsing feature within Text Miner has embedded capabilities that enable the user to:

- Break sentences or documents into terms.
- Extract particular entities that were meaningful to specific applications.
- Find the root form/stem of words and specify synonyms.
- Remove extraneous words that provided no additional value such as *a*, *an*, and *the.*
- Identify the term's part of speech.
- Create a quantitative representation for the collection of documents.

These features, which are associated with the preprocessing of the data/text mining step, are labor intensive for users. Software vendors continue to try to simplify the preprocessing phase, but unless the 'created rules' are generalized for all domains, the risk of producing 'findings' that are inaccurate based on faulty preprocessing. Text Miner allows the user to specify its level of parsing capability. (See Figure 8)

The text parsing feature breaks down text into components beyond the level of simple-words. This is an extremely import feature that distinguishes concept extraction from simple word extraction. An example of how Text Miner parses terms is available in Figure 9. Additional features are available that have the ability to handle entities (concepts). To date the concepts that Text Miner has been able to 'single out' are included in Table 5.

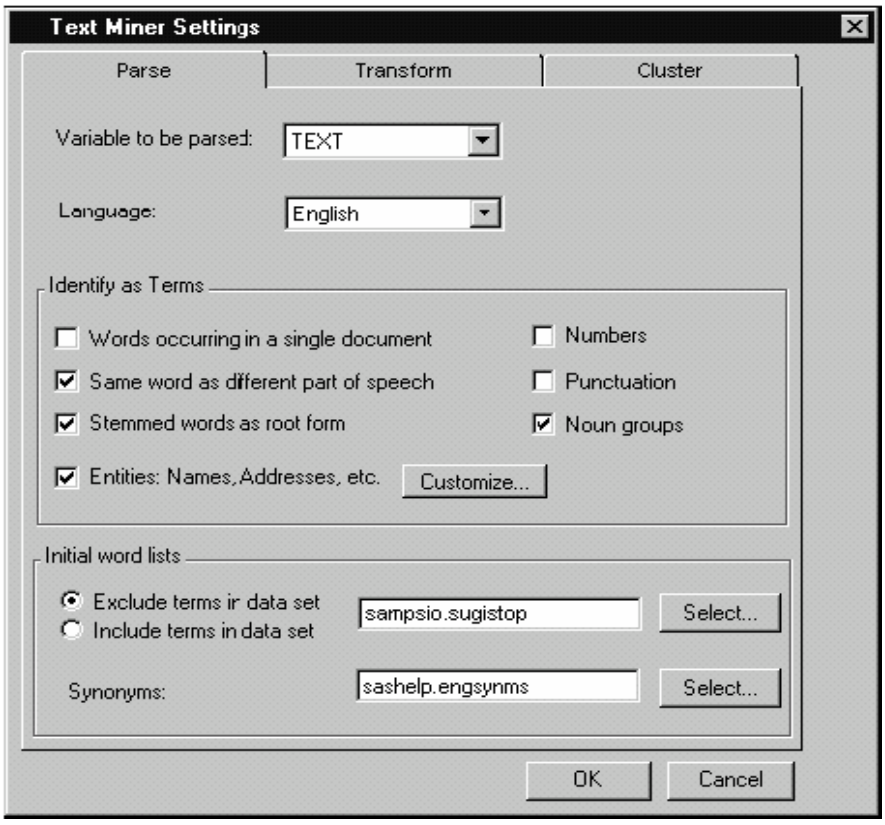


Figure 8. Text Miner Parsing Settings (From: SAS. "Getting Started with SAS Text Miner Software Release 8.2". SAS Publishing.)

The ability to find the root form/stem which SAS calls stemming can be interpreted as finding and returning the base form of a word. For ease of interpretation for the user and system, this feature provides the ability to work through the specific tense of a word and returns the basic word that was evaluated. The feature is valuable when specific words are identified as 'key'. A simple word search on a key term may return zero 'hits', but with this feature, 'hits' are returned no matter what the tense of the word that is present. In addition to the stemming feature, Text Miner has a synonym feature, but this is user defined. The same stemming capability exists with the synonym feature.

| | |
|---|---|
| • Address | • Percent |
| • Company | • Person |
| • Currency | • Phone |
| • Date | • Product |
| • Internet | • Social security number |
| • Location | • Time |
| • Measure | • Time period |
| • Noun group | • Title |

Table 5. Enterprise Miner Concept Categories

Once the parsing portion of the text mining is complete, Text Miner transforms the text into a numerical form for further analysis. Based on weighted functions, a matrix is created that is used to reduce the parsed text documents into similar groups or clusters. Because the goal is to classify documents based on a known taxonomy, which SAS calls expectation-maximization clustering, training documents (pre-categorized documents) are used to define the taxonomy.

Through manipulation of specified events (concepts) on which to classify and the selected confidence thresholds for the categories different results are attained. This is where human intervention is applicable; settings must be manipulated to try and find an optimal setting for the application based on the defined categories. During testing, several settings will be evaluated and corresponding results produced so that empirical evidence can dictate the accuracy of results.

| Sentence | Parsed Terms |
|---|---|
| Coca-Cola announced earnings on Saturday, Dec. 12, 2000. Profits were up by 3.1% as of 12/12/1999. | coca-cola<br>announced<br>earnings<br>on<br>saturday<br>december<br>12/12/2000<br>profits<br>were<br>up<br>by<br>3.1%<br>as of<br>12/12/1999<br>Coca-Cola<br>Saturday<br>1999-12-12 |
| Douglas Ivester earned $12.8 million. | douglas<br>ivester<br>earned<br>$12.8 million<br>Douglas Ivester<br>12800000 USD |

Figure 9. Text Miner Parsing Capability (From: SAS. "Getting Started with SAS Text Miner Software Release 8.2". SAS Publishing.)

## C. CONCLUSIONS

Each product (Reel Two, SERglobalBrain, PolyAnalyst, LexiQuest Categorize, Enterprise Miner) presents a solution to flat text document categorization. Each system will be tested for its ability to perform the task of categorizing documents into know categories using identical test data. The software package that performs the best for the given application will be determined through the defined evaluation process mentioned in Section III of Chapter III.

Since it has been determined through literature review that the use of multiple categorization methodologies provides the best overall results, it is no surprise that the five applications discussed use multiple methodologies in their approaches. Each application will have advantages and disadvantages associated with them, but accuracy will be the determining factor for the identified 'best' system.

THIS PAGE INTENTIONALLY LEFT BLANK

# IV. METHODOLOGY AND EXPERIMENTAL SET UP

## A. SOFTWARE TESTING METHODOLOGY

Before embarking on a full explanation of how the hypothesis will be supported or disproved based on the testing and evaluation of the selected software solutions, it is important to highlight the level of complexity of the overall problem. The hypothesis is that the laborious task of identifying which psychological category people fall into (based on a personal profile) can be accomplished through IT means. It is very important to establish both a testing methodology and evaluation criteria for what was considered a success and what was not. Thus, various testing iterations, evaluation decision points, and overall stopping points for the hypothesis acceptance or decline were needed.

Regardless of the method however, there were certain common requirements that had to be met for each system prior to beginning the testing phase. Some of these requirements were: robustness of data (minimum number of records to evaluate), commonality of language (English only), and pre-defined categories. Based on the system under evaluation additional requirements were levied. This includes systems that used neural networks that had to be initiated with a learning process.

The data attained (individual profiles) for testing was unstructured and required significant preprocessing. Although this is a common starting point for the applications under evaluation, further recommendations will be made for simplification of the pre-processing. Additionally, there were significant holes in the data sets that required resolution prior to testing and evaluation. These included profiles that were incomplete or had low psychological significance and were primarily a chronology of a person's life rather than a profile that included a person's motivations, beliefs, and thoughts. For each of these holes, data manipulation had to occur and a social psychologist was brought in to assist. Each required manipulation will be addressed in a sequential manner.

The minimum requirements in the methodology fell below recommended levels from the application vendors, but this does not discount the capability. The vendors of the applications built the systems for the task of text mining on large data sources (greater than a thousand), but the evaluation was conducted on numbers less than five hundred.

The overall minimum requirements that were accepted for this experiment were: at least three (ideally mutually exclusive) categories, profiles of personnel that were more then a simplistic chronology of a person's life, one hundred overall profiles in flat text format, at least thirty profiles per category for overall evaluation.

The purpose of these minimums was for the significance of mathematics and psychology. Because the profiles are only samples of a target population statistical analysis was used to describe the characteristics of the population. In order to assume that a sample is representative of its population it must have a minimum sample size of thirty based on the Central Limit Theorem[65]. For a profile to have significant attributes about a person's 'being', it must be more robust then the accomplishments of a person over the course of their life. Although life events tell a story about the type of person someone is: thrill seeker, recluse, aggressive, etc. it is only representative of their actions and not their motivations.

The software solutions under examination were addressed in this chapter in Section II. For each system, the specific requirements and inputs that were entered are addressed in the testing/evaluation and results portion of this thesis, Chapter IV.

Testing will involve multiple iterations with varying parameters for each system. We will modify all applicable parameters, diagnostic and visualization tools to monitor, maintain, and fine-tune the performance of each product. The goal is to find optimum settings for our applications without large amounts of insignificant testing. Since each test/iteration requires significant computational time and effort, each system will be limited to no more than ten iterations. The settings that result in the most accurate outcome will be reported along with the results of each system.

In an attempt to find the optimum settings for each tool, applicable parameters will be adjusted and the results for each iteration will then be compared to both previous iterations of that system and 'true' results[66]. Concurrent settings will depend on the

---

65 McCabe, George P. and David S. Moore. *Introduction to the Practice of Statistics, 4th Ed.* W H Freeman and Company, 2002.

66 'True' results: Prior to testing all the records will have pre-assigned categories. External professional(s) will have compiled the categories and assignment of records. The results obtained from the external professionals will be considered to be the 'absolute' correct answers for which each system will be evaluated against.

results of the previous iterations. This method will prevent unnecessary testing. For example, if a parameter is set to a higher sensitivity in iteration #2 and then even higher for iteration #3 and the results are becoming less accurate, there will be no reason to pursue further adjustment of that parameter in that direction. Rather, the parameter will be adjusted in the opposite direction until results diminish again. This will give an approximate setting of optimization. Since some parameters have infinite possibilities it is an impossible to test every possible combination. Therefore, only specific applicable parameters will be adjusted during testing; the first iteration settings will be based upon literature review and suggested settings from the software vendors.

Once testing is complete for each system and the 'best' results are attained, a verification process will be undertaken using those 'best' results. The verification process is designed to validate the original output by applying the exact settings that originally produced those results. If the results from there-test produce the same overall accuracy as was produced by the exact settings in the original testing phase, then it is evident that the specified settings for a particular system will perform at that threshold consistently.

Evaluation will involve the comparison of the best verified results from each system. Each system will be scored will an overall percentage of fit. The percentage of fit is attained by taking the total number of correct matches (records to categories) divided by the total number of records.

The score for each system is then compared against the other system's score and the largest numerical value will determine the 'best' system tool. This will not be the only scoring of overall system performance, but will be the determining factor for accuracy. Other intangible, non-accuracy related, factors such as usability will be considered during final selection of the 'best' tool.

## B. THE EVALUATION DATA SET

In order to evaluate the software packages, suitable data, that provided sufficient psychological information, was required. Real world data of individual profiles was found and attained from a sensitive source in order to conduct the categorization/segmentation. The data consisted of 349 leadership profiles from a

population of unknown size. (See Appendix C for a classified description of the data) Although approximately 375 profiles were randomly selected from the population, only 349 profiles consisted of usable data. For example, some profiles did not provide an adequate description of the person; rather, they consisted of chronologies listed by date in a bulleted format. Other discarded profiles consisted of only a single paragraph that offered insufficient insight into a persons beliefs, goals, attitudes, or psychological characteristics. Only profiles that a human could categorize were extracted from the population. Therefore there was an injected bias as to what profiles could be considered for further testing. This type of discrimination was consistent with the goal of the research since it was assumed that usable data existed for any individual that would be psychologically targeted. Without information on a particular individual, it is not possible to know what targeting method would be most successful.

Each profile consisted of a one to three page description of a single person. Each contained various facts about an individual to include chronologies, career information, personality traits, social interaction, and examples of how that person responded in various situations.

### 1. Generated Categories

A social psychologist, specializing in influence, from the University of California at Santa Cruz, was provided with the 349 profiles. His task was to determine if the profiles could be categorized into specific categories based on their content. After approximately 20 to 30 profiles, he determined that four categories were adequate to describe the major personality types in the data set. These categories were consistent with prior research in this area.[67],[68] The four categories were: Power, Personal Standards, Social approval, and Other. The following is a descriptive list of the four categories.[69]

---

[67] Greenwald, A.G., and Breckler, S. To Whom is the Self Presented? In B. Schlenker (Ed.), *The Self and Social Life*. New York: McGraw-Hill, 1985.

[68] Greenwald, A. G., and Pratkanis, A. R. The Self. In R. S. Wyer and T. Srull (Eds.), *The Handbook of Social Cognition*. Hillsdale, New Jersey: Lawrence Erlbaum, 1984.

[69] Pratkanis, Anthony, R. (in press). Social Influence Analysis: An Index of Tactics. In A. R. Pratkanis (Ed.), *The Science of Social Influence: Advances and Future Progress*. Philadelphia: Psychology Press.

Power: the power hungry is interested in promoting his or her power and will most likely be calculating the benefits and cost of any course of action in terms of reaching that goal. In general, these are people who would betray others for personal gain and would not be considered trustworthy.

Personal (private) and achievement orientation: has strong personal principles; sees political activity as a way to achieve those principles; internal locus of control; honors commitments; gains self-esteem from living consistent with principles and making achievements that he or she finds important.

Social approval / Public orientation: gains self-esteem by approval from other in general; motivated by being approved of by others in general.

Other: individuals that exhibited varying personality traits. Often these individuals were a conglomerate of the other categories; thus, these individuals were too hard to categorize into one group. Additionally, the Other group contained Collective oriented people. The Collective orientation gains self-esteem from a specific group (family, ethnic group, religion, etc.) and is motivated to perform consistent with the group.[70] Collective was not used as a separate category because too few profiles (less than ten) were found to fit this description. The Other group also consisted of profiles that did not provide adequate information for the social psychologist to determine which category they best fit into; although their profiles were several pages long, there were no behavioral descriptions or events described that explained motive and thus prevented a distinct categorization. The Other category was therefore a default category to which any profile that did not fit into the three primary categories was placed and was considered undefined and dimensionless in nature.

Along with these descriptions, tactics were identified to influence each type of category; these tactics are found in Appendix C.

### 2. Expert Results

After roughly 60 hours of reading profiles, the human expert on social psychology categorized 256 documents into the four predefined categories. Figure 10 shows the overall distribution of the 256 profiles that were categorized via the only current existing

[70] Ibid.

method: manual human effort. Given the vast effort required to categorize the documents, the remaining 93 profiles were not manually categorized for the testing because an adequate number of profiles had been categorized in order to test the IT tools; the additional time required to manually categorize these documents was not cost effective. *This reinforces the purpose and the need for this research. If the sample had consisted of 10,000 profiles, the social psychologist would have only categorized 2.5% and required over a week to do it.*

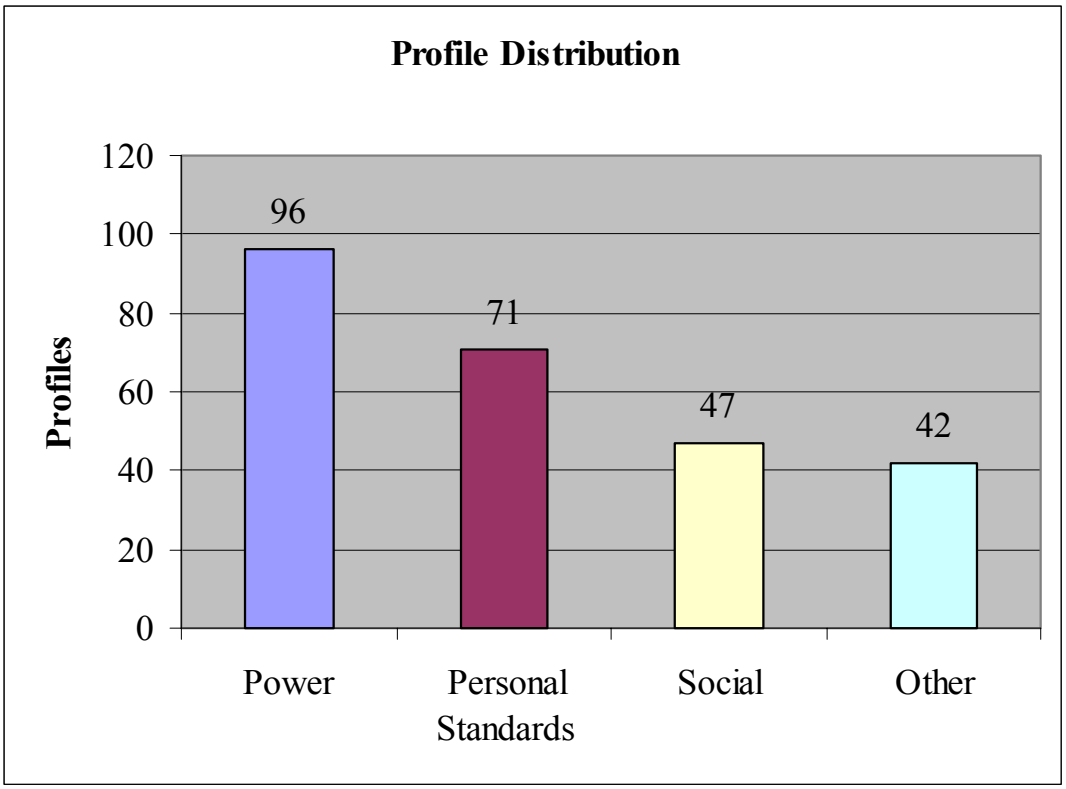


Figure 10. Profile Distribution

Table 6 shows the percentage of each category within the data set. By definition of the Central Limit Theorem, it is expected that the same distribution of categories exists within the population.

> The Central Limit Theorem says that the distribution of a sum or average of many small random quantities is close to normal. This true even if the quantities are not independent (as long as they are not too highly correlated) and even if they have different distribution (as long as no one random quantity is so large that it dominates the other) the central limit

> theorem suggests why the normal distributions are common models for observed data. Any variable that is a sum of many small influences will have approximately a normal distribution.[71]

However, this distribution is for the population under evaluation. Other populations are expected to have varying distributions; this is further discussed in Appendix C.

| Power | Personal Standards | Social | Other |
|---|---|---|---|
| 37.5% | 27.7% | 18.4% | 16.4% |

Table 6. Profile Distribution Percentages

Table 6 also shows how likely a profile, randomly placed into one of the four categories, was categorized correctly. For example, if a population of 100 profiles was randomly sorted into the four categories (each category receiving 25 documents), the distribution percentages in Table 6 illustrates how many profiles would likely have been placed in the correct category. To further illustrated this example, consider the Power category; it is expected that 37.5% of the 25 documents would be correctly categorized since 37.5% of the population is know to be Power; therefore, about 9 of the 25 profiles randomly placed into the power category would have been categorized correctly. By adding up the number of documents that are likely to be correctly categorized for each category (~9 for Power, ~7 for Personal Standard, ~5 for Social, ~4 for social), the percentage of documents likely to be categorized correctly via guessing would be 25. Therefore, the expected accuracy of randomly placing the documents into one of the four categories is 25%. In order for a software tool to show a better performance than guessing, it must have an overall accuracy that is statistically more significant than 25% within a 95% confidence interval (statistical significance and confidence intervals are discussed in subsequent sections).

### 3. The Training/Evaluation Sets

From the 256 categorized documents, twenty profiles per category were needed to train the software packages. Thus 60 profiles (20 each for Power, Personal Standards, and Social) were randomly selected to represent the predefined categories. Each software

[71] McCabe, George P. and David S. Moore. *Introduction to the Practice of Statistics, 4th Ed.* W. H. Freeman and Company, 2002.

package used the same 60 profiles as training documents in their supervised learning systems. Because the Other category had no definitive characteristics, it was not trained as a specific category. This technique was in-line with the existing features of the software tools; each tool had the ability to build an additional category for documents that did not exhibit the attributes within the predefined categories established by the user.

The remaining 196 manually categorized documents were then randomly segregated into two evaluation sets. The first set consisted of 97 profiles and the second set consisted of 99 profiles.[72] Once a software package had been trained using the 60 training profiles, an evaluation set was then entered into the tool. The system was then manipulated using the Graphical User Interface (GUI) to automatically sort the evaluation data set. Each evaluation data set was categorized independently of one another. Figure 11 shows the distribution of how the 256 manually categorized profiles were divided into training documents and evaluation sets.

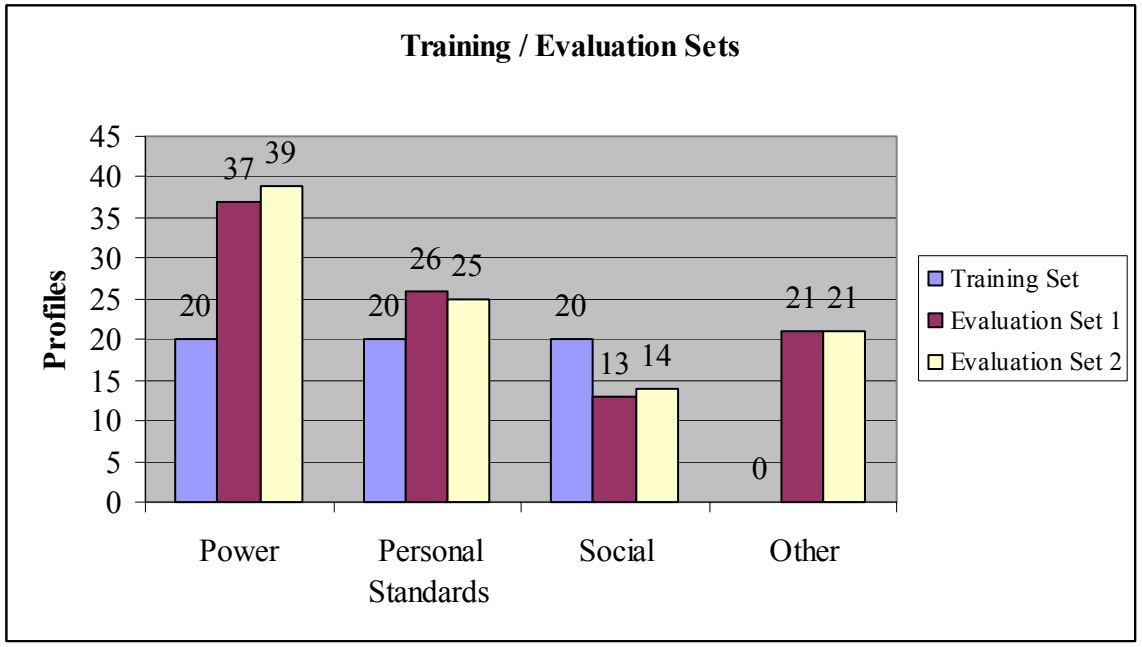


Figure 11. Training/Evaluation Sets

[72] There is no specific reason why the evolution sets did not contain an equal number of profiles. The only concern was that there would be enough social profiles (the minority category) in each set. The remaining profiles were randomly assigned to an evaluation set.

After the categorization was complete, the evaluation documents were counted to determine how many documents were placed into each group, by the software tool, and whether or not each document was categorized correctly (see Appendix A). The accuracies of the tools were then determined through statistical analysis.

## C. EVALUATION METHODS USED

In order to evaluate each software tool fairly, an assessment method was generated that would examine the results from two separate perspectives with two separate trials. The first perspective evaluated the *goodness of fit* between the actual results generated by the social psychologist and the specific software package. The 'goodness of fit' among proportions was a comparison of the produced distributions; this was considered the first level of confidence. An example of this first level confidence can be shown by email messages. If a person normally received between 50 and 70 email messages per day, but one day received only 5 messages, that person would suspect that there was a problem with the email system. The same is true for the distribution of profiles within the IT tools; if they cannot produce similar distributions to the manual human effort, it can be suspected that those IT systems are not performing the task in a similar manner. The second perspective evaluated the *overall accuracy* of the IT results versus the true results obtained from the social psychologist; this was considered the second level of confidence. This was the proportion of the total number of correctly categorized documents divided by the total number of documents in the evaluation set. This proportion was the accuracy of the IT system.

The null hypothesis for the level one confidence was that the distribution produced by a human would not be different from the distribution produced by an IT system. The alternate hypothesis was that the distribution and the method that produced the distribution (human or IT system) were not independent. This type of statistical evaluation compared an observed value to an expected value for which a chi-square test[73] $\chi^2 = \Sigma \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$ is specifically designed to measure. The result of

[73] McCabe, George P. and David S. Moore. *Introduction to the Practice of Statistics, 4th Ed.* W. H. Freeman and Company, 2002.

the chi-square test, does not identify causation or justification, but rather provides evidence to reject, or fail to reject, the null hypothesis based upon the chi-square test statistic produced.

The second level of confidence derived the accuracy of each software tool evaluated. Accuracy was derived by the comparison of the actual categorization of a profile verses the placement done by the software system. For each software system, accuracy was captured both by category and overall. The method for determining accuracy for a category was to calculate the number of correct placements within the category and divide them by the total number of documents placed into that category. The same method was used to determine overall accuracy; divide the total correct placements in all categories by the total of all documents in the evaluation set.

To determine the most accurate IT system, each tool was compared against each other to determine if there was a statistically significant difference between them. Thus, for each of these comparisons a separate statistical test was performed. In order to compare the two populations (IT system vs. the social psychologist) the difference in the two populations was used. $D = \hat{\rho}_1 - \hat{\rho}_2$ Because the sample size for the evaluation were sufficiently large (>30) it could be assumed that the sampling distribution of the difference between the two distributions was approximately normal. In order to prove that the differences were or were not statistically significant, confidence intervals were created. If the confidence interval included zero (because the distributions could be assumed to be normal with a mean of zero and a standard deviation of one) then it was proven not to be statistically significant; simply put, the two systems that were compared performed approximately equal. The confidence intervals were generated with the following formulas.

$$\left(\hat{\rho}_1 - \hat{\rho}_2\right) \pm z \times SE_D$$

$$SE_D = \sqrt{\frac{\hat{\rho}_1\left(1-\hat{\rho}_1\right)}{n_1} + \frac{\hat{\rho}_2\left(1-\hat{\rho}_2\right)}{n_2}}$$

In addition to comparing the two proportions by providing confidence intervals, it was useful to test the null hypothesis that the two population proportions were equal. The alternate hypothesis was that the two proportions were not equal.

$$H_0 : \rho_1 = \rho_2$$

$$H_a : \rho_1 \neq \rho_2$$

In order to test this hypothesis the pooled estimate, pooled standard error and z statistic were computed.[74]

$$\hat{\rho} = \frac{\text{number of successes in both samples}}{\text{number of observations in both samples}} = \frac{X_1 + X_2}{n_1 + n_2}$$

$$SE_{D_p} = \sqrt{\hat{\rho}\left(1 - \hat{\rho}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$z = \frac{\hat{\rho}_1 - \hat{\rho}_2}{SE_{D_p}}$$

Once the z statistic was computed it was compared to the critical z (derived from the desired level of confidence) and if it was above or below the positive and negative critical z the null hypothesis was rejected. Otherwise the null hypothesis could not be rejected. This test showed whether or not the software tools provided approximately the same level of accuracy or that there was a statistically significant difference between them.

Once each tool was compared against each other the same statistical tests were used to compare each tool to the overall accuracy of guessing. Guessing was defined as the outcome of a correct placement of a profile into its proper category based on randomness. The overall proportion of correct placements with guessing was computed through probability and reinforced through simulation.

From the manual categorization, the overall proportions of profiles were: .375 power, .277 personal standards, .184 social, and .164 other (Table 6). The probability of a profile being randomly placed into any one category had a probability of .25. By using

[74] Ibid., pp. 604-605.

these probabilities through simulation, it was determined that the overall number of successes in one hundred trials would be approximately twenty-five. Therefore, the expected accuracy of guessing based on one hundred profiles and four categories would be 25%. The comparison between each software tool and guessing provided the needed reinforcement that each application was either legitimately categorizing documents or simply providing its guess as to where profiles should be placed.

## D. SOFTWARE SYSTEM RESULTS

The results for each tool are summarized in Tables 7, 8, 9, and 10. For both trials all the software applications failed the level I confidence evaluation using the chi-square test (see Figures 12, 13, 14, 15). The failure of level one confidence shows that the IT systems were not performing the task of categorization in the same manner as the social psychologist. Similar to the email example discussed above, there was a statistically significant difference between what was expected and what was observed. This shows that the IT systems and the social psychologist perform differently but does not answer the question of how accurate the IT system performed. Individual system accuracy is addressed by level II confidence.

Level II confidence is a measure of accuracy of the IT system; it is the proportion of profiles that were accurately segmented into their correct categories. The proportion successful was defined as the total number of correctly placed profiles in the trial divided by the total number of profiles in the trial. Level II confidence also provides the metric to determine if an IT system performed better than guessing. Findings are summarized in Appendix A.

Enterprise Miner (EM) was one of the five software applications that was to be evaluated. Unfortunately, EM was unable to produce measurable results that could be compared to the other applications. The shortfall in EM was its inability to categorize the specified profiles in a supervised learning manner. EM was capable of conducting unsupervised learning, but this would provide the system autonomy to categorize based on its own determination of valuable attributes. EM does possess value in this research domain, but did not produce measurable outputs for comparison with other software tools. The capabilities that EM possesses (due to it foundations in data mining) is

desirable for further research, but in the application of text mining only, it provides unsupervised learning and concept extraction which is short of the goal of automated categorization based on learning sets.

| **Software Package** | **Version/Release** | **Settings** | **Results** | |
|---|---|---|---|---|
| | | | Level I Confidence | Level II Confidence (Proportion Successful) |
| ReelTwo | Classification System 2.4.6 | Taxonomy depth: full<br>Exclude empty directory: No<br>Subsumption: No<br>Check for duplicates: No<br>Split Documents: No | Reject null hypothesis | .3299 |
| SER Globalbrain | Personal Edition 1.7.0 | Minimum pattern size: 3<br>Maximum dictionary size: 100,000<br>Include numbers: Yes<br>Absolute relevance value: 51<br>Relative distance value: 1 | Reject null hypothesis | .2990 |
| Lexiquest Categorize | Lexiquest Taxonomy Manager for Categorization 1.5 | Minimum number of categories per document: 1<br>Maximum number of categories/document: 3<br>Minimum confidence level for single response: 1 | Reject null hypothesis | .2990 |
| PolyAnalyst 4.6 | 4.6.498 | Key sense type: All<br>Find collocations: No<br>Taxonomy[75] | Reject null hypothesis | .3608 |

Table 7. Trial I Results

The null hypothesis for level I confidence (there is no difference between distributions of the categories and the method that produced them) was rejected for all four software tools. Figure 12 explores the actual distributions for the categories across

[75] PolyAnalyst was not capable of performing supervising learning without user defined taxonomies. For each category, descriptors (words) had to be inputted that defined to the software what to look for and match. The system would accept individual words, phrases, synonyms, and root forms of words. For the specified taxonomy. See Appendix B for all descriptions.

the software tools in order to determine which category (or categories) caused the null hypothesis to be rejected. In the level II confidence the tools overall accuracy was determined.

In viewing the actual distributions in Figure 12 it can be seen that the categories of power and personal standards are not significantly different than what would be expected. However, the categories of social and other are significantly different than what would be expected. This provides evidence that the category of social caused the null hypothesis to be rejected by all the software tools.

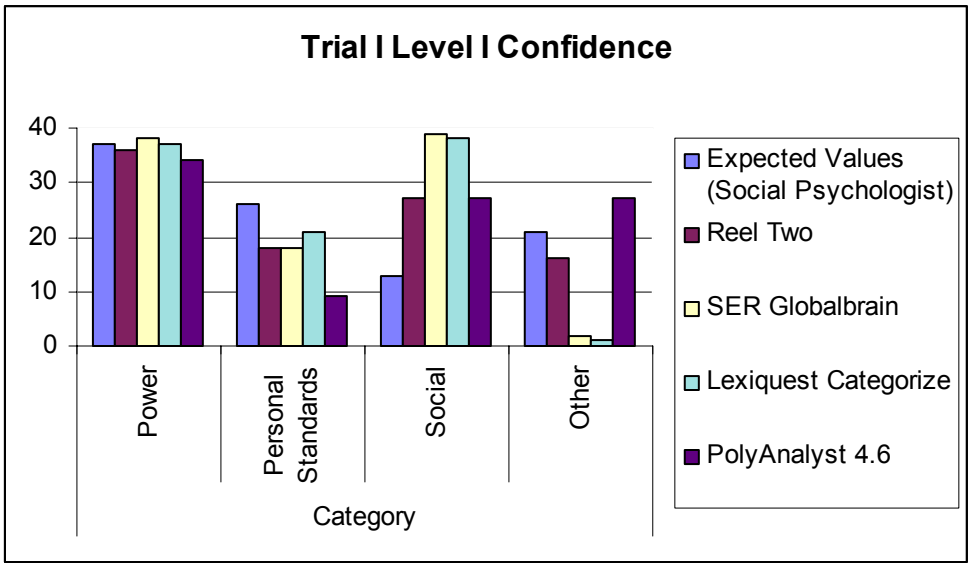


Figure 12. Trial I Level I Confidence

| **Software Package** | **Version/Release** | **Settings** | **Results** | |
|---|---|---|---|---|
| | | | Level I Confidence | Level II Confidence (Proportion Successful) |
| ReelTwo | Classification System 2.4.6 | Same as trial I | Reject null hypothesis | .3535 |
| SER Globalbrain | Personal Edition 1.7.0 | Same as trial I | Reject null hypothesis | .3030 |
| Lexiquest Categorize | Lexiquest Taxonomy Manager for Categorization 1.5 | Same as trial I | Reject null hypothesis | .3232 |
| PolyAnalyst 4.6 | 4.6.498 | Same as trial I | Reject null hypothesis | .3232 |

Table 8. Trial II Results

For the second trial the null hypothesis was rejected for all tools again. Figure 13 provides the detailed breakdown of the distribution for each application. Again the social category caused the rejection of the null hypothesis. Level II confidence was not significantly different than the results attained in the first trial.

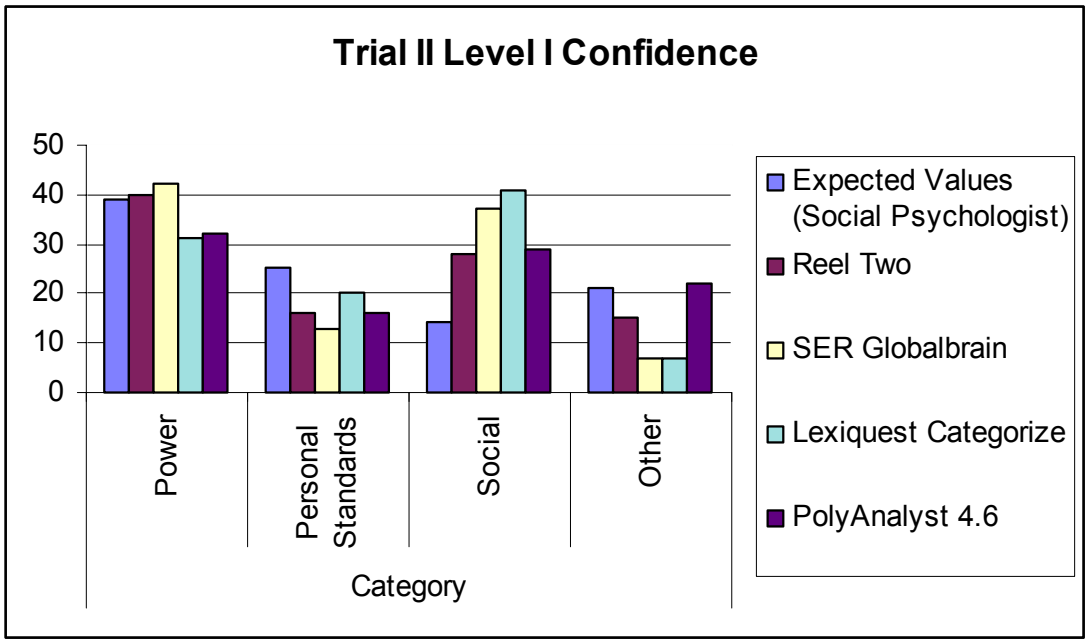


Figure 13. Trial II Level I Confidence

At the completion of two independent trials, that both rejected the null hypothesis for level I confidence, further exploration of the data was conducted. The social category was creating the most difficulty for the software tools. A new hypothesis to explain why the social category was problematic pointed towards the concepts (words) in the social profiles. The concepts in a social profile are similar to those of other categories when simply comparing frequency. It is believed that each application was text mining the documents and extracting concepts independent of their context or overall document sentiment. A social profile could possess many words that are similar to the key words in other categories. For example, the words power or aggressive could be used in many profiles that span the categories, but it is the context in which they are used that determines the sentiment of the profile and category in which it belongs.

In an attempt to simplify the task and to continue to evaluate the accuracy of the software, a collective category was created that encompassed both the social and the 'other' category. Results for this evaluation with the same data sets and trials are described in Tables 9 and 10, and Figures 14 and 15.

| **Software Package** | **Version/Release** | **Settings** | **Results** | |
|---|---|---|---|---|
| | | | Level I Confidence | Level II Confidence (Proportion Successful) |
| ReelTwo | Classification System 2.4.6 | Taxonomy depth: full<br>Exclude empty directory: No<br>Subsumption: No<br>Check for duplicates: No<br>Split Documents: No | Reject null hypothesis | .3918 |
| SER Globalbrain | Personal Edition 1.7.0 | Minimum pattern size: 3<br>Maximum dictionary size: 100,000<br>Include numbers: Yes<br>Absolute relevance value: 51<br>Relative distance value: 1 | Reject null hypothesis | .3608 |
| Lexiquest Categorize | Lexiquest Taxonomy Manager for Categorization 1.5 | Minimum number of categories per document: 1<br>Maximum number of categories/document: 2<br>Minimum confidence level for single response: 1 | Reject null hypothesis | .3918 |
| PolyAnalyst 4.6 | 4.6.498 | Key sense type: All<br>Find collocations: No<br>Taxonomy[76] | Reject null hypothesis | .4742 |

Table 9. Trial I Results with Collective Category

For the first trial with the collective category the null hypothesis was rejected for all tools again. Figure 14 provides the detailed breakdown of the distribution for each application. Here it is no longer clear which category is responsible for the rejection of the null hypothesis. Level II confidence provided higher accuracies, but it was over fewer categories.

[76] PolyAnalyst was not capable of performing supervising learning without user defined taxonomies. For each category descriptors (words) had to inputted that defined to the software what to look for and match. The system would accept individual words, phrases, synonyms, and root forms of words. For the specified taxonomy. See Appendix B for descriptions of power and personal standards.

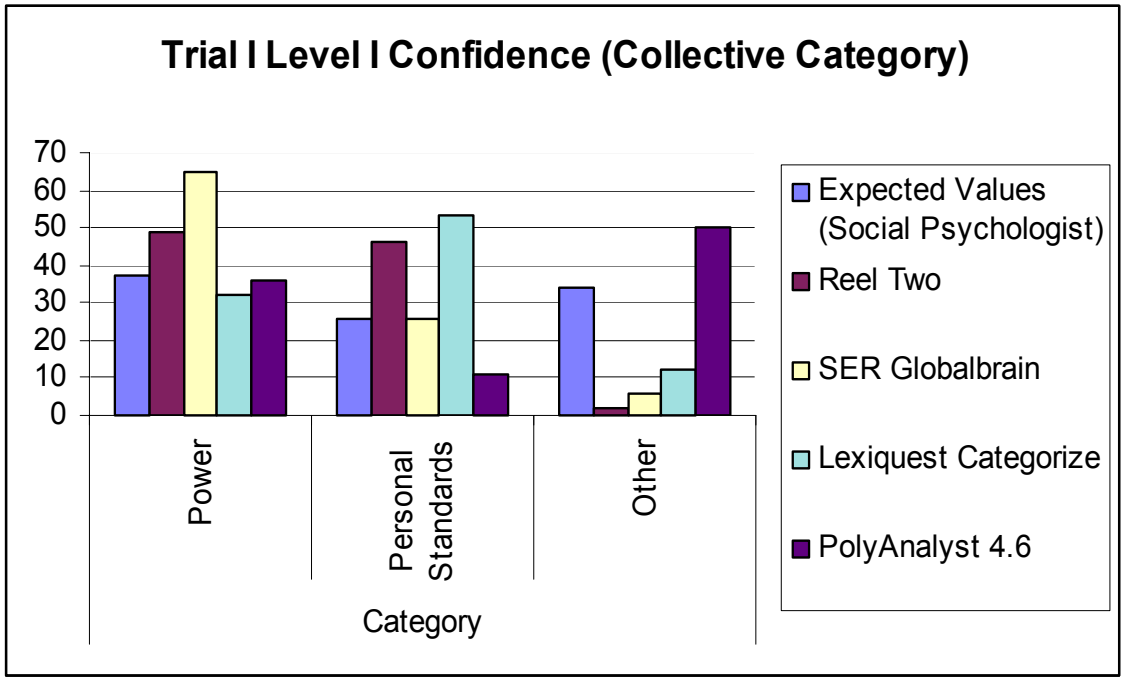


Figure 14. Trial I Level I Confidence (Collective Category)

| Software Package | Version/Release | Settings | Results | |
|---|---|---|---|---|
| | | | Level I Confidence | Level II Confidence (Proportion Successful) |
| ReelTwo | Classification System 2.4.6 | Same as trial I with collective category | Reject null hypothesis | .4242 |
| SER Globalbrain | Personal Edition 1.7.0 | Same as trial I with collective category | Reject null hypothesis | .3636 |
| Lexiquest Categorize | Lexiquest Taxonomy Manager for Categorization 1.5 | Same as trial I with collective category | Reject null hypothesis | .3838 |
| PolyAnalyst 4.6 | 4.6.498 | Same as trial I with collective category | Reject null hypothesis | .3939 |

Table 10. Trial II Results with Collective Category

For the second trial with a collective category the null hypothesis was rejected for all tools again. Figure 15 provides the detailed breakdown of the distribution for each application. It remains unclear where the cause for the rejection of the null hypothesis

lies. Level II confidence was similar to the results attained in the first trial. All the level II confidence proportions were not a significant improvement given that there was a reduced amount of categories.

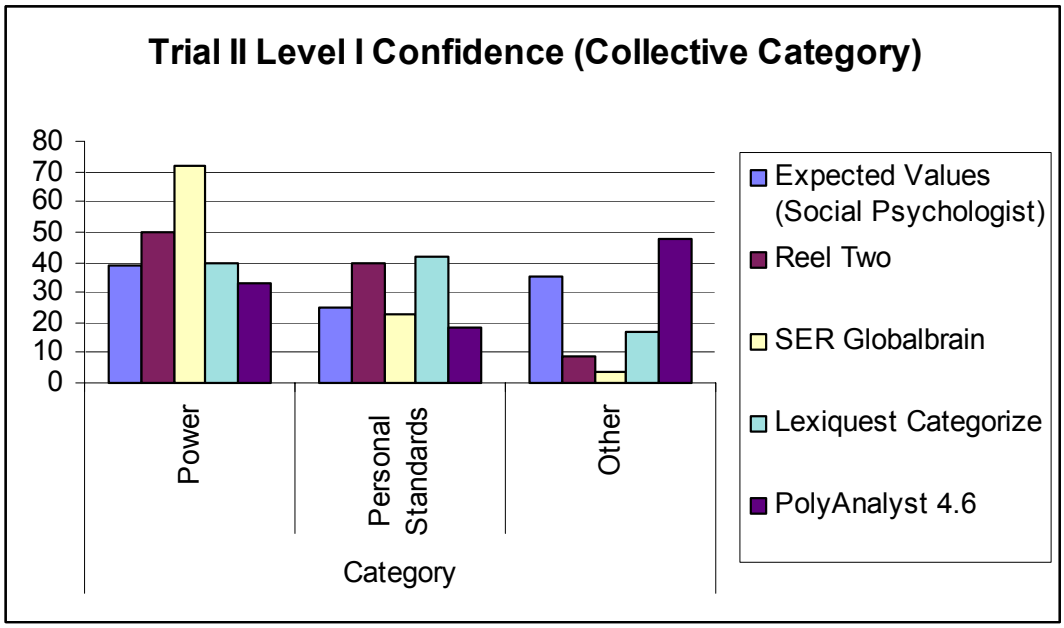


Figure 15. Trial II Level I Confidence (Collective Category)

Through these results it was concluded that even though none of the software tools met level one confidence, they all improved in their overall accuracy of placing profiles into their correct categories. The improved accuracy with reduced categories can not be compared directly with prior testing but must be evaluated separately versus the ability to randomly guess the correct category for any profile.

## E. COMPARISON OF RESULTS ACROSS SOFTWARE APPLICATIONS

In order to determine which of the software tools performed the best, individual comparisons were conducted between each tool and guessing. The results of the comparison are graphically represented in Figure 16.

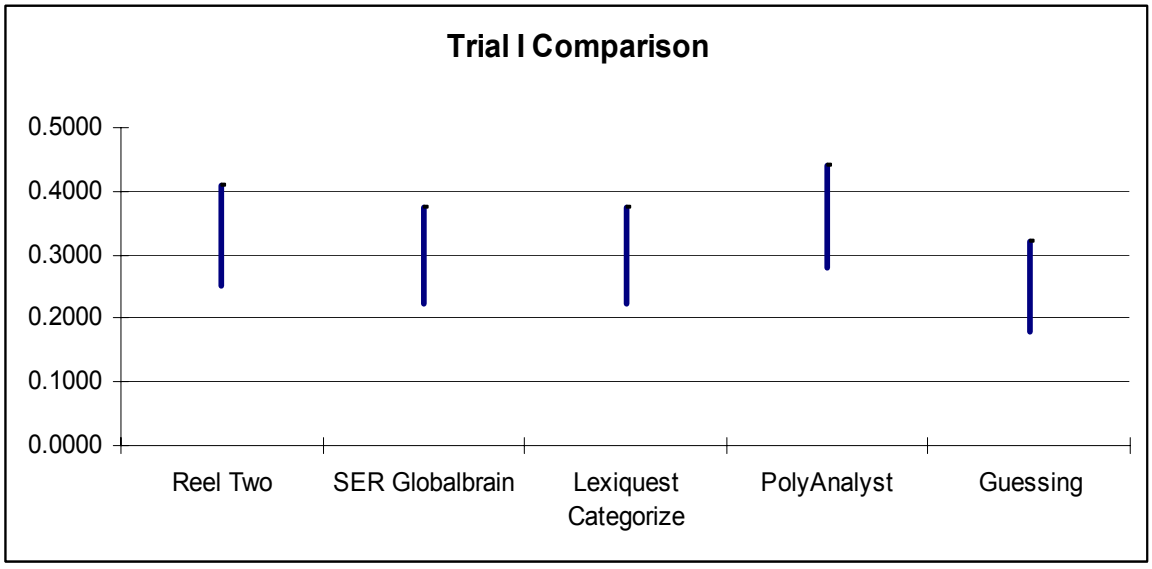


Figure 16. Trial I Accuracy Comparison

Figure 16 represents the successful proportions of each software application along with its 95% confidence interval. When each tool was evaluated against one another using a significance test described earlier in the chapter, it was determined that none of the software tools were significantly different than any other. Additionally, there was no statistical significance between guessing and any of the software tools with the exception of PolyAnalyst.

On the second trial, all the software applications performed approximately the same with ReelTwo performing the best. The same significance test was conducted and they all failed to be significantly different to include guessing. See Figure 17 for the results.

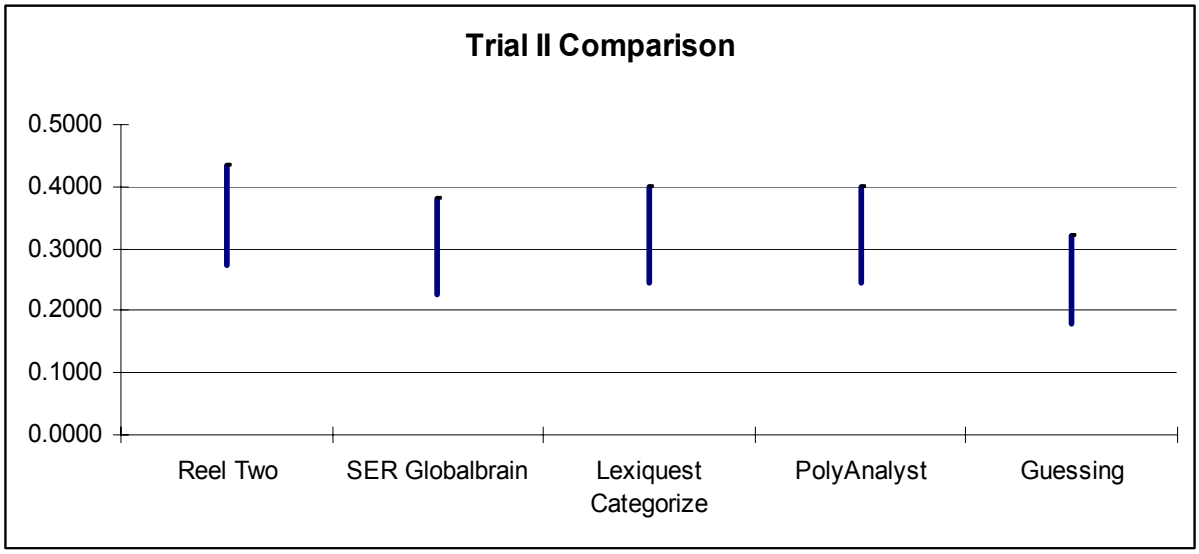


Figure 17. Trial II Comparison

When the software applications had only two categories to learn from, the overall proportion of accuracies were improved, but in comparison to each other they were not significantly different. Additionally, the only application that performed better than guessing and was statically significant was PolyAnalyst. Figure 18 show the results of trial I.

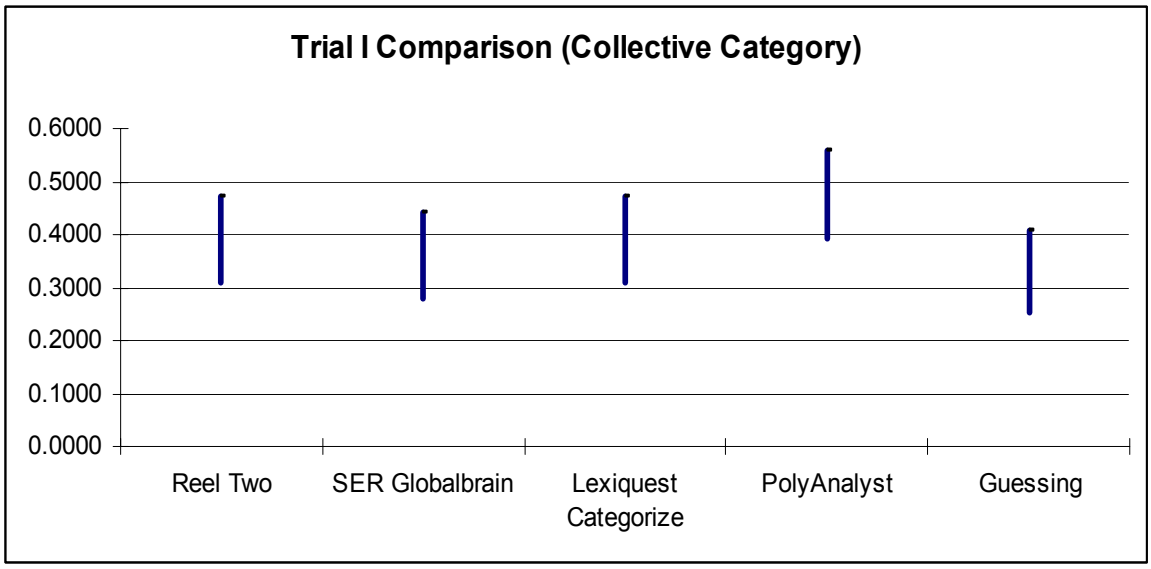


Figure 18. Trial I Comparison (Collective Category)

The trial II results can be seen in Figure 19. In trial II all the software tools performed approximately the same without any statistical significance among them. For trail II all tools performed on par with guessing.

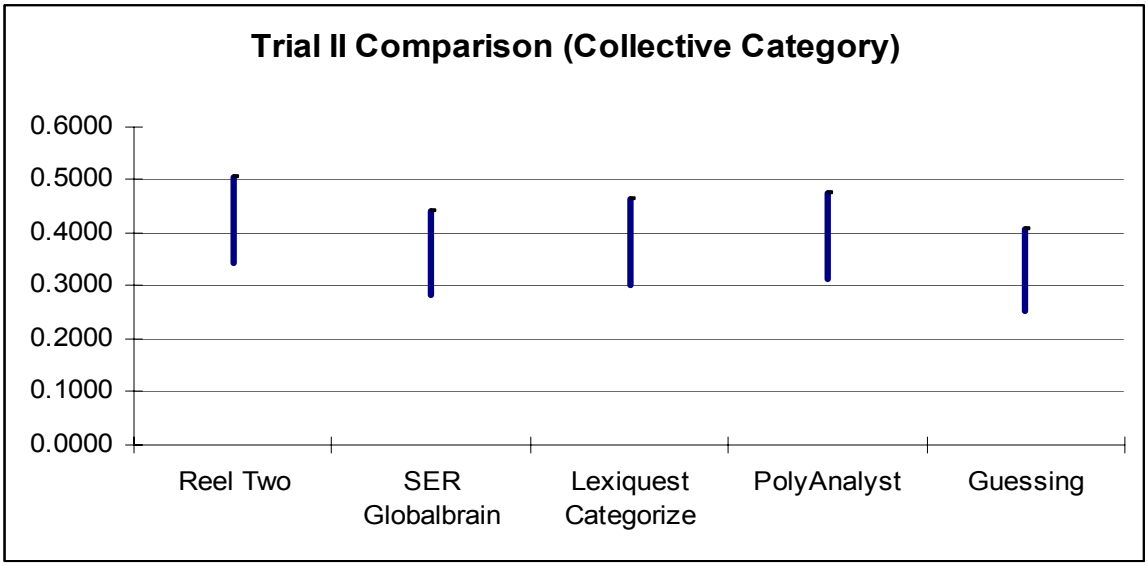


Figure 19. Trial II Comparison (Collective Category)

THIS PAGE INTENTIONALLY LEFT BLANK

# V. SUMMARY, CONCLUSIONS, AND FUTURE RESEARCH

## A. RESEARCH FINDINGS

Data and text mining are widely used within the commercial world. The technology of mining data has become fairly mature and thus additional applications have been employed within the market place as well as in government sectors. The value of this technology is in the development of more informed decisions based on discovered patterns that were previously lost in the overwhelming size of the data. The capability to analyze data from any conceivable angle and mold it into actionable information is desired for any organization that wants to survive in the information age. IT currently enables categorization based on any number of variables to include text, but the accuracy of those results is dependent on the techniques used.

Text mining is not a simplistic task even with seemingly uncomplicated methods such as word matching. In fact, text mining is an extremely complex task for computers to undertake especially from the perspective of reproducing the actions of a human mind. To ask IT to read a document, produce a condensed result based on the essence of the document, and then place it into a user defined category is daunting. Many methods have been developed to help IT systems produce these kinds of results, but the accuracy is always a factor of the inputs as well as the capability of the tool. No matter what the method of categorization, certain attributes have significant impact on the abilities of the IT systems whether it is the learning data set for a neural network or the specified rules for a rule-based approach. It is intuitive that the best results would be generated by a combination of systems, but the simple definition of 'best results' may be significantly short of its intended goal.

This research attempted to define whether or not IT systems could categorize psychological profiles into defined categories; the results were bleak. None of the software solutions passed the first level of confidence because none of them gave a distribution that was close enough to the true composition of the population. For the second level of confidence, the comparison between the IT solutions and the probabilities of guessing also showed that the IT solutions did not perform as desired. Although, in

trial one, PolyAnalyst did yield a statistically significant more accurate result than guessing, none of the tools were seen to have a statistical difference between themselves and PolyAnalyst. Thus, even though PolyAnalyst showed greater accuracy than guessing, with 95% confidence, it did not perform statistically better than the other tools. Additionally, PolyAnalyst failed to show any statistical difference against guessing in Trial II or the testing of Trial I and Trial II with only two categories.

It was determined that no statistical evidence existed to prove a difference between the tools and guessing with the exception of PolyAnalyst in trial one. However, even though no statistical difference existed, it is still evident that every tool performed better that guessing when looking at the raw results (Appendix A). In essence, the tools perform the same as guessing from a statistical standpoint, but they consistently show slightly better accuracy than guessing when looking at the actual accuracies of each tool. Therefore, the tools performed some operation that allowed them to be slightly better than guessing on every trial. This slight advantage appeared to come from simple word matching.

Because every software solution that was tested was proprietary in nature, each tool was essentially a black box. It was not possible to perform a detailed analysis to determine the reasons for their poor performance. However, by viewing the results of each tool, it became evident that the tools were looking for words (representing concepts) to match the profiles with the learned categories. Given the haphazard format of the profiles and the limitless use of descriptors within them, the tools had difficultly discerning to which category a particular profile belonged.

It was identified that the IT systems could extract concepts from the profiles, but it was how the concepts were being used that potentially was causing problems. The desire was for the IT system to use the extracted concepts in context and develop an accurate sentiment for the overall profile. Webster's dictionary defines sentiment as: the emotional significance of a passage or expression as distinguished from its verbal context.[77] This was what was really being asked of an IT system: the production of a sentiment for the purposes of classification.

[77] Merriam-WebsterOnlineDictionary. http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=sentiment. August 2004.

The results of the research provide evidence that sentiments were not being created and therefore were not being used for the classifying task. Most of the software applications provided the ability and specifically mentioned that the user can view the concepts that were extracted and modify their importance in the overall task of classification; however, this demonstrates the concept of word matching and not the theme of the individual texts. If concepts are extracted and used independent of the context in which they apply, it is understandable that a system would not know the theme of a sentence, paragraph, or profile.

What is not occurring within the systems can be explained by the old axiom of "the whole is greater than the sum of its parts." The IT systems are very good at finding the numbers to sum but once the numbers are removed from their context, the value diminished and they no longer summed to the actual value of the document. An example of this concept could be equated to a standard pneumatic office chair. If the chair was not assembled and was just a pile of parts on the floor, the IT system would identify all the parts and probably provide a level of detail beyond that of a regular human, but it would not know that it was a chair. Conversely, humans could easily view the parts and constructively assemble the parts in their head and come to the conclusion that it was a chair, only not assembled yet. The human probably would not be able to identify some of the parts, nor provide specifications of parts, but that would not hamper the final outcome: the discernment that the pieces make a chair. IT systems are not only being asked to determine that the pieces can form a chair but are being asked "What type of chair is it?" Is it an armchair, conversation chair, dining chair, office chair, etc...? This is a daunting task for a software tool.

When working with text only, it is difficult to segment entire documents based on extracted concepts that do not take into consideration how the concepts were used. If weights (levels of importance) are assigned to each concept, the concepts are then treated independent of context. The research focused on software tools that conducted concept extraction because that is the basic building block of text mining; it is also the cutting edge of text mining technology. To continue to capitalize on the capabilities of text mining, further development needs to occur in the area of sentiment generation based on extracted concepts.

## B. FURTHER RESEARCH

Current IT solutions are not yet robust enough to perform automated categorization based on the parameters defined in this research. This leaves two options available if the desire remains to create an automated psychological characterization tool: 1) wait until the software tools for commercial use (although intended for different purposes) become powerful enough to perform the task, or 2) embrace the strengths that exist within the current tools. Since waiting for technology enhancements is not an attractive option, the alternate method requires exploration. In using the strengths of the tools, the rule-based method in conjunction with data mining becomes the most promising approach for future research.

PolyAnalyst could not perform automated characterization under a supervised environment based on training documents. However, it was capable of performing automated characterization based on user-defined rule sets (Appendix B). By using rule sets, the software was able to match words, phrases, and concepts that were defined by the user rather than relying on its own extracted concepts. In this regard, the system was given less responsibility for recognizing which category a profile fit into and was able to instead match words, phrases, and concepts that it determined were similar. This type of approach produced the only test results that were statistically more significant than guessing.

Rule-based methods were discussed in detail in Chapter III. However, a distinction must be made with regard to how the rule sets are created. In order to maximize the IT system strengths, robust rule sets must be created manually with inputs from expert personnel. Although it is possible to have the software solutions create the rule sets, this would not improve accuracy since a weak rule set (such as a rule based on an insignificant attribute) is thought to be blamed for the poor performance of the IT systems in this research. The concepts extracted by the IT systems could be used by the users to form the rule sets, but the system must not be tasked with the combination of concepts to form sentiment or context.

Rather than having psychologists analyze hundreds of profiles and manually categorize them, their efforts would focus on the creation of rule sets that contain key phrases, words, and concepts that would then be fed into the systems to conduct the classification of the uncategorized documents. A limited number of profiles would still require analysis in order to determine the categories present in a population, but this would be a significant departure from current methods. An example of this is described in Chapter IV.

Once the rule set is created, it can be used with multiple tools. Although PolyAnalyst mandated the use of rule sets during the testing phase, it is not the only tool that can perform in such a manner. It is not suggested that PolyAnalyst is the best solution with regard to a rule based approach. Rather, it is one of several solutions that need to be further researched for their ability in a rule-based environment.

In order to use data mining, additional variables would be needed in order to compliment the raw text that currently exists. This would require an additional field to current style that is used to report the data. Instead of reporting the information about a person in a written summary alone, questions would be added to the documents to provide additional variables. More variables would enable data mining vice text mining alone. With the advent of additional variables, where the data becomes multifaceted, multiple angles of analysis become possible. The additional variables could be input via scalable ordinal results, true/false or yes/no answers, and multiple-choice selections. These questions and answers would be added to the text documents that already exist. Thus, the only change that would need to occur in the creation of the profiles would include a questionnaire regarding each person that would be added to the end of the document. No additional data collection would need to be performed; rather, the data would be used in multiple ways and provide the additional robustness needed for automated categorization.

If this option is to be attempted, a series of questions needs to be developed. These questions should be created in conjunction with professional analysts that can assign proper precedence and weighting to the questions. A starting point for the questionnaire was developed and can be found in Appendix C. By using the feedback

from such questions (which would generate multiple variables) and supplementing textual write-ups, more concrete information can be processed by an IT system. These types of applications would work well for information discovery, character/individual clustering, and categorical prediction. Although the inputs for the questions would need to change depending on the desired categorization, the outcomes would be based on rules.

Along with further research, vendor support should be sought and exploited. Every vendor of the five software solutions tested in this project was instrumental in making the tool available, ensuring its functionality, and answering questions on technical issues. Some vendors were willing to discuss design changes and provide face-to-face support, but this was not an option due to the sensitivity of the raw data. Future success is heightened by vendors that are willing to design a product around the specific requirements associated with automatic psychological characterization. Although they would be drawing on the core engines of their products, a simplified user interface and additional functionality for the automation of profiles based on rule sets must be explored. One particular tool that deserves further research is a rule-based method developed by ClearForest Corporation; ClearForest is a rule-based tool that requires substantial investment and training time.

## C. FINAL COMMENTS

Limited resources will always present challenges to organizations. By leveraging IT, human burdens can be diminished with acceptable results. This research focused on a limited resource involving individual psychological categorization. It was found that IT solutions are not currently able to produce results that would enable automatic categorization of individual profiles based on supervised linguistic processing. However, significant findings show potential in further research that employs rule-based approaches to this problem.

Future IT systems will continue to increase in functionality. As these systems are developed, interaction with vendors will allow tools to be designed with automatic

categorization as a primary concern. Future efforts should be focused on leveraging the advantages these systems currently contain while working with vendors to produce a tailored approach to the automatic categorization.

THIS PAGE INTENTIONALLY LEFT BLANK

# APPENDIX A. SOFTWARE SYSTEM ACCURACY RESULTS

## A. DISCUSSION

This appendix provides the individual details of the result produced for each software application. In level I confidence, the null and alternate hypotheses are clearly articulated below in section B of this appendix. Level I provides a 'goodness of fit'; a method that can be used to quickly see whether or not things look correct. The true value of the results lies within the level II confidence. The level II results provide individual proportions of accuracy by category and software application. Since the research addressed the capabilities of IT to categorize correctly, the individual category results were not as important as the cumulative results of the tool itself. Individual accuracies for the predominant categories did produce the highest proportion correct, but this result was expected due to the conclusions discussed in Chapter V. By utilizing the results produced with the cumulative category (Figures 27 and 30) it is shown that individual accuracies among the predominant categories (power and personal standards) did not improve by removing the troubled category, social.

## B. TRIAL I DATA

### 1. Level I Confidence

$H_0$: There is no association between the column and row variables (Method / Category). This can also be stated that the distributions of the categories are the same regardless of the method that produced them.

$H_a$: There is an association between the row and column variables.

| | Method | Category | | | | |
|---|---|---|---|---|---|---|
| | | Power | Personal Standards | Social | Other | |
| Reel Two | Expected Values (Social Psychologist) | 37 | 26 | 13 | 21 | 97 |
| | Observed Value | **36** | **18** | **27** | **16** | 97 |
| | Chi Square Test Stat | 18.75478925 | | | | |
| | P-value | 0.000307072 | | **REJECT Null Hypothesis** | | |

| | Method | Category | | | | |
|---|---|---|---|---|---|---|
| | | Power | Personal Standards | Social | Other | |
| SER Globalbrain | Expected Values (Social Psychologist) | 37 | 26 | 13 | 21 | 97 |
| | Observed Value | **38** | **18** | **39** | **2** | 97 |
| | Chi Square Test Stat | 71.67904168 | | | | |
| | P-value | 1.86495E-15 | | **REJECT Null Hypothesis** | | |

| | Method | Category | | | | |
|---|---|---|---|---|---|---|
| | | Power | Personal Standards | Social | Other | |
| Lexiquest Categorize | Expected Values (Social Psychologist) | 37 | 26 | 13 | 21 | 97 |
| | Observed Value | **37** | **21** | **38** | **1** | 97 |
| | Chi Square Test Stat | 68.08608059 | | | | |
| | P-value | 1.0965E-14 | | **REJECT Null Hypothesis** | | |

| | Method | Category | | | | |
|---|---|---|---|---|---|---|
| | | Power | Personal Standards | Social | Other | |
| PolyAnalyst | Expected Values (Social Psychologist) | 37 | 26 | 13 | 21 | 97 |
| | Observed Value | **34** | **9** | **27** | **27** | 97 |
| | Chi Square Test Stat | 28.22875977 | | | | |
| | P-value | 3.3783E-06 | | **REJECT Null Hypothesis** | | |

Figure 20. Trial I Level I Confidence Data

## 2. Level II Confidence

| Method | Category | Count | Count of Successes | Proportion Successful | Standard Error | 95% Confidence Interval | | |
|---|---|---|---|---|---|---|---|---|
| Reel Two | Power | 36 | **18** | 0.5000 | 0.0833 | 0.3629 | - | 0.6371 |
| | Personal Standards | 18 | **6** | 0.3333 | 0.1111 | 0.1506 | - | 0.5161 |
| | Social | 27 | **4** | 0.1481 | 0.0684 | 0.0357 | - | 0.2606 |
| | Other | 16 | **4** | 0.2500 | 0.1083 | 0.0719 | - | 0.4281 |
| | Total (n) | 97 | 32 | **0.3299** | 0.0477 | **0.2514** | **-** | **0.4084** |
| SER Globalbrain | Power | 38 | **16** | 0.4211 | 0.0801 | 0.2893 | - | 0.5528 |
| | Personal Standards | 18 | **8** | 0.4444 | 0.1171 | 0.2518 | - | 0.6371 |
| | Social | 39 | **5** | 0.1282 | 0.0535 | 0.0401 | - | 0.2163 |
| | Other | 2 | **0** | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 |
| | Total (n) | 97 | 29 | **0.2990** | 0.0465 | **0.2225** | **-** | **0.3754** |
| Lexiquest Categorize | Power | 37 | **17** | 0.4595 | 0.0819 | 0.3247 | - | 0.5942 |
| | Personal Standards | 21 | **6** | 0.2857 | 0.0986 | 0.1236 | - | 0.4479 |
| | Social | 38 | **6** | 0.1579 | 0.0592 | 0.0606 | - | 0.2552 |
| | Other | 1 | **0** | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 |
| | Total (n) | 97 | 29 | **0.2990** | 0.0465 | **0.2225** | **-** | **0.3754** |
| PolyAnalyst | Power | 34 | **19** | 0.5588 | 0.0852 | 0.4188 | - | 0.6989 |
| | Personal Standards | 9 | **4** | 0.4444 | 0.1656 | 0.1720 | - | 0.7169 |
| | Social | 27 | **4** | 0.1481 | 0.0684 | 0.0357 | - | 0.2606 |
| | Other | 27 | **8** | 0.2963 | 0.0879 | 0.1518 | - | 0.4408 |
| | Total (n) | 97 | 35 | **0.3608** | 0.0488 | **0.2806** | **-** | **0.4410** |
| Guessing | Total | 100 | 25 | **0.2500** | 0.0433 | **0.1788** | **-** | **0.3212** |

Figure 21. Trial I Level II Confidence Data

Count: The total number of profiles that the software application placed in the category.

Count of Successes: Total correct number of profiles that the software application placed in the category.

## 3. Comparison

$$H_0 : \rho_1 = \rho_2$$

$$H_a : \rho_1 \neq \rho_2$$

| Software Packages | Difference between p<br>z score on difference | SE of difference | Margin of Error | Confidence Interval<br>Judgement |
|---|---|---|---|---|
| Reel Two vs SER Globalbrain | 0.0309 | 0.0666 | 0.1096 | -0.0787 - 0.1405 |
| **Significance** | **z = 0.4639** | 0.0667 | 0.3144 | **Fail to Reject Null Hypothesis** |
| Reel Two vs Lexiquest Categorizer | 0.0309 | 0.0666 | 0.1096 | -0.0787 - 0.1405 |
| **Significance** | **z = 0.4639** | 0.0667 | 0.3144 | **Fail to Reject Null Hypothesis** |
| Reel Two vs PolyAnalyst | -0.0309 | 0.0682 | 0.1122 | -0.1432 - 0.0813 |
| **Significance** | **z = -0.4530** | 0.0683 | 0.3454 | **Fail to Reject Null Hypothesis** |
| Reel Two vs Guessing | 0.0799 | 0.0645 | 0.1060 | -0.0261 0.1859 |
| **Significance** | **z = 1.2364** | 0.0646 | 0.2893 | **Fail to Reject Null Hypothesis** |
| SER Globalbrain vs Lexiquest Categorize | 0.0000 | 0.0657 | 0.1081 | -0.1081 - 0.1081 |
| **Significance** | **z = 0.0000** | 0.0657 | 0.2990 | **Fail to Reject Null Hypothesis** |
| SER Globalbrain vs PolyAnalyst | -0.0619 | 0.0674 | 0.1108 | -0.1727 - 0.0490 |
| **Significance** | **z = -0.9162** | 0.0675 | 0.3299 | **Fail to Reject Null Hypothesis** |
| SER Globalbrain vs Guessing | 0.0490 | 0.0635 | 0.1045 | -0.0555 - 0.1535 |
| **Significance** | **z = 0.7703** | 0.0636 | 0.2741 | **Fail to Reject Null Hypothesis** |
| Lexiquest Categorize vs PolyAnalyst | -0.0619 | 0.0674 | 0.1108 | -0.1727 - 0.0490 |
| **Significance** | **z = -0.9730** | 0.0675 | 0.3299 | **Fail to Reject Null Hypothesis** |
| Lexiquest Categorize vs Guessing | 0.0490 | 0.0635 | 0.1045 | -0.0555 - 0.1535 |
| **Significance** | **z = 0.7703** | 0.0636 | 0.2741 | **Fail to Reject Null Hypothesis** |
| PolyAnalyst vs Guessing | 0.1108 | 0.0652 | 0.1073 | 0.0036 - 0.0036 |
| **Significance** | **z = 1.6897** | 0.0656 | 0.3046 | **Reject Null Hypothesis** |

Figure 22. Trial I Comparison Data

## C. TRIAL II DATA

### 1. Level I Confidence

$H_0$: There is no association between the column and row variables (Method / Category). This is too say that the distributions of the categories are the same regardless of the method that produced it.

$H_a$: There is an association between the row and column variables.

| Reel Two | Method | Category: Power | Personal Standards | Social | Other | |
|---|---|---|---|---|---|---|
| | Expected Values (Social Psychologist) | 39 | 25 | 14 | 21 | 99 |
| | Observed Value | **40** | **16** | **28** | **15** | 99 |
| | Chi Square Test Stat | 18.97959554 | | | | |
| | P-value | 0.000276024 | | | | |
| | | | | **REJECT Null Hypothesis** | | |

| SER Globalbrain | Method | Category: Power | Personal Standards | Social | Other | |
|---|---|---|---|---|---|---|
| | Expected Values (Social Psychologist) | 39 | 25 | 14 | 21 | 99 |
| | Observed Value | **42** | **13** | **37** | **7** | 99 |
| | Chi Square Test Stat | 53.10981685 | | | | |
| | P-value | 1.73714E-11 | | | | |
| | | | | **REJECT Null Hypothesis** | | |

| Lexiquest Categorize | Method | Category: Power | Personal Standards | Social | Other | |
|---|---|---|---|---|---|---|
| | Expected Values (Social Psychologist) | 39 | 25 | 14 | 21 | 99 |
| | Observed Value | **31** | **20** | **41** | **7** | 99 |
| | Chi Square Test Stat | 64.04578755 | | | | |
| | P-value | 8.02506E-14 | | | | |
| | | | | **REJECT Null Hypothesis** | | |

| PolyAnalyst | Method | Category: Power | Personal Standards | Social | Other | |
|---|---|---|---|---|---|---|
| | Expected Values (Social Psychologist) | 39 | 25 | 14 | 21 | 99 |
| | Observed Value | **32** | **16** | **29** | **22** | 99 |
| | Chi Square Test Stat | 20.61286116 | | | | |
| | P-value | 0.00012652 | | | | |
| | | | | **REJECT Null Hypothesis** | | |

Figure 23. Trial II Level I Confidence Data

## 2. Level II Confidence

| Method | Category | Count | Count of Successes | Proportion Successful | Standard Error | 95% Confidence Interval | | |
|---|---|---|---|---|---|---|---|---|
| Reel Two | Power | 40 | **22** | 0.5500 | 0.0787 | 0.4206 | - | 0.6794 |
| | Personal Standards | 16 | **6** | 0.3750 | 0.1210 | 0.1759 | - | 0.5741 |
| | Social | 28 | **4** | 0.1429 | 0.0661 | 0.0341 | - | 0.2516 |
| | Other | 15 | **3** | 0.2000 | 0.1033 | 0.0301 | - | 0.3699 |
| | Total (n) | 99 | 35 | **0.3535** | 0.0480 | **0.2745** | **-** | **0.4326** |
| SER Globalbrain | Power | 42 | **16** | 0.3810 | 0.0749 | 0.2577 | - | 0.5042 |
| | Personal Standards | 13 | **6** | 0.4615 | 0.1383 | 0.2341 | - | 0.6890 |
| | Social | 37 | **7** | 0.1892 | 0.0644 | 0.0833 | - | 0.2951 |
| | Other | 7 | **1** | 0.1429 | 0.1323 | -0.0747 | - | 0.3604 |
| | Total (n) | 99 | 30 | **0.3030** | 0.0462 | **0.2271** | **-** | **0.3790** |
| Lexiquest Categorize | Power | 31 | **14** | 0.4516 | 0.0894 | 0.3046 | - | 0.5986 |
| | Personal Standards | 20 | **10** | 0.5000 | 0.1118 | 0.3161 | - | 0.6839 |
| | Social | 41 | **7** | 0.1707 | 0.0588 | 0.0741 | - | 0.2674 |
| | Other | 7 | **1** | 0.1429 | 0.1323 | -0.0747 | - | 0.3604 |
| | Total (n) | 99 | 32 | **0.3232** | 0.0470 | **0.2459** | **-** | **0.4006** |
| PolyAnalyst | Power | 32 | **14** | 0.4375 | 0.0877 | 0.2933 | - | 0.5817 |
| | Personal Standards | 16 | **4** | 0.2500 | 0.1083 | 0.0719 | - | 0.4281 |
| | Social | 29 | **6** | 0.2069 | 0.0752 | 0.0832 | - | 0.3306 |
| | Other | 22 | **8** | 0.3636 | 0.1026 | 0.1949 | - | 0.5323 |
| | Total (n) | 99 | 32 | **0.3232** | 0.0470 | **0.2459** | **-** | **0.4006** |
| Guessing | Total | 100 | 25 | **0.2500** | 0.0433 | **0.1788** | **-** | **0.3212** |

Figure 24. Trial II Level II Confidence Data

Count: The total number of profiles that the software application placed in the category.

Count of Successes: Total correct number of profiles that the software application placed in the category.

### 3. Comparison

$H_0 : \rho_1 = \rho_2$

$H_a : \rho_1 \neq \rho_2$

| Software Packages | Difference between p<br>z score on difference | SE of difference<br>pooled SE | Margin of Error<br>Z | Confidence Interval<br>Judgement |
|---|---|---|---|---|
| Reel Two vs SER Globalbrain | 0.0505 | 0.0666 | 0.1096 | -0.0591 - 0.1601 |
| **Significance** | **z = 0.7567** | 0.0667 | 0.3283 | Fail to Reject Null Hypothesis |
| Reel Two vs Lexiquest Categorizer | 0.0303 | 0.0672 | 0.1106 | -0.0803 - 0.1409 |
| **Significance** | **z = 0.4506** | 0.0673 | 0.3384 | Fail to Reject Null Hypothesis |
| Reel Two vs PolyAnalyst | 0.0303 | 0.0672 | 0.1106 | -0.0803 - 0.1409 |
| **Significance** | **z = 0.4506** | 0.0673 | 0.3384 | Fail to Reject Null Hypothesis |
| Reel Two vs Guessing | 0.1035 | 0.0647 | 0.1064 | -0.0029 - 0.2099 |
| **Significance** | **z = 1.5913** | 0.0651 | 0.3015 | Fail to Reject Null Hypothesis |
| SER Globalbrain vs Lexiquest Categorizer | -0.0202 | 0.0659 | 0.1084 | -0.1286 - 0.0882 |
| **Significance** | **z = -0.3065** | 0.0659 | 0.3131 | Fail to Reject Null Hypothesis |
| SER Globalbrain vs PolyAnalyst | -0.0202 | 0.0659 | 0.1084 | -0.1286 - 0.0882 |
| **Significance** | **z = -0.3065** | 0.0659 | 0.3131 | Fail to Reject Null Hypothesis |
| SER Globalbrain vs Guessing | 0.0530 | 0.0633 | 0.1041 | -0.0511 - 0.1572 |
| **Significance** | **z = 0.8364** | 0.0634 | 0.2764 | Fail to Reject Null Hypothesis |
| Lexiquest vs PolyAnalyst | 0.0000 | 0.0665 | 0.1093 | -0.1093 - 0.1093 |
| **Significance** | **z = 0.0000** | 0.0665 | 0.3232 | Fail to Reject Null Hypothesis |
| Lexiquest Categorizer vs Guessing | 0.0732 | 0.0639 | 0.1051 | -0.0319 - 0.1784 |
| **Significance** | **z = 1.1425** | 0.0641 | 0.2864 | Fail to Reject Null Hypothesis |
| PolyAnalyst vs Guessing | 0.0732 | 0.0639 | 0.1051 | -0.0319 - -0.0319 |
| **Significance** | **z = 1.1425** | 0.0641 | 0.2864 | Fail to Reject Null Hypothesis |

Figure 25. Trail II Comparison Data

## D. TRIAL I WITH COLLECTIVE CATEGORY DATA

### 1. Level I Confidence

$H_0$ : There is no association between the column and row variables (Method / Category). This is too say that the distributions of the categories are the same regardless of the method that produced it.

$H_a$ : There is an association between the row and column variables.

| | Method | Category | | | |
|---|---|---|---|---|---|
| | | Power | Personal Standards | Other | |
| Reel Two | Expected Values (Social Psychologist) | 37 | 26 | 34 | 97 |
| | Observed Value | **49** | **46** | **2** | 97 |
| | Chi Square Test Stat | 49.39415434 | | | |
| | P-value | 1.88016E-11 | | **REJECT Null Hypothesis** | |

| | Method | Category | | | |
|---|---|---|---|---|---|
| | | Power | Personal Standards | Other | |
| SER Globalbrain | Expected Values (Social Psychologist) | 37 | 26 | 34 | 97 |
| | Observed Value | **65** | **26** | **6** | 97 |
| | Chi Square Test Stat | 44.24801272 | | | |
| | P-value | 2.46414E-10 | | **REJECT Null Hypothesis** | |

| | Method | Category | | | |
|---|---|---|---|---|---|
| | | Power | Personal Standards | Other | |
| Lexiquest Categorize | Expected Values (Social Psychologist) | 37 | 26 | 34 | 97 |
| | Observed Value | **32** | **53** | **12** | 97 |
| | Chi Square Test Stat | 42.94943133 | | | |
| | P-value | 4.71682E-10 | | **REJECT Null Hypothesis** | |

| | Method | Category | | | |
|---|---|---|---|---|---|
| | | Power | Personal Standards | Other | |
| PolyAnalyst | Expected Values (Social Psychologist) | 37 | 26 | 34 | 97 |
| | Observed Value | **36** | **11** | **50** | 97 |
| | Chi Square Test Stat | 18.78991334 | | | |
| | P-value | 0.000301982 | | **REJECT Null Hypothesis** | |

Figure 26. Trial I Level I Confidence Data with Collective Category Data

## 2. Level II Confidence

| Method | Category | Count | Count of Successes | Proportion Successful | Standard Error | 95% Confidence Interval | | |
|---|---|---|---|---|---|---|---|---|
| Reel Two | Power | 49 | **23** | 0.4694 | 0.0713 | 0.3521 | - | 0.5867 |
| | Personal Standards | 46 | **15** | 0.3261 | 0.0691 | 0.2124 | - | 0.4398 |
| | Other | 2 | **0** | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 |
| | Total (n) | 97 | 38 | **0.3918** | 0.0496 | **0.3102** | **-** | **0.4733** |
| SER Globalbrain | Power | 65 | **24** | 0.3692 | 0.0599 | 0.2708 | - | 0.4677 |
| | Personal Standards | 26 | **10** | 0.3846 | 0.0954 | 0.2277 | - | 0.5416 |
| | Other | 6 | **1** | 0.1667 | 0.1521 | -0.0836 | - | 0.4169 |
| | Total (n) | 97 | 35 | **0.3608** | 0.0488 | **0.2806** | **-** | **0.4410** |
| Lexiquest Categorize | Power | 32 | **15** | 0.4688 | 0.0882 | 0.3236 | - | 0.6139 |
| | Personal Standards | 53 | **17** | 0.3208 | 0.0641 | 0.2153 | - | 0.4262 |
| | Other | 12 | **6** | 0.5000 | 0.1443 | 0.2626 | - | 0.7374 |
| | Total (n) | 97 | 38 | **0.3918** | 0.0496 | **0.3102** | **-** | **0.4733** |
| PolyAnalyst | Power | 36 | **19** | 0.5278 | 0.0832 | 0.3909 | - | 0.6646 |
| | Personal Standards | 11 | **5** | 0.4545 | 0.1501 | 0.2076 | - | 0.7015 |
| | Other | 50 | **22** | 0.4400 | 0.0702 | 0.3245 | - | 0.5555 |
| | Total (n) | 97 | 46 | **0.4742** | 0.0507 | **0.3908** | **-** | **0.5576** |
| Guessing | Total | 100 | 33 | **0.3300** | 0.0470 | **0.2527** | **-** | **0.4073** |

Figure 27. Trial I Level II Confidence with Collective Category Data

Count: The total number of profiles that the software application placed in the category.

Count of Successes: Total correct number of profiles that the software application placed in the category.

### 3. Comparison

$H_0 : \rho_1 = \rho_2$

$H_a : \rho_1 \neq \rho_2$

| Software Packages | Difference between p<br>z score on difference | SE of difference<br>pooled SE | Margin of Error<br>Z | Confidence Interval<br>Judgement |
|---|---|---|---|---|
| Reel Two vs SER Globalbrain | 0.0309 | 0.0695 | 0.1144 | -0.0834 - 0.1453 |
| **Significance** | **z = 0.4446** | 0.0696 | 0.3763 | Fail to Reject Null Hypothesis |
| Reel Two vs Lexiquest Categorize | 0.0000 | 0.0701 | 0.1153 | -0.1153 - 0.1153 |
| **Significance** | **z = 0.0000** | 0.0701 | 0.3918 | Fail to Reject Null Hypothesis |
| Reel Two vs PolyAnalyst | -0.0825 | 0.0709 | 0.1166 | -0.1991 - 0.0341 |
| **Significance** | **z = -1.1592** | 0.0711 | 0.4330 | Fail to Reject Null Hypothesis |
| Reel Two vs Guessing | 0.0118 | 0.0694 | 0.1141 | -0.1024 0.1259 |
| **Significance** | **z = 0.1694** | 0.0694 | 0.3858 | Fail to Reject Null Hypothesis |
| SER Globalbrain vs Lexiquest Categorize | -0.0309 | 0.0695 | 0.1144 | -0.1453 - 0.0834 |
| **Significance** | **z = -0.4446** | 0.0696 | 0.3763 | Fail to Reject Null Hypothesis |
| SER Globalbrain vs PolyAnalyst | -0.1134 | 0.0703 | 0.1157 | -0.2291 - 0.0023 |
| **Significance** | **z = -1.6014** | 0.0708 | 0.4175 | Fail to Reject Null Hypothesis |
| SER Globalbrain vs Guessing | -0.0192 | 0.0688 | 0.1132 | -0.1323 - 0.0940 |
| **Significance** | **z = -0.2786** | 0.0688 | 0.3706 | Fail to Reject Null Hypothesis |
| Lexiquest vs PolyAnalyst | -0.0825 | 0.0709 | 0.1166 | -0.1991 - 0.0341 |
| **Significance** | **z = -1.1889** | 0.0711 | 0.4330 | Fail to Reject Null Hypothesis |
| Lexiquest Categorize vs Guessing | 0.0118 | 0.0694 | 0.1141 | -0.1024 - 0.1259 |
| **Significance** | **z = 0.1694** | 0.0694 | 0.3858 | Fail to Reject Null Hypothesis |
| PolyAnalyst vs Guessing | 0.0942 | 0.0702 | 0.1155 | -0.0212 - -0.0212 |
| **Significance** | **z = 1.3369** | 0.0705 | 0.4264 | Fail to Reject Null Hypothesis |

Figure 28. Trial I Comparison with Collective Category Data

## E. TRIAL II WITH COLLECTIVE CATEGORY DATA

### 1. Level I Confidence

$H_0$ : There is no association between the column and row variables (Method / Category). This is too say that the distributions of the categories are the same regardless of the method that produced it.

$H_a$ : There is an association between the row and column variables.

| | Method | Category: Power | Personal Standards | Other | |
|---|---|---|---|---|---|
| Reel Two | Expected Values (Social Psychologist) | 39 | 25 | 35 | 99 |
| | Observed Value | **50** | **40** | **9** | 99 |
| | Chi Square Test Stat | 31.41684982 | **REJECT Null Hypothesis** | | |
| | P-value | 1.50632E-07 | | | |

| | Method | Category: Power | Personal Standards | Other | |
|---|---|---|---|---|---|
| SER Globalbrain | Expected Values (Social Psychologist) | 39 | 25 | 35 | 99 |
| | Observed Value | **72** | **23** | **4** | 99 |
| | Chi Square Test Stat | 55.54021978 | **REJECT Null Hypothesis** | | |
| | P-value | 8.70151E-13 | | | |

| | Method | Category: Power | Personal Standards | Other | |
|---|---|---|---|---|---|
| Lexiquest Categorize | Expected Values (Social Psychologist) | 39 | 25 | 35 | 99 |
| | Observed Value | **40** | **42** | **17** | 99 |
| | Chi Square Test Stat | 23.64218235 | **REJECT Null Hypothesis** | | |
| | P-value | 2.97884E-05 | | | |

| | Method | Category: Power | Personal Standards | Other | |
|---|---|---|---|---|---|
| PolyAnalyst | Expected Values (Social Psychologist) | 39 | 25 | 35 | 99 |
| | Observed Value | **33** | **18** | **48** | 99 |
| | Chi Square Test Stat | 9.714436022 | **REJECT Null Hypothesis** | | |
| | P-value | 0.02115616 | | | |

Figure 29. Trial II Level I Confidence with Collective Category Data

## 2. Level II Confidence

| Method | Category | Count | Count of Successes | Proportion Successful | Standard Error | 95% Confidence Interval | | |
|---|---|---|---|---|---|---|---|---|
| Reel Two | Power | 50 | **24** | 0.4800 | 0.0707 | 0.3638 | - | 0.5962 |
| | Personal Standards | 40 | **14** | 0.3500 | 0.0754 | 0.2260 | - | 0.4740 |
| | Other | 9 | **4** | 0.4444 | 0.1656 | 0.1720 | - | 0.7169 |
| | Total (n) | 99 | 42 | **0.4242** | 0.0497 | **0.3425** | **-** | **0.5059** |
| SER Globalbrain | Power | 72 | **28** | 0.3889 | 0.0575 | 0.2944 | - | 0.4834 |
| | Personal Standards | 23 | **7** | 0.3043 | 0.0959 | 0.1465 | - | 0.4622 |
| | Other | 4 | **1** | 0.2500 | 0.2165 | -0.1061 | - | 0.6061 |
| | Total (n) | 99 | 36 | **0.3636** | 0.0483 | **0.2841** | **-** | **0.4432** |
| Lexiquest Categorize | Power | 40 | **19** | 0.4750 | 0.0790 | 0.3451 | - | 0.6049 |
| | Personal Standards | 42 | **14** | 0.3333 | 0.0727 | 0.2137 | - | 0.4530 |
| | Other | 17 | **5** | 0.2941 | 0.1105 | 0.1123 | - | 0.4759 |
| | Total (n) | 99 | 38 | **0.3838** | 0.0489 | **0.3034** | **-** | **0.4642** |
| PolyAnalyst | Power | 33 | **14** | 0.4242 | 0.0860 | 0.2827 | - | 0.5658 |
| | Personal Standards | 18 | **5** | 0.2778 | 0.1056 | 0.1041 | - | 0.4514 |
| | Other | 48 | **20** | 0.4167 | 0.0712 | 0.2996 | - | 0.5337 |
| | Total (n) | 99 | 39 | **0.3939** | 0.0491 | **0.3132** | **-** | **0.4747** |
| Guessing | Total | 100 | 38 | **0.3800** | 0.0485 | **0.3002** | **-** | **0.4598** |

Figure 30. Trial II Level II Confidence with Collective Category Data

Count: The total number of profiles that the software application placed in the category.

Count of Successes: Total correct number of profiles that the software application placed in the category.

### 3. Comparison

$$H_0 : \rho_1 = \rho_2$$

$$H_a : \rho_1 \neq \rho_2$$

| Software Packages | Difference between p<br>z score on difference | SE of difference | Margin of Error Z | Confidence Interval Judgement |
|---|---|---|---|---|
| Reel Two vs SER Globalbrain | 0.0606 | 0.0693 | 0.1140 | -0.0534 - 0.1746 |
| **Significance** | **z = 0.8727** | 0.0694 | 0.3939 | **Fail to Reject Null Hypothesis** |
| Reel Two vs Lexiquest Categorize | 0.0404 | 0.0697 | 0.1146 | -0.0742 - 0.1550 |
| **Significance** | **z = 0.5793** | 0.0697 | 0.4040 | **Fail to Reject Null Hypothesis** |
| Reel Two vs PolyAnalyst | 0.0303 | 0.0698 | 0.1149 | -0.0846 - 0.1452 |
| **Significance** | **z = 0.4336** | 0.0699 | 0.4091 | **Fail to Reject Null Hypothesis** |
| Reel Two vs Guessing | 0.0442 | 0.0694 | 0.1142 | -0.0700 0.1585 |
| **Significance** | **z = 0.6365** | 0.0695 | 0.4020 | **Fail to Reject Null Hypothesis** |
| SER Globalbrain vs Lexiquest Categorize | -0.0202 | 0.0687 | 0.1131 | -0.1333 - 0.0929 |
| **Significance** | **z = -0.2938** | 0.0688 | 0.3737 | **Fail to Reject Null Hypothesis** |
| SER Globalbrain vs PolyAnalyst | -0.0303 | 0.0689 | 0.1134 | -0.1437 - 0.0830 |
| **Significance** | **z = -0.4395** | 0.0689 | 0.3788 | **Fail to Reject Null Hypothesis** |
| SER Globalbrain vs Guessing | -0.0164 | 0.0685 | 0.1127 | -0.1291 - 0.0963 |
| **Significance** | **z = -0.2388** | 0.0685 | 0.3719 | **Fail to Reject Null Hypothesis** |
| Lexiquest vs PolyAnalyst | -0.0101 | 0.0693 | 0.1140 | -0.1241 - 0.1039 |
| **Significance** | **z = -0.1466** | 0.0693 | 0.3889 | **Fail to Reject Null Hypothesis** |
| Lexiquest Categorize vs Guessing | 0.0038 | 0.0689 | 0.1133 | -0.1095 - 0.1171 |
| **Significance** | **z = 0.0557** | 0.0689 | 0.3819 | **Fail to Reject Null Hypothesis** |
| PolyAnalyst vs Guessing | 0.0139 | 0.0690 | 0.1136 | -0.0996 - -0.0996 |
| **Significance** | **z = 0.2019** | 0.0691 | 0.3869 | **Fail to Reject Null Hypothesis** |

Figure 31. Trail II Comparison with Collective Category Data

THIS PAGE INTENTIONALLY LEFT BLANK

# APPENDIX B. POLYANALYST TAXONOMY

| Category | Descriptors |
|---|---|
| Power | Power or manipulate! or "power hungry" or conniving or aggressive or devious or scheming or shrewd and not honest and not morals and not trustworthy and not just and not decent |
| Personal Standards | "personal standards" or character! Or morals or (honest) or honor or logical or devoted or justice or ethical or decent and not (manipulate) and not "power hungry" and not aggressive and not devious and not scheming and not social and not collective |
| Social | Social or cog or approval! Or shy or "eager to please" or societal or collective or cooperative and not manipulative and not power and not conniving and not aggressive and not scheming and not shrewd and not "personal standards" |

Table 11. PolyAnalyst Taxonomy

PolyAnalyst searches for the individual words that are defined as descriptors. In order for the system to identify phases quotes must be used. For words that the user defines as important, all versions and tenses can be used by surrounding the word with parenthesis. PolyAnalyst also allows the one time use of an explanation point; this is used to specify that all known synonyms of that word are found.

The descriptors used in the defined taxonomy were extracted from existing profiles that belonged to the defined category. These terms and phrases had been isolated by the social psychologist during his manual categorization.

THIS PAGE INTENTIONALLY LEFT BLANK

# APPENDIX C. CLASSIFIED RESULTS

For access to the classified results, please contact Professor Raymond Buettner in the Department of Information Sciences at the Naval Postgraduate School.

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF REFERENCES


Aixploratorium. "Information Gain Seeking Small consistent Decision Trees." http://www.cs.ualberta.ca/~aixplore/learning/DecisionTrees/InterArticle/4-DecisionTree.html. October 2003.

An, A. "Learning Classification Rules from Data." International Journal of Computers and Mathematics with Applications, Vol.45, No.4-5. 2003. http://www.cs.yorku.ca/~aan/research/paper/cam03.pdf. December 2003.

Anderson, Dave and George McNeil. "Artificial Neural Networks Technology: A DACS State-of-the-Art Report." Kamon Sciences Corporation, Utica New, York. August 1992. http://www.gaianxaos.com/PdfChaosLibrary_files/ArtificialNeuralNetworksTechnology.pdf August 2004.

Ankerst, Mihael, Chrisian Elsen, Martin Ester and Hans-Peter Kriegel. "Visual Classification: An Interactive Approach to Decision Tree Construction." Institute for Computer Science. University of Munich Germany. http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/Kdd-99.final.pdf. August 2004.

Applied Psychology Research. "The Best of Both Worlds: Probabilistic and Rules-Based Text Mining." http://www.itri.bton.ac.uk/projects/euromap/Text%20Mining%20Event/Dan_Brown.pdf August 2004.

Apte, C., Damerau, F. and Weiss, S. M. (1994): "Towards Language Independent Automated Learning of Text Categorization Models." http://researchweb.watson.ibm.com/dar/papers/pdf/sigir94_with_cover.pdf. August 2004.

Association of the United States Army. *Army*. Vol. 13, No. 5. 1529 Eighteenth Street, N.W., Washington 6, D.C. December 1962.

Brucher, Heide, Gerhard Knolmayer and Marc-Andre Mittermayer. "Document Classification Methods for Organizing Explicit Knowledge." http://www.alba.edu.gr/OKLC2002/Proceedings/pdf_files/ID237.pdf. August 2004.

Casillas, J., O. Cordon, M. J. del Jesus and F. Herrera. "Genetic Feature Selection in a Fuzzy Rule-Based Classification System Learning Process for High Dimensional Problems." http://citeseer.ist.psu.edu/cache/papers/cs/18338/ftp:zSzzSzdecsai.ugr.eszSzpubzSzaraizSztech_repzSzga-flzSztr-000122.pdf/casillas00genetic.pdf. November 2003.

Chang, Chin-Chung and Chih-Jen Lin. LIBSVM -- A Library For Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/. October 2003.

Charniak, Eugene. *Bayesian Networks Without Tears*. http://www.ai.mit.edu/~murphyk/Bayes/Charniak_91.pdf. October 2003.

Cheng, Yi, Jianye Ge, Jun Liang and Sheng Yu. "Comparison of Web Page Classification Algorithms." http://www.softlab.ece.ntua.gr/facilities/public/AD/Text%20Categorization/Comparison%20of%20Web%20Page%20Classification%20Algorithms.ppt. August 2004.

Cooper, G. and E Herskovitz. *A Bayesian Method for the Induction of Probabilistic Networks from Data*. Machine Learning. 9:309-347, 1992.

Cordon, Oscar, Maria Jose del Jesus and Francisco Herrera. "Analyzing the Reasoning Mechanisms in Fuzzy Rule Based Classification Systems." http://sci2s.ugr.es/publications/byType.php?typeName=International%20Journals&typeNumber=01. August 2004.

Cortes, C. and V. Vapnik. "Support -Vector Networks." Machine Learning, 20:273-297. http://citeseer.ist.psu.edu/cache/papers/cs/23317/http:zSzzSzwww.research.att.comzSz~corinnazSzpaperszSzsupport.vector.pdf/cortes95supportvector.pdf. August 2004.

Davison, Brian D. *Classification Using an Online Genetic Algorithm.* http://www.cse.lehigh.edu/~brian/pubs/1998/aaai/aaai98stuabs.html. October 2003.

Diez F. J. "Local Conditioning in Bayesian Networks." Technical Report R-181, Cognitive Systems Lab., Dept. of Computer Science, UCLA, July 1992. http://citeseer.ist.psu.edu/context/149372/130346. August 2004.

Department of Defense, United States of America. "The Creation and Dissemination of All Forms of Information in Support of Psychological Operations (PSYOP) in Time of Military Conflict." Office of the Under Secretary of Defense For Acquisition, Technology and Logistics, Washington, D.C. 20301-3140, May 2000. http://www.fi.uib.no/~antonych/deza/deza.html. August 2004.

Dumais, S. Platt, J., Heckermann, D. and Sahami, M. (1998). "Inductive Learning Algorithms and Representations for Text." http://robotics.stanford.edu/users/sahami/papers-dir/cikm98.pdf. August 2004.

Fernandez, Jaime. *The Genetic Programming Notebook.* http://www.geneticprogramming.com. October 2003.

Furnkranz, Johannes. "A Pathology of Bottom-Up Hill-Climbing in Inductive Rule Learning." Austrian Research Institute for Artificial Intelligence Schottengasse 3, A-1010 Vienna, Austria. http://www.ke.informatik.tu-darmstadt.de/~juffi/publications/alt-02.pdf. August 2004.

Furnkranz, Johannes. "An Analysis of Rule Learning Heuristics." Austrian Research Institute for Artificial Intelligence Schottengasse 3, A-1010 Vienna, Austria. http://www.cs.bris.ac.uk/Publications/Papers/1000694.pdf. August 2004.

Ghosh, Joydeep and Yoan Shin. “Efficient Higher-Order Neural Networks for Classification and Function Approximation.” Int'l Jl. of Neural Systems , Vol. 3, No. 4, 1992. March, 1995.

Greenwald, A. G. and Pratkanis, A. R. The Self. In R. S. Wyer and T. Srull (Eds.), *The handbook of social cognition*. Hillsdale, J. F.: Lawrence Erlbaum, 1984.

Greenwald, A.G. and Breckler, S. To Whom is the Self Presented? In B. Schlenker (Ed.), *The self and social life*. New York: McGraw-Hill, 1985.

Grzymala-Busse, J. W. *Knowledge Acquisition Under Uncertainty – A Rough Set Approach*. Journal of Intelligent and Robot Systems. 1:3-16, 1988.

Han, Jiawei, Pei, Jian and Yin, Yiwen. “Mining Frequent Patterns without Candidate Generation.” School of Computing Science, Simon Fraser University, Vurnaby, British Columbia, Canada.
http://citeseer.ist.psu.edu/cache/papers/cs/14568/ftp:zSzzSzftp.fas.sfu.cazSzpubzSzcszSzhanzSzpdfzSzsigmod00.pdf/han99mining.pdf. August 2004.

Heckerman D. *A Tutorial on Learning Bayesian Networks*. Technical Report MSR-TR-95-06, Microsoft Research, 1996.

Heckerman D., D. M. Chickering, C. Meek, R. Rounthwaite and C. Kadie. *Dependency Networks for Inference, Collaborative Filtering, and Data Visualization*. Journal of Machine Learning Research, 1:49-75, October 2000.

Higgins, Jr., Charles M. “Classification and Approximation with Rule-Based Networks.” Department of Electrical Engineering California Institute of Technology
Pasadena, California 1993.
http://neuromorph.ece.arizona.edu/~higgins/pubs/oldpubs/thesis.pdf. August 2004.

Ho, Tin Kam. “The Random Subspace Method for Constructing Decision Forests.” IEEE Computer Society Washington, DC, USA.
http://portal.acm.org/citation.cfm?id=284986&dl=ACM&coll=portal. August 2004.

Hull, D. A. Stemming Algorithms: A Case Study For Detailed Evaluation. Journal of the American Society for Information Science, 47(1): 70-84, 1996.

http://www-users.cs.umn.edu/~mjoshi/hpdmtut/sld001.htm. August 2004.

Jain, Sonal and Gaurav Rathi. “Key Issues in Construction of Decision Trees for Classification in Data Mining.” http://www-scf.usc.edu/~sonaljai/sonal_rathi.PDF. August 2004.

Joachims, T. *Text Categorization With Support Vector Training.* In Proceedings of the 1997 NIPS Workshop on Support Vector Machines, 1998.

Joachims, T. (1998). "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." http://citeseer.ist.psu.edu/cache/papers/cs/26885/http:zSzzSzranger.uta.eduzSz~alpzSzixzSzreadingszSzSVMsforTextCategorization.pdf/joachims97text.pdf. August 2004.

Johansen, M. M. *Topics of Evolutionary Computation 2002 – Collection of Student Reports*. "Evolving Neural Networks for Classification." Department of Computer Science, University of Aarhus, Denmark. Fall 2002. http://www.evalife.dk/bbase/show_bibitem.php?bib_id=22588&idx=24. August 2004.

Kamber, Micheline, Lara Winstone, Wan Gong, Shan Cheng and Jiawei Han. "Generalization and Decision Tree Induction: Efficient Classification in Data Mining." http://www-faculty.cs.uiuc.edu/~hanj/pdf/ride97.pdf. August 2004.

Kdnuggets™. http://www.kdnuggets.com/software/classification-multi.html. August 2004.

Kothari, Ravi and Ming Dong. "Decision Trees for Classification: A Review and Some New Results." in *Lecture Notes in Pattern Recognition*, S.R. Pal and N.R. Pal, (Eds.), Singapore, 2001, World Scientific Publishing Company. http://www.cs.wayne.edu/~mdong/papers/paper_review.pdf. August 2004.

Krusinska, E., R. Slowinski, and J. Stefanowski. *Discriminant Versus Rough Set Approach To Vague Data Analysis*. Applied Stochastic Models and Data Analysis. 8:43-56, 1992.

Kumar, Vipin and Mahesh Joshi. "High Performance Data Mining." ACM Press, New York, NY, USA 2000. http://delivery.acm.org/10.1145/350000/349109/p309-kumar.pdf?key1=349109&key2=0081983901&coll=GUIDE&dl=GUIDE&CFID=26372550&CFTOKEN=42743664. August 2004.

Lam, W., Ho, C. Y. (1998). "Using Bayesian Network Induction Approach for Text Categorization." http://www-ai.ijs.si/DunjaMladenic/papers/PWW/pwwAAAI98.ps. August 2004.

Langley, Pat and Herbert A. Simon. "Applications of Machine Learning and Rule Induction."http://citeseer.ist.psu.edu/rd/70644833%2C109872%2C1%2C0.25%2CDownload/http://citeseer.ist.psu.edu/cache/papers/cs/512/http:zSzzSzwww.isle.orgzSz%7ElangleyzSzpaperszSzapp.cacm.pdf/langley95applications.pdf. August 2004.

Larkey, L. S. and Croft, W. B. (1996). "Combining Classifiers in Text Categorization." http://citeseer.ist.psu.edu/cache/papers/cs/97/http:zSzzSzciir.cs.umass.eduzSzinfozSzpsfileszSzirpubszSzcombo.pdf/larkey96combining.pdf. August 2004.

Lewis, D. D. and Ringuette, M. (1994). "A Comparison of Two Learning Algorithms for Text Categorization." http://citeseer.ist.psu.edu/cache/papers/cs/508/http:zSzzSzwww.cs.cmu.eduzSzafszSzcs.cmu.eduzSzuserzSzmnrzSzwwwzSzpaperszSzcateg.pdf/lewis94comparison.pdf. August 2004.

Li, Ruey-Hsia, Geneva G. Belford. "Instability of Decision Tree Classification Algorithms".http://www.cs.uiuc.edu/Dienst/Repository/2.0/Body/ncstrl.uiuc_cs/UIUCDCS-R-2001-2230/pdf. August 2004.

Lim, Tjen-Sien, Wei-Yin Loh and Yu-Shan Shih. "An Empirical Comparison of Decision Trees and Other Classification Methods." University of Wisconsin, Madison. January 1998.

McCabe, George P. and David S. Moore. *Introduction to the Practice of Statistics, 4th Ed*. W. H. Freeman and Company, 2002.

Megaputer. PolyAnalyst 4 User Manual. Megaputer Intelligence, Inc. August 2003.

Merriam-WebsterOnlineDictionary. http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=sentiment. August 2004.

Neural Network Toolbox. www.mathworks.com/access/helpdesk/help/toolbox/nnet/nnet.shtml. August 2004.

Nguyen Sinh Hoa and Nguyen Hung Son. "Some Efficient Algorithms for Rough Set Methods." To appear in Proc. of the IPMU-96, Granada, Espana. 6

Office of the Under Secretary of Defense For Acquisition, Technology and Logistics. Report of the Defense Science Board Task Force on *The Creation and Dissemination of All Forms of Information in Support of Psychological Operations (PSYOP) in Time of Military Conflict*. Washington, D.C. 20301-3140. May 2000.

Page, D. D., A. F. Koschan, S. R. Sukumar, B. Roui-Abidi and M. A. Abidi. "Shape Analysis Algorithm Based on Information Theory." University of Tennessee, Knoxville, TN. http://imaging.utk.edu/publications/papers/2003/page_icip03.pdf. August 2004.

Pal, Mahesh and Paul Mather. "Decision Tree Based Classification of Remotely Sensed Data." School of Geography, University of Nottingham, Notingham, U.K. http://www.crisp.nus.edu.sg/~acrs2001/pdf/046PAL.PDF. August 2004.

Pawlak, Z. Rough Sets*: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishers, Dordrecht, 1991.

Penny, William D. and Stephen J. Roberts. "Bayesian Neural Networks for Classification: How Useful is the Evidence Framework?". Department of Electrical and Electronic Engineering, Imperial College, London, U.K. http://citeseer.ist.psu.edu/rd/70644833%2C85509%2C1%2C0.25%2CDownload/http://citeseer.ist.psu.edu/cache/papers/cs/950/http:zSzzSzwww.ee.ic.ac.ukzSzresearchzSzneuralzSzwpennyzSzpublicationszSz.zSzevidence-tech.pdf/penny99bayesian.pdf. August 2004.

Platt. J. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization.* In Advances in Kernel Methods -- Support Vector Learning, pp. 42-65. MIT Press, January 1999.

Polkowski, A. S. L. *Rough Sets in Knowledge Discovery*. Physica-Verlag, 1998.

Post, Jerrold M. Assessing Leaders at a Distance: The Political Personality Profile. In J. Post (Ed.), *The Psychological Assessment of Political Leaders.* Ann Arbor: The University of Michigan Press, 2003.

Pratkanis, A. R. 'How to Sell a Pseudoscience', Skeptical Enquirer 19(4): 19-25 (1995).

Pratkanis, Anthony, R. (in press). Social Influence Analysis: An Index of Tactics. In A. R. Pratkanis (Ed.), *The Science of Social Influence: Advances and Future Progress*. Philadelphia: Psychology Press.

Radvanyi, Janos. *Psychological Operations and Political Warfare in Long-term Strategic Planning*. Praeger Publishers, New York, New York. 1990.

Reel Two. Classification System White Paper; The Reel Two Solution for Automatic Text Categorization. Reel Two, Inc. June 2003.

Reel Two. http://www.reeltwo.com/index.html. August 2004.

Reel Two. Reel Two Help Guide for the Reel Two Classification System. Version 2.4.6.

Reel Two. The Reel Two Solution for Automatic Text Categorization. Classification System White Paper; Reel Two, Inc., June 2003.

Reel Two. Weighted Confidence Learner Algorithm. Reel Two Technical Brief; Reel Two, Inc.

Rhoad, Kelton. "Working Psychology." Los Angeles, California. www.workingpsychology.com. August 2004.

Ruiz, M. E. and Srinivasan, P. (1998). "Automatic Text Categorization Using Neural Network." http://informatics.buffalo.edu/faculty/ruiz/publications/sigcr97/sigcrfinal2.html. August 2004.

SAS. “Getting Started with SAS Text Miner Software Release 8.2.” SAS Publishing. SAS Institute Inc. Cary, North Carolina. 2002.

SER Solutions, Inc., www.ser.com. May 2004.

SERbrainware. SERbrainware: The Full Perspective. Version 2.1 White Paper; SER Solutions, Inc., October 15, 2001.

SERglobalBrain. Intelligent Search and Retrieval. SERglobalBrain-0404 White Paper; SER Solutions, Inc., 2004.

SERglobalBrain. SERglobalBrain Personal Edition User Guide. Version 1.7.0.

SERglobalBrain. SERglobalBrain Technical White Paper; SER Solutions, Inc. April 2004.

Shannon, C. E. “A Mathematical Theory of Communication.” http://www.essrl.wustl.edu/~jao/itrg/shannon.pdf. August 2004.

Siolas, G. and d’Alche-Buc, F. (2000). “Support Vector Machines Based on a Semantic Kernel for Text Categorization”. In IEEE-IJCNN 2000.

SPSS. “LexiQuest Categorization System Algorithms.” SPSS. Chicago, IL. October 2003.

Szladow, A. and W. Ziarko. *Rough Sets: Working With Imperfect Data*. AI Expert. 7:36-41, 1993.

*The Page of Genetic Programming Inc*. http://www.genetic-programming.com. August 2004.

Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

Wall, Mathew. *Intro To Genetic Programming.* http://lancet.mit.edu/~mbwall/presentations/IntroToGAs/. October 2003.

Wedel, Michel and Wagner A. Kamakura. *Market Segmentation: Conceptual and Methodological Foundations*. International Series in Quantitative Marketing, Vol. 8. Kluwer Academic Publishers; 1st Edition, November 1999.

*www.genetic-programming.com, The Page of Genetic Programming Inc*., http://www.genetic-programming.com. August 2004.

Xu, J. and W. B. Croft. Corpus-based Stemming Using Co-Occurrence of Word Variants. ACM TOIS, 16(1):61-81, January 1998.

Yang, Y. and Liu, X. (1999). "A Re-Examination of Text Categorization Methods." http://citeseer.ist.psu.edu/cache/papers/cs/26885/http:zSzzSzranger.uta.eduzSz~alpzSzixzSzreadingszSzYangSigir99CategorizationBenchmark.pdf/yang99reexamination.pdf. August 2004.

Yang, Y. *An Evaluation of Statistical Approaches to Text Categorization*. Technical Report CMU-CS-97-127, Carnegie Mellon University.

Yin, Xiaoxin and Jiawei Han. "CPAR: Classification Based on Predictive Association Rules." University of Illinois, Urbana-Champaign. http://www.ncassr.org/projects/sift/papers/xiaoxinCPAR_siam2003.pdf. August 2004.

# INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Marine Corps Representative
   Naval Postgraduate School
   Monterey, California

4. Director, Training and Education, MCCDC, Code C46
   Quantico, Virginia

5. Director, Marine Corps Research Center, MCCDC, Code C40RC
   Quantico, Virginia

6. Marine Corps Tactical Systems Support Activity (Attn: Operations Officer)
   Camp Pendleton, California

7. Professor Raymond Buettner
   Naval Postgraduate School
   Monterey, California

8. Professor Magdi Kamel
   Naval Postgraduate School
   Monterey, California

9. Professor Russell Gottfried
   Naval Postgraduate School
   Monterey, California

10. Professor Anthony Pratkanis
    University of California at Santa Cruz
    Santa Cruz, California

11. Professor Dan C. Boger
    Naval Postgraduate School
    Monterey, California