Title

Who searches the internal tobacco industry documents and why: Using informatics to improve public health

Author

Michel, Martha C

Publication Date

2005

Peer reviewed|Thesis/dissertation

Who Searches the Internal Tobacco Industry Documents and Why:

Using Informatics to Improve Public Health

by

Martha C. Michel

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Date University Librarian

Degree Conferred:.....

ACKNOWLEDGMENTS

I give my deep appreciation and gratitude to all who have helped me through this dissertation process. I am blessed to work with some of the finest people at UCSF, my committee, Dr. Lisa Bero, Dr. Donna Hudson, and Dr. Stan Glantz. My committee has been helpful and thoughtful. I am grateful for their advice, comments, and commitment.

My advisor, Dr. Lisa Bero, has been fantastic. She has allowed me the freedom to pursue my own ideas but has prevented me from going too far off track. She is enthusiastic and encouraging, yet has high expectations, which has helped me push myself further.

In addition, I am grateful for discussions and insight about my dissertation from Dr. Marti Hearst and Dr. Ida Sim. Dr. Hearst has been helpful in development of the "Tobacco Flamenco".

Also, I am extremely grateful to my dissertation coach, Dorothy Brown, for her practical advice and good tips. I've enjoyed many conversations with Kirsten Neilsen and appreciate her bright ideas and dedication to making the tobacco industry documents more accessible.

My friends have been great for providing support and much needed breaks. I'd like to thank Annette Mears, Judy Goldstein, Sharon Longo, May Yamate, Diana Campbell, Miki Hong, Mariya Strauss, Beth Herman, Tina Thornton, Chi Kao, and others for their encouragement and support.

My friends in the Biological and Medical Informatics program have been great for both sharing ideas and having fun. I will always hold a special place in my heart for my Women in Life Science group: Meg Byrne, Michelle Lent, Rachael Friedman and Morgan Royce-Tolland. In addition, special thanks to Dani Behonick, JoAnn Lopez, and Maureen Conway. You all are the best!

I thank all the members of the Center for Tobacco Control and Education for their advice and friendship. In addition, the members of the Bero lab have been great friends. I'd like to thank Bonnie Glaser, Peggy Loppipero, Erika Campbell, Rebecca Wilson-Loots, Steve Batilero, Dan Cook, Jenny White, Dorie Appollonio, Josh Dunsby, Annamaria Baba, Miki Hong, Christopher Jewell, Kirby Lee, Dale Rose, Fieke Oostovogel, and Tiffani Bright for their support.

I'd like to thank my husband, David, who has given me support, advice, and encouragement throughout the process. I would also like to thank my in-laws, Ilana and Erhard Konerding, and my sister-in-law, Rebecca, for their love and support. My parents have always provided love and guidance without judgment. Thanks also to Hoover and Norm.

DISSERTATION

WHO SEARCHES THE INTERNAL TOBACCO INDUSTRY DOCUMENTS AND WHY: USING INFORMATICS TO IMPROVE PUBLIC HEALTH

DOCTOR OF PHILOSOPHY 2005

MARTHA C. MICHEL, B.A., UNIVERSITY OF CALIFORNIA SANTA CRUZ

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF CALIFORNIA SAN FRANCISCO

Directed by: Professor Lisa A. Bero

Table of Contents

Who Searches the Internal Tobacco Industry Documents and Why: Using Informatics to Improve Public Health

Appendices

LIST OF TABLES

Chapter 1 **Page**

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST OF FIGURES

Chapter 1 **Page**

Chapter 2

Chapter 3

Chapter 4

Chapter 5

WHO SEARCHES THE INTERNAL TOBACCO INDUSTRY DOCUMENTS AND WHY: USING INFORMATICS TO IMPROVE PUBLIC HEALTH

MARTHA MICHEL
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
DIRECTED BY: DR. LISA BERO

ABSTRACT
2005

In 1998, millions of internal tobacco industry documents were released onto the Internet. Access to these documents led to discoveries about how the tobacco industry behaves and their motivations. The utilization of the internal tobacco industry documents has changed public health. However, along with this access came problems in accessing and utilizing the documents to their maximum capacity.

The aim of this dissertation is to study 1.) who uses the internal tobacco industry documents 2.) how they are used, 3.) problems using the documents 4.) methods to improve searching. These questions are examined using both quantitative and qualitative methods via a survey and a usability study of the Legacy Tobacco Documents Library. One of the major findings of the study is that teachers, lawyers, and public health advocates are under-utilizing searching of the internal tobacco industry documents.

Then I explored researching the tobacco industry documents using different text mining algorithms. Using SAS Text Miner™ and Leximancer, I compared the results of concepts derived from the documents to a gold standard. The results showed that similar concepts were obtained using Leximancer when comparing it to a gold standard. I concluded that text mining the tobacco industry documents may be a useful technique for researchers.

There are many different reasons for searching the internal tobacco industry documents. The development of new user interfaces may help to encourage different user groups to continue to utilize this tremendous resource.

Lisa Bero

PART I: FRAMING THE PROBLEM

Chapter 1.
Introduction

"We accept an interest in people's health as a basic responsibility, paramount to every other consideration in our business. We believe the products we make are not injurious to health. We always have and always will cooperate closely with those whose task it is to safeguard the public health."

- From "A Frank Statement to Cigarette Smokers" a full page advertisement produced by the tobacco industry, 1954(1).

A. PUBLIC HEALTH INFORMATICS

Thanks to litigation, whistleblowers, advocates, and public health professionals, we now have access on the web to millions of internal documents of the tobacco industry. The documents give us an unprecedented look into their deception, as shown in the above quote from an advertisement directed to cigarette smokers. However, there are millions of these documents and with them comes a problem that has commonly been referred to as information overload. In the digital age, more information is available, but with less filtering. Even scholars have difficulty in using this amazing resource.

This dissertation utilizes methods and techniques from public health informatics to study digital collections of the internal tobacco industry documents. Public health informatics is an interdisciplinary field that derives many tools from medical informatics. However, it is useful to examine the history of such new fields to put the findings into context. Initially a brief history public health informatics will be discussed, followed by its applicability to tobacco control as it relates to the internal tobacco industry documents.

1. The Beginning of Public Health Informatics

The Soho section of London was the site of a terrible and devastating outbreak of cholera, which killed 600 people within a quarter of a mile in the course of a few days in 1854. It was initially assumed that cholera was airborne, another form of miasma or bad air, but an epidemiologist, Dr. John Snow, was sure that this was not true. He had treated many patients without getting cholera himself and argued that the infection always seemed to affect the stomach before the patient felt generally ill. These facts suggested to him that the disease agent was ingested rather than airborne (2).

In late August 1854, there were only a few cholera deaths in Soho. But on August 31 and September 1, there was a significant increase in the disease. Fifty-six new cases were reported that night; the next day there were 143 new cases and on September 2, 116 more were reported. The deaths of those victims followed soon after their diagnosis (3).

When he heard of the outbreak, Snow was determined to investigate it. To account for such a rapid, violent epidemic, he became certain that the water must be contaminated. There was a popular water pump that stood at the junction of Broad Street and Cambridge Street. He examined the water pump but found only minimal visible contamination. Because this was not enough evidence, he then went to the Register of Deaths and obtained details of all the deaths from cholera from other London districts. With that data, Snow returned to the streets to find out what had really happened. The obvious conclusion came only after mapping the death data, which showed that most of the deaths were close to the pump. In fact, of the 89 who died by September 2, only ten lived closer to any other water pump. His investigation of the statistics provided strong evidence that the water from the Broad Street pump was contaminated. When the handle

from the Broad Street pump was removed the number of cases immediately started to diminish (3).

The importance of Snow's work was his recognition of the power of gathering data about the characteristics associated with the cholera deaths. Instead of gathering anecdotal evidence, Snow was examining cumulative data. This was the beginning of public health informatics and surveillance and the end of the cholera epidemic in Britain (3).

Figure I displays Snow's map of London and each dot represents a cholera case. He was the first person to link the display of information on a map to disease state, which in his time was quite novel. While some may argue that the beginning of medical informatics could not occur before the invention of the computer, Snow's map and the research of others as early medical geographers and epidemiologists, show the beginning of assembling and displaying information in a novel way, a core aspect of medical informatics.

Figure 1: John Snow's map of London displaying cholera cases. (4)

2. History of Surveillance Systems and Organizations

In 1878, Congress authorized the U.S. Marine Hospital Service, the predecessor of the U.S. Public Health Service, to collect reports of deaths from cholera, smallpox, plague, and yellow fever from U.S. consulates in other countries. The information was used for establishing measures to quarantine victims and to prevent the spread of diseases from other nations into the U.S. States. In 1879, Congress appropriated funds for the collection and publication of reports of these diseases. The authority for weekly reporting and publication of these reports was expanded in 1893. To increase the uniformity of the

data, another law was enacted in 1902 directing the development and provision of forms for the aggregation of data and for the publication at the national level.

By 1912, state and territorial health authorities and the U.S. Public Health Service (PHS) suggested telegraphic reporting of five infectious diseases and the monthly reporting of ten additional diseases. The first annual summary in 1912 of "The Notifiable Diseases" included reports of ten diseases from 19 states. In 1928, all states and territories were participating in national reporting of 29 specified diseases. At their annual meeting in 1950, the State and Territorial Health Officers authorized a conference of state and territorial epidemiologists whose purpose was to determine which diseases should be reported to PHS. In 1961, the Centers for Disease Control (CDC) assumed responsibility for the collection and publication of data concerning nationally notifiable diseases and a reliable surveillance system was born (5).

With the growth of surveillance grew a need for informatics and the organizations to support a new field. The American Medical Informatics Association (AMIA) was formed in 1990 by the merger of three organizations: the American Association for Medical Systems and Informatics (AAMSI), the American College of Medical Informatics (ACMI), and the Symposium on Computer Applications in Medical Care (SCAMC). AMIA is the official representative of the United States in the International Medical Informatics Association (IMIA).

Medical informatics endeavors to provide both the theoretical and scientific basis for computer applications in biomedicine and other health areas. It provides for the development of biomedical information, data, and storage retrieval, and problem solving and decision-making (6). Public health informatics is an interdisciplinary field that

combines theories of public health, statistics, computer science, information science and medical informatics with the goal of improving health at the population level. It is not represented by a single organization; rather different organizations, such as AMIA and the American Public Health Association (APHA), have subgroups specifically for researchers in public health informatics.

3. Informatics and Public Health

The American Public Health Association (APHA) is the oldest and largest organization of public health professionals in the world. It represents more than 50,000 members from over fifty occupations of public health. APHA has had influence over policies and setting priorities in public health for over a century. Its history shows that it was at the vanguard of numerous efforts to prevent disease and promote health.

APHA is concerned with broad issues in personal and environmental health, including federal and state funding for health programs, pollution control, programs and policies related to chronic and infectious diseases, promoting a smoke-free society, and professional education in public health (7). Many of these goals are furthered by information technology. Public health informatics is defined as "the systematic application of information and computer science and technology to public health practice, research and learning"(8).

As a field, public health informatics attempts to join experts in public health, health information systems, and informatics to establish a paradigm for developing health information systems. Its goal is to foster effective health information systems through collaboration, innovation, and action (9). The approach to public health information

systems has been to combine practices in informatics with knowledge and experience in public health and health care. Skills in governance, project planning, change management, and communication made certain that activities achieved their goals and the outcomes in health (9). Public health informatics also began to foster research on supporting a smoke-free society(9).

4. Informatics and Tobacco Control

Tobacco control has advanced significantly since the 1970s. Ever since the release of the 1964 Surgeon General's Report entitled Smoking and Health, the movement has grown (10). The tobacco control movement has been assisted by the growth of public health informatics, which has helped by enhancing rapid communication by email and also enabling large data collection and processing.

Two large surveys highlight the influence of informatics on tobacco control. In late 1998, The National Center for Chronic Disease Prevention and Health Promotion's Office on Smoking and Health and the World Health Organization's (WHO) Tobacco Free Initiative launched a Global Youth Tobacco Survey as part of a cooperative project between WHO and the United Nations Children's Education Fund - supported project on youth and tobacco (11). The Global Youth Tobacco Survey was a school-based, tobacco-specific survey that focuses on adolescents aged 13-15 years, providing an in-depth assessment of students' knowledge, attitudes, and behaviors related to tobacco. The Global Youth Tobacco Survey (GYTS) was focused on tracking tobacco use among young people across countries using a common methodology and core questionnaire. The GYTS surveillance system was intended to enhance the capacity of countries to design, implement, and evaluate tobacco control and prevention programs

(11). Without the coordinated efforts of many countries agreeing on standards and protocols to exchange the data, such a large survey would not be possible.

In early 2000, the Office of Smoking and Health, the World Health Organization, and Tobacco Free Initiative added a Global School Personnel Survey (GSPS) to the existing youth tobacco survey initiative. The GSPS, which was piloted in several countries between 2000 and 2001, was a tobacco-specific survey of school personnel providing in-depth assessment of behavior, knowledge, attitude, and school curriculum and policies regarding tobacco. The Centers for Disease Control and Office of Smoking and Health supported the development and review of the final questionnaires. These organizations provided data collection answer sheets. They also processed, edited and weighted all data. Further, they produced detailed tables; and provided a data diskette for other analyses. The OSH staff also made available ongoing technical assistance for further analyses, interpretation of results, and report writing (11). These efforts require incredible collaboration and communication that is enhanced by what has been learned over time in public health informatics.

B. TEXT MINING – A NEW PUBLIC HEALTH INFORMATICS TOOL

Text mining is a new tool used in a variety of applied fields, specifically biology, but which has not been applied to public health data. It is a set of methods used to train, a computer to extract new information from huge amounts of unstructured textual data. It is still a "nascent field" (12); however, various innovations have been made in statistical language processing, information science, and computer science to make the field more accessible.

In a seminal paper on natural language processing, it was noted that significant words were powerful in their primary texts (13). Another researcher also found that text mining and word synthesizing research methods were natural characterizations and organizations of information. Natural language processing (NLP) came from an analysis of frequencies and distributions of words (14). NLP articulated ideas that were already contained in the scientific literature, and were natural phenomena worthy of exploration and synthesis. However, because NLP is wholly dependent on the computer to find connections it was contrasted unfavorably with scientific attitudes toward information and the activities of intelligence analysts (15). The scientific approach was for the scientist herself to read and synthesize texts and research, not depend on computers. Therefore, researchers may see NLP as "fishing" for results.

By using the computer, scientists were called upon to be more like intelligence analysts. They had to perceive value in examining the thoughts from scientific literature and see in it new knowledge that is as valuable as knowledge from the laboratory. That challenge prompted Swanson to develop a system to discover meaningful new knowledge in the biomedical literature (16). Swanson developed a text-mining program, Arrowsmith, that extracts new knowledge from published literature and made it available on the web (17).

The program works by finding common keywords and phrases in both "complementary and noninteractive" sets of articles or literatures and juxtaposing representative citations likely to reveal interesting co-occurrences. Two sets of literatures were considered complementary if together they can reveal useful information not apparent in the two sets considered separately. Swanson discovered at least three

biomedically important relationships using this system: an association between fish oil consumption and Raynaud's syndrome; magnesium, migraines, and epilepsy; and arginine and somatomedin C (18-20). Most recently he has used the Arrowsmith program to identify several dozen viruses as potential bioweapons (21).

Lindsay and Gordon and Kostoff (1999) extended Swanson's approach without calling it text mining(18). However, Kostoff's other work explicitly used the term text mining at one point. Swanson's system was essentially as follows: Medline searches were done on two subjects and the results were cast into Arrowsmith, which generated a list of all significant words and phrases common to the two result sets. They then used this information juxtaposed with pairs of text passages for the user to consider as possibly related (16). Lindsay and Gordon added lexical frequency statistics (see Chapter 6 re: term frequency*inverse document frequency) to rank the common words and phrases by probable discriminatory value (18). Their system, like Swanson's, still requires human filters at several points.

Kostoff and co-workers published several papers on the Web describing various text mining systems and applications. Losiewicz, Oard, and Kostoff described a text data mining (TDM) architecture that unified information retrieval with text collections, information extraction from individual texts, knowledge discovery in databases, knowledge management in organizations, and visualization of data and information(22). What they meant by "unified" information retrieval was unclear, but this statement unmistakably attempts to generate a broad view of text mining as including many things beyond mere collection of word frequencies. It appears to be a synonym for the entire range of nontraditional information retrieval strategies. The TDM architecture they

described included subsystems for data collection, data warehousing information extraction and data storage, and data exploitation, data mining and presentation. Kostoff's paper suggests a system for extracting and analyzing metadata. Metadata is commonly referred to as data about data. The authors discussed linguistic analysis and numerous exotic pattern-finding techniques, but these appeared to be long-range goals.

Current work in text mining and statistical natural language processing has focused on the more pedestrian challenges of relevance feedback, bibliometrics, and phrase extraction and statistics. The methods continue to be time and labor intensive, requiring the close involvement of technical domain experts at every level of processing. The intent is still to automate text mining, but technical experts are still indispensable. Part of the intention of my study has been to explore automation in the area of information retrieval and text mining, with regard to the tobacco industry documents. This exploration may serve to broaden the horizons and perspectives of technical experts who wish to communicate their expertise to non-technical users. The further goal of my study is to make the methods of text mining accessible to the non-technical users themselves, as well as allowing them to understand the experts. To that end, I have included information on the text mining tools and techniques.

C. INTERNAL TOBACCO INDUSTRY DOCUMENTS

1. Origins and Existence of the Documents

In 1994, previously secret, internal tobacco industry documents were at the center of several analyses and research studies that changed public health (23). Recent litigation and the Master Settlement Agreement of 1998 made millions of tobacco industry internal

documents available on the Internet (24). As required by the Master Settlement Agreement, tobacco companies maintain websites where many of the documents are located. Government and private funding sources have begun to support the archiving and indexing of these documents on other Internet sites(25, 26).

Multiple locations on the Internet contain different types of internal tobacco industry documents (Table 1). The websites include tobacco industry sites, university sites, and private sites all of which have advantages and disadvantages to their search methods.

Table 1: Electronic Tobacco Industry Document Archives

Name	*Website address*
British American Tobacco Documents Archive	http://bat.library.ucsf.edu/
Brown and Williamson /American Tobacco Company	http://www.bwdocs.com/
CDC Tobacco Industry Documents	www.cdc.gov/tobacco/industrydocs/
Council for Tobacco Research Document website	www.ctr-usa.org/ctr
Legacy website	www.legacy.library.ucsf.edu
Lorillard	www.lorillarddocs.com
Philip Morris website	www.pmdocs.com
RJ Reynolds website	www.rjrtdocs.com
Tobacco Archives	www.tobaccoarchives.com/
Tobacco Control Archives	www.library.ucsf.edu/tobacco/
Tobacco Documents Online	www.tobaccodocuments.org
Tobacco Institute Document website	www.tobaccoinstitute.com

The University of California at San Francisco (UCSF) library is creating a permanent Internet archive of all the tobacco industry documents. The UCSF library archived the Legacy collection, which as of June 13, 2005 is 7.2 million documents with an estimated 42 million pages; approximately 1.4 terabytes of data(26). In addition, the UCSF library contains four separate collections, called the Tobacco Control Archives.

The British American Tobacco Company (BATCo) documents, a paper collection of approximately 8 million documents, is housed in Guildford, England. The British American Tobacco Documents Archive is a new collection of documents from BAT that

consists of 2,251,092 pages in 536,413 documents, approximately 125 gigabytes, which were minimally indexed and converted into text using optical character recognition algorithms(27). These documents have been gathered by researchers throughout the world and systematically compiled and organized. The collection is still growing. It is estimated that there are 7 million additional documents in the Guilford collection, which will eventually be online in 2006. The Guilford collection is administered by British American Tobacco and unlike the Minnesota depository, access to the documents is quite controlled (28).

An additional important collection is the BATCo documents which UCSF currently hosts an online collection of 16,554 documents with about 52,000 pages. BATCo is a separate collection from BATDa and contains documents collected from public health groups, including Health Canada, the British Columbia Ministry of Health, Physicians for a Smokefree Canada, and the World Health Organization (29).

These documents have been manually indexed and have had controlled vocabulary terms assigned to them (30). The vocabulary contains hierarchical relationships and related terms (see Appendix G). The other documents on the Legacy website were not indexed using a controlled vocabulary. Since the BATCo documents were assigned controlled vocabulary by professional indexers, they were an ideal set of documents that serve as a gold standard for testing machine-learning algorithms.

2. Use of the Documents in Public Health

The internal tobacco industry documents have changed public health. They have demonstrated the tobacco industry motives and how the industry operates away from the

public eye, behind closed doors. An article by Dr. Lisa Bero summarizes the research to date from the tobacco industry documents (Figure 2) (31).

Figure 2: Findings from the Internal Tobacco Industry documents

Tobacco industry motives

- Profit
- Fear of litigation
- Protect tobacco from regulation
- Concerns about credibility/image of the industry

How the tobacco industry operates

- Deceive the public and policy makers
- Hide information from the public and policy makers
- Create controversy
- Involve lawyers in decisions – from scientific research to marketing to public relations
- Use third parties or front groups to hide political lobbying
- Coordinate action and communication among tobacco companies globally
- Influence practices/procedures that affect a variety of corporate interests
- Use financial ties with other corporations to pressure those organizations to support tobacco industry goals

The truth about tobacco and tobacco advertising

- Nicotine is a drug
- Nicotine is addictive
- Secondhand smoke exposure is harmful to health
- Industry attempts to develop less harmful tobacco products have been a failure
- Tobacco advertising, promotions, and product design target youth
- Tobacco advertising aims to increase consumption of tobacco products

The internal tobacco industry documents are the largest glimpse people have had into the workings of any industry in the United States. These findings shown in Figure 2 show the thinking of an industry responsible for millions of deaths. Uncovering these

documents was of tremendous use to public health in that when the truth came out about the industry, public opinion began to turn against them. That change was in part forwarded by what was contained within the documents.

3. Use of the Documents in Advocacy

As will be demonstrated in the dissertation, advocates currently underutilize the internal tobacco industry documents. However, there are some examples where the documents have been successfully used. The documents have been successfully used in tobacco control advertisements in the state campaigns of Massachusetts and California. Another example of a successful use of the documents in tobacco control is the Truth campaign. The Truth campaign had a number of commercials that directly utilized the documents.

4. Problems with Searching the Documents

Searching the internal tobacco industry documents remains difficult, although it has become easier. Malone and Balbach state that the documents "...may prove to be either a treasure trove of information valuable for tobacco control research and advocacy, or a quagmire of quantity into which researchers sink in despair."(32). It takes researchers months to search through the documents. Part of the problem with searching the documents includes misspelling, variations in spellings and spacing in the text being searched. In addition, the sheer volume of documents makes it difficult to find all of the relevant information about a topic. In information retrieval, finding all the documents relevant to a topic is called precision. In searching the documents these problems are common when searching other large datasets.

In addition the lack of being able to search the text of documents is a difficulty. Currently the documents hosted on the Legacy Tobacco Documents Library can only be searched by their metadata. Metadata is data about data and was the original data by which the documents were indexed for court cases. The metadata contained in most of the tobacco industry documents include title, document author, document date, Bates number (a unique identifier for each document), mentioned names, characteristics of documents (e.g. illegible, handwritten etc.), document type (e.g., memo, letter, email, scientific report), copied, recipient, request number, and litigation usage.

Finally in searching the documents, there is a lot of noise to signal. Many of the documents are receipts, newspaper articles, published journal articles and other publicly accessible documents. While these documents might be interesting or have certain handwritten notations, for the most part researchers cannot distinguish these documents and have to go through them by hand.

D.) DEFINITION OF THE RESEARCH

1. Profile of Potential Users

There are many potential users of the tobacco industry documents. Some are currently using the documents; others could benefit by using the documents but are currently not using them, or not using them as much as they could. The potential categories of document users that we identified in context of this dissertation are as follows: lawyers, legal analysts, teachers, researchers, advocates, and students.

2. Objective

Based on the observations of the users of internal tobacco industry documents housed online at the University of California-San Francisco, the objective of this dissertation was: to determine the characteristics of a user interface that would improve access to data from the two tobacco data sets, the Legacy collection and the British American Tobacco Documents Archive.

The study proposed to gather information to improve user accessibility and user searches of the tobacco documents. It utilized a mix of research techniques from medical informatics, public health, and surveillance techniques. A case study was conducted with features of quasi-experimental research using a set of questions to gather information from users. Further, a survey focused on problems in the system (33, 34). Finally this dissertation tested two text mining systems using internal tobacco documents.

3. Significance

a) For users of the tobacco documents - There are different reasons that people come the website to search for internal tobacco industry documents. Some are researchers while others are curious. This dissertation explores the different types of searchers and investigates tools that could be useful to then. There are some users who would benefit from alternative user interfaces and additional tools while searching the documents. For advocates, lawyers or laypeople a new interface which encourages browsing the documents may be helpful (see Appendix E). Researchers would benefit from tools specifically relevant to their searching needs. Chapter 6 further investigates some clustering tools useful for document researchers.

b) For interface design of text mining tools - Text mining is still a new field which has not been widely adopted. Therefore, there are opportunities not only to investigate new methods for text mining, but also to create new user interfaces specifically designed for text mining.

This dissertation suggests that the designers of user interfaces first discover what the users of the dataset to be mined want or need in a system. Users can be interviewed or questioned by surveys using both qualitative and quantitative methods. By determining the different potential user groups and assessing their needs, the designer can create a text mining system that will be used and save resources.

c) Applicability to use of other databases -

The tobacco industry documents are a unique collection within public health; however, they are similar to other large datasets. Methods developed for the tobacco industry documents can translate into tools that are useful for other datasets and other programs and tools created for other large datasets may be useful for the internal tobacco industry document.

The foremost example is PubMed. PubMed is a biomedical database with over 15 million citations that date back to the 1950s. Various clustering algorithms have been applied to PubMed and are used regularly in searching it. Particularly, a product called Vivismo clusters the results of PubMed searches (35). Findings from this current dissertation suggest various products similar to Vivismo would be useful for searching the internal tobacco industry documents. Improvements suggested in this study would be useful for other large datasets.

My objective is to develop new methods of searching the tobacco industry documents by using informatics tools to discover new information about the tobacco industry. Using these methods will enable people to examine the industry in a different way and facilitate connections that were not easily made using a traditional search engine. The tobacco industry documents have already been valuable for advancing public health objectives of reducing tobacco use and exposure(31). However, tobacco control researchers have done most document analyses. Developing new ways of searching could open up analysis to other research disciplines and advocates.

4. Thesis roadmap: tackling an informatics problem in a human way

This thesis contains two sections. The first section frames the problem. Public health informatics is a new field with an old history. While there have been many people in public health concerned with gathering data, information, and knowledge in an efficient and effective way, they have not referred to themselves as informaticists until recently. PHI is truly an interdisciplinary field and therefore requires a thorough background and justification of my approach.

Chapter 2 addresses the current state of the literature in the many fields that comprise public health informatics; human-computer interaction, computer science, statistics, computer networking and administration, public health, biostatistics, epidemiology and public policy. It begins with a description of human-computer interaction and selectively follows the literature through to text data mining, another interdisciplinary field.

Chapter 2 ends with some discussion about approaching informatics from the human angle, first and foremost. There are numerous case studies in medical informatics that describe the failure of cleverly designed systems that ignore the human need. Some systems are built without a need and others are built with a need in mind but the necessities of the program do not mesh with the daily workflow. These issues will be further explored in Chapter 2.

The second section describes 3 different studies. Chapter 3 describes primary study of who currently uses the internal tobacco industry documents and why. The tobacco industry documents are a fascinating resource that is underutilized by various populations. While steps are being initiated to rectify this, there is still much work to be done to reach all target audiences. Chapter 4 describes in-depth interviews conducted to obtain incite about how tobacco document researchers do their work and what tools could assist them in their job. Chapter 5 is about using a specific text data mining tool, clustering, to explore the internal tobacco industry documents.

The final chapter explains further directions to investigate in text mining the tobacco industry documents. The chapter discusses why we want to assess peoples' needs before developing informatics systems and how those systems will likely be more cost effective.

E. REFERENCES

1. Tobacco Industry Research Comm TIRC, New York Herald Tribune NYHT. A Frank Statement to Cigarette Smokers. In. Tobacco Institute; 1954. p. TIMN0040888.

2. Snow J. On the mode of communication of cholera. 2nd ed. London: Church; 1855.

3. BBC. Historic Features: John Snow. 2001 [cited 2005 August 17]; Available from: http://www.bbc.co.uk/history/historic_figures/snow_john.shtml

4. Tufte E. The Visual Display of Quantitative Information. 2nd ed. Cheshire, Connecticut: Graphics Press; 1983.

5. Division of Public Health Surveillance and Informatics - CDC. National Notifiable Diseases Surveillance System. In.; 2004.

6. American Medical Informatics Association. Frequently Asked Questions. 2005 [cited 2005 September 5]; Available from: http://www.amia.org/inside/faq/

7. APHA. About APHA. [cited 2005 July 21]; Available from: http://www.apha.org/about/

8. Yasnoff W, O'Carroll PW, Koo D, Linkins RW, Kilborne EM. Public Health Informatics: Improving and Transforming Public Health in the Information Age. Journal of Public Health Management and Practice 2000;6(6):67-75.

9. Public Health Informatics Institute. 2005 [cited 2005 August 17]; Available from: http://www.phii.org/

10. U.S. Surgeon General. Reducing the Health Consequences of Smoking. In: Public Health Service, editor.: U.S. Government; 1964.

11. Centers for Disease Control. Global Youth Tobacco Survey (GYTS) - Introduction. 2005 [cited 2005 July 21]; Available from: http://www.cdc.gov/tobacco/global/GYTS/GYTS_intro.htm

12. Hearst M. Untangling Text Data Mining. In: Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics; 1999 June 20-26; University of Maryland; 1999.

13. Luhn H. The automatic creation of literature abstracts. IBM Journal of Research and Development 1958;2:159-165.

14. Doyle L. Semantic road map for literature searchers. Journal of the Association for Computing Machinery 1961;8(4):553 - 578.

15. Swanson DR. Historical note: Information retrieval and the future of an illusion. Journal of the American Society for Information Science 1988;39:92-98.

16. Swanson DR, Smalheiser NR. Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. Library Trends 1999;48:48-51.

17. Smalheiser N, Torvik V, Lugli G, Zhang W, Zhou W, Hulth M, et al. The Arrowsmith Project Homepage. 2002 [cited 2005 August 28]; Available from: http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html

18. Lindsay R, Gordon M. Literature-Based Discovery by Lexical Statistics. Journal of the American Society for Information Science 1999;50(7):574-587.

19. Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med 1988;31(4):526-57.

20. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 1986;30(1):7-18.

21. Swanson D, Smalheiser N, Bookstein A. Information Discovery from Complementary Literatures: Categorizing Viruses as Potential Weapons. Journal of the American Society for Information Science 2001;52(10):797-813.

22. Losiewicz P, Oard DW, Kostoff RN. Textual data mining to support science and technology management. Journal of Intelligent Information Systems 2000;15(2):99-119.

23. Glantz SA, Slade J, Bero L, Hanauer P., Barnes DE. The Cigarette Papers [Online]. 1996 [cited 2002 March 4, 2002]; Available from: www.library.ucsf.edu/tobacco/cigpapers/

24. Chapman S, Cummings K. Impact of new technologies in tobacco control: call for papers. Tobacco Control 1998;7(3):222.

25. Center for Disease Control. Tobacco Industry Documents. 2002 [cited 2002 March 05, 2002]; Available from: http://www.cdc.gov/tobacco/industrydocs/index.htm

26. University of California San Francisco, American Legacy Foundation. Legacy Tobacco Documents Library. 2004 [cited 2004 June 18, 2004]; Available from: http://legacy.library.ucsf.edu/

27. University of California San Francisco. British American Tobacco Documents Archive. [cited 2005 September 5]; Available from: http://bat.library.ucsf.edu/

28. Muggli ME, LeGresley EM, Hurt RD. Big tobacco is watching: British American Tobacco's surveillance and information concealment at the Guildford depository. Lancet 2004;363(9423):1812-9.

29. UCSF Library and Center for Knowledge Management. British American Tobacco Document Collection. 2005 [cited 2005 August 28];

30. University of California San Francisco. The British-American Tobacco Document Collection. 2002 January 28, 2002 [cited 2002 January 13, 2002]; Available from: http://www.library.ucsf.edu/tobacco/batco/

31. Bero L. Implications of the tobacco industry documents for public health and policy. Annual Rev Public Health 2003;24:267-88.

32. Malone RE, Balbach ED. Tobacco industry documents: treasure trove or quagmire? Tobacco Control 2000;9(3):334-8.

33. Shortliffe E, Perreault L, Weiederhold G, Fagan L. Medical Informatics: Computer Applications in Health Care and Biomedicine. New York: Springer; 2001.

34. O'Carroll P, Yasnoff W, Ward E, Ripp L, Martin E. Public Health Informatics and Information Systems. New York: Springer; 2003.

35. Vivisimo Clustering Demonstration. 2005 [cited 2005 August 28]; Available from: http://demos.vivisimo.com/projects/medline

Chapter 2.
Conceptual Framing of the Informatics Theory

"...nicotine is addictive. We are, then, in the business of selling nicotine, an addictive drug...But cigarettes...have certain unattractive side effects: they cause, or predispose to, lung cancer, they contribute to cardiovascular disorders...they may well be truly causative in emphysema."

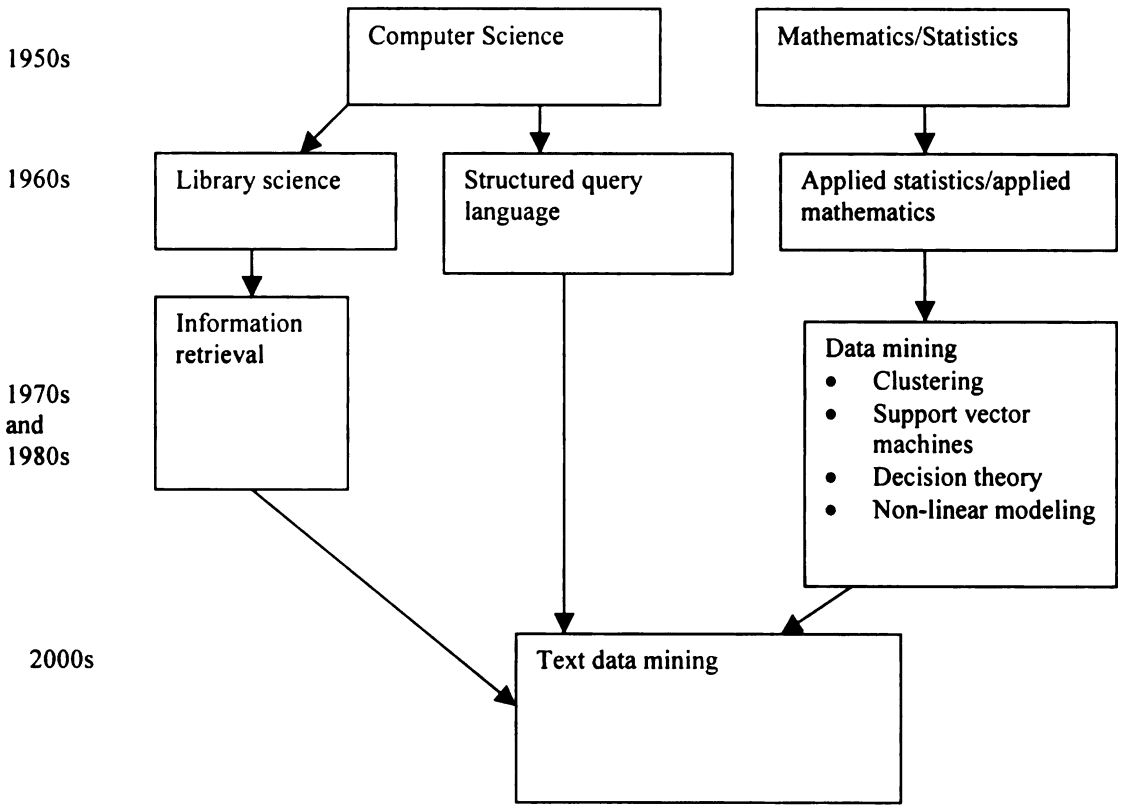
- Brown and Williamson Tobacco Company,
Internal document, 1963

A. INTRODUCTION

The purpose of this chapter is to provide the framework for the approach taken in the dissertation. The roots of the dissertation are in human computer interaction, text data mining and public health informatics. These are broad fields but can be tied together in context of the dissertation. Text mining is a relatively new field but has been exploding in the past 5 years and is a combination of data mining, statistics, information retrieval and computer science (see Figure 1; see Appendix F for definitions under text data mining). Public health informatics has been called other things in the past, such as surveillance. For this review, I will discuss the background for human computer interaction, text mining and public health informatics from the perspective of what they have to contribute to searching the internal tobacco industry documents. In addition, I will discuss the origin and scientific importance of the internal tobacco industry

documents and informatics tools developed specifically for the tobacco industry documents.

Figure 1 Fields that comprise text mining



B. HUMAN-COMPUTER INTERACTION

Human-computer interaction is a broad field within computer science that studies "...how people design, implement, and use interactive computer systems and how computers affect individuals, organizations and society (1)." Ever since the invention of computers, there have been people interested in how we interact with them.

Unfortunately, digital and human interfaces are typically not well designed and cause frustration, at best. For example, Neilsen, a usability expert, claims that it is no surprise that medical errors are traceable directly to human-computer interaction (2).

The following diagram (see Figure 2) gives an overview of all the areas in which human computer interaction can come into play and in which difficulties can arise (3). The diagram shows a person in front of a computer, a scene with which we are all familiar. However, the scene is also filled with a less familiar context, the ergonomic setup, the input device, the dialogue technique, the social organization and work, the implementation and design process. For most systems, the end user is not aware of the context and it remains hidden, which may prevent him or her from realizing why they are having trouble using the program or website. However, all of these contextual aspects are important for helping the end user work the most efficiently with the product. These potential problems and inefficiencies will be further explored in chapters 3 and 4, with respect to users' needs in accessing the tobacco industry documents.

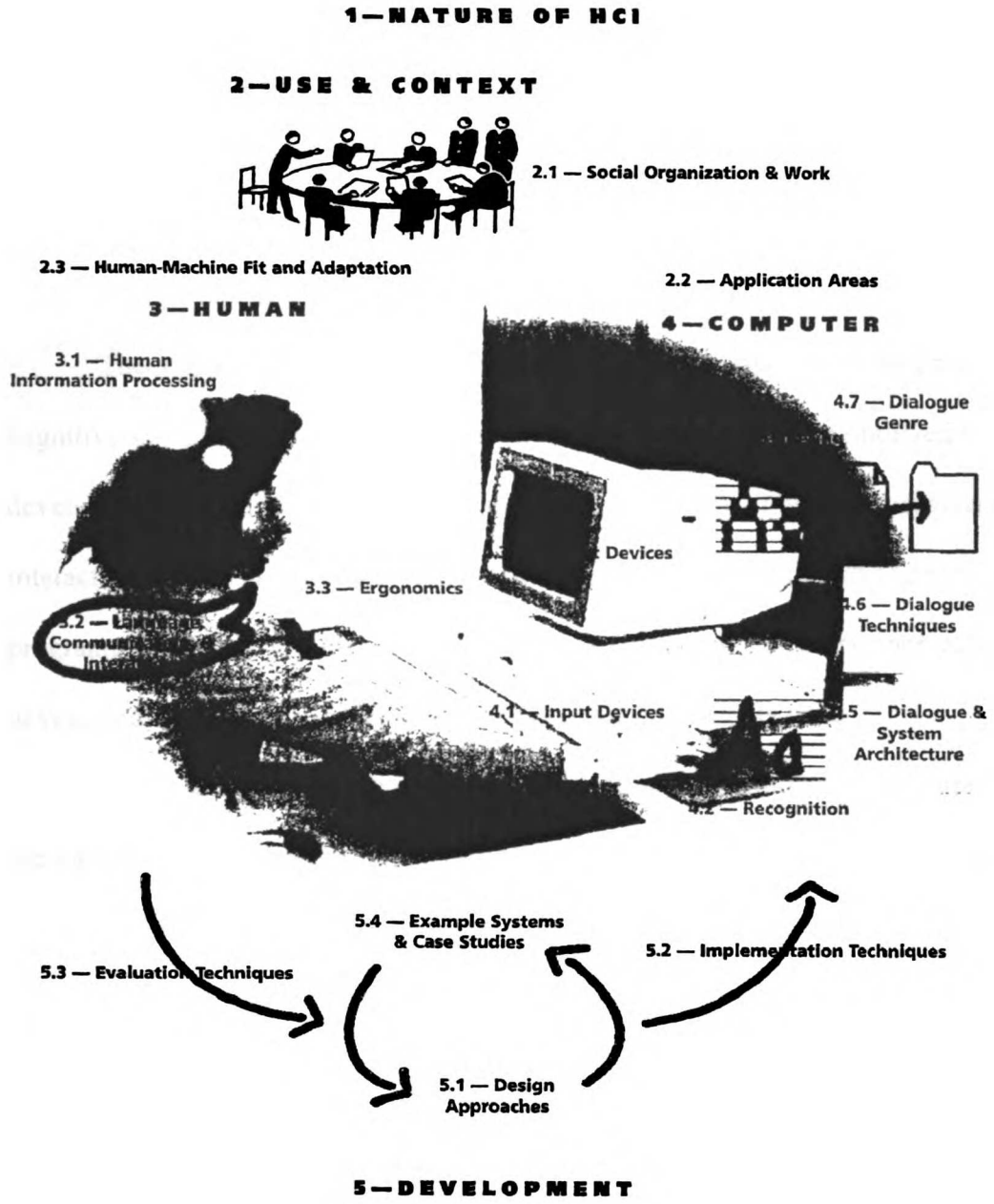


Figure 2: Human-Computer Interaction(3)

Figure 2 is taken from an article about human computer interaction written ten years ago by Dr. Gary Strong (3), showing the importance of all the component parts that are involved when a person interacts with a computer, and the context in which the user interacts with the system. In addition, the user comes to the program or computer with preconceived ideas and notions about how things should work based on prior experience. These multiple factors are the reasons that many systems fail.

Theories about how to develop computer programs were developed mainly from cognitive scientists in the 1980s(4). However, they remain largely unutilized when developing programs. Sutcliffe states an important mission of human-computer interaction (HCI) is to bring psychology and sociology into the design and utilization of programs(4). However, in practice, this theory of a cross-disciplinary approach when developing computer programs or websites remains not widely used.

Social scientists have been aware that notions of community and computers affect the view through which one interacts with a program; however, only recently have people who are HCI specialists began to focus on sociologists like Berg or Timmerman.

Berg and Wears wrote an editorial in JAMA that likens computer technology in the medical field as a bright future that never arrives (5). There are many promises when it comes to computerized systems in health care; however, an estimated 75% of information systems fail (5).

C. TOBACCO DOCUMENTS

The public became aware of the internal tobacco industry document collection in 1994 when a series of articles were published in the New York Times (6). Soon after the

articles a box of documents arrived in front of the office of Dr. Stan Glantz, anonymously labeled, "Mr. Butts". The story has been chronicled in many popular media outlets such as Frontline. Following the release of the story in the press a series of articles in the Journal of the American Medical Association were published(7-11). These articles, with additional documents were published in a book titled The Cigarette Papers(12).

The implications of the internal tobacco industry documents have been explored and highlighted in an article by Dr. Lisa Bero (13). The major findings to date of the British American Tobacco documents have been described in another article (14). This editorial also called for the public health community to act to obtain the Guildford documents before the depository closes its doors in 2008(14). The public health community did act and thanks to funding from multiple sources and work from many people, the gathering of thousands of documents has been successful (15).

D. ORIGIN AND DESCRIPTION OF TEXT MINING

Text mining (also referred to as text data mining) is a relatively new technique, derived from data mining and information retrieval, that uses statistical algorithms and user guided analysis to come up with new information and knowledge from a corpus (16). Data mining and text data mining are also known as knowledge discovery in data or databases (KDD); however, data mining usually refers to numerical data only. Both fields are interdisciplinary, using techniques from computer science, artificial intelligence, statistics and other related fields (17).

In the literature, text mining continues to mean different things to different authors, including the use of any algorithm applied to free text. Algorithms used in text

mining include support vector machines, clustering, linear discriminate analysis and others. Common problems exist in that research projects are tested and trained on a specific dataset and are thus not applicable to new data.

Text data mining can be used to discover new "nuggets" of information in a set of documents (16). It will have increasing importance to clinicians in the future as medicine becomes increasingly digital. Electronic medical records, picture archiving (e.g., CT scans and X-rays), communication systems, and the medical literature are all areas where text mining can have a significant role in the future (see Table 2).

Table 1: Text and Numerical data retrieval methods

	Finding Patterns	**Finding Nuggets**	
		Novel	**Non-Novel**
Non-textual data	Data mining	?	Database queries
Textual data	Computational linguistics	Text data mining	Information retrieval

From (16)

Table 1 shows the different types of analysis that can be conducted with text and non-text data. Finding patterns in numerical data is data mining. Data mining is a well developed field that uses different statistical algorithms to detect patterns in data (18). Following along the table, non-textual data can be queried to find pieces of "non-novel" information using database queries. The idea behind the information being non-novel is that although the information might not be currently located it does exist somewhere in the database and people know to look for it ("novel" means that nobody has yet made the

connection that the information exists). Database queries are commonly executed using a variation of structured query language or SQL.

Textual data is far more complex. Text contains intricacies, context, abbreviations, double-meanings, and other features that make it more difficult to be able to apply straight out data mining algorithms in order to perform text mining algorithms. Finding patterns in textual data is considered computational linguistics, which is a rich field from which text mining derives many methods (19).

E. APPLICATIONS OF TEXT MINING

Since text mining has developed substantially since it was created in the early 90s, there have been attempts to apply text mining to real data. Much of the early text mining work was done in literature text data mining. However, people have started to apply text mining to other data sets or corpii. I will cover the key articles that have successfully applied text mining for cutting edge research and how text mining can be used in searching the tobacco documents.

There have been a number of applied research articles that have used text mining to detect trends in the literature (20-32). Viator and Pestorius analyzed acoustics research citations from the years 1970, 1980, 1990 and 1999 using the INSPEC database (28). INSPEC is a database that indexes articles in physics, computer science and electrical engineering. The authors retrieved articles published in JASA, the Journal of Acoustical Society of America to determine research trends over the past 30 years. They used software developed at Georgia Tech called the Technology Opportunities Analysis of Scientific Information System (Tech OASIS) (33). This software uses cluster analysis to

create similar groups and measures internal and external quality. Oasis (later, Vantage Point), which is the commercial project, is based on an innovative forecasting technique. The results gave an overview of articles with United States versus non-United States affiliatied authors or institutions, research areas by year, by world region, and the breadth of coverage in JASA compared to other journal publishers (28). This analysis of research trends provides different information than can be obtained by simple tabulating software, since it exposes patterns that might be unanticipated or unexpected.

One of the more interesting applications of text mining in the biomedical literature was by Swanson who derived the hypothesis that fish oil can be used for the treatment of Raynaud's syndrome. He defined an application of text mining, before text mining had even been thought up. He states in his research "Scientific articles can be seen as clustering into more or less independent sets of 'literatures'". Before text mining, which is connecting documents based on statistical natural language processing, Swanson was considering literature as a group of independent ideas or islands that were "logically related" and may contain connections that would remain "unintended, unnoticed, and unknown".

Table 2: Future applications of text mining in research and clinical settings

Category	**Examples**
Research	• Discover relations between genes, proteins, metabolic pathways etc. • Generate new hypotheses in the medical and biological literature • Uncover protein-protein interactions
Clinical	• Mine free text electronic medical records within and between patients
Emergency medicine and	• Use in conjunction with a decision support system to

general practice	diagnose uncommon diseases
Public health	• Outbreak detection • Symptom surveillance
Ambulatory care	• Use to synthesize symptoms from an EMR and draw on other patient experiences for machine hypothesis and learning

Table 3 shows potential future applications for text mining specifically for the internal tobacco industry documents. Text mining could be used for automatic categorization, document summarization, and concept abstraction. Some researchers do not consider text categorization as text mining. However the rationale is that somebody has read the document and knows how to categorize it. In the case of the internal tobacco industry documents, the knowledge is in the public domain and we are in effect deriving new information by automatically categorizing documents. The other applications of text mining such as text summarization would be tremendously useful because some of the documents are very long. Clustering would be helpful for the internal documents because they could cluster similar projects or topics together. It could help detect name variations to deal with the problem of uncertain name recognition.

Table 3: Future applications of text mining in internal tobacco documents research

Category	**Algorithm**	**Examples**
Clustering	K-means algorithm	Similar project topics together (e.g. Project alpha/Premier, environmental tobacco smoke (ETS)) Similar characteristics together (e.g. non-smokers, ETS chamber, use of employees) Variations in names together
Automatic Categorization	Support vector machines Naïve Bayesian classifiers Support vector machines	Categorization of the full text of documents based on the Tobacco thesaurus
Connection discovery	K-means algorithm	Observe proximity of different categories of documents

F. CONCLUSIONS ON TEXT MINING

While text mining is no longer a nascent field, reports of its use are still relegated to obscure computer science journals and conference proceedings, and remain largely out of the public eye; it also continues to be slow in reaching applied academic fields. It has the potential to influence biomedical research and patient care from bench to bedside. However, creating more awareness of how text mining can be used practically is necessary to encourage the substantial investment from hospitals, private practice and health maintenance organizations.

The barriers to carrying out text mining in a non-research institution are substantial; however, they are worth attempting to overcome. We need informaticists and

physicians who can translate research into practice from computer science into patient care and public health knowledge.

Text mining currently has a number of software products, commercial and open source, which can be used. Examples such as SAS Text Miner, SPSS Clementine, and IBM are the leaders commercially. Open source products include Natural Language Tool Kit in a programming language called Python. The program can be adapted to the users' needs.

However, text data mining is not a panacea. It, as other new technologies, should be approached with caution. As in numerical data mining problems, the analyst must pay particular attention to the increased potential of spurious results and not place too much "faith in the black box" (34). The black box will lead to incorrect conclusions if the results are not further investigated in a rigorous manner. Another concern with numerical data mining, which is valid with text mining as well, is that the analyst will find an inappropriate model to fit the data, rather than examine the data and carefully consider whether the underlying assumptions of a particular distribution are true.

With regard to this dissertation research, I explore the opportunities of text mining of the tobacco documents. Table 3 presents some of these possibilities. Although some of the examples require further research, others are investigated in Chapter 5.

G. CONCEPTUAL FRAMING – PEOPLE-CENTERED PUBLIC HEALTH INFORMATICS

The framing of the dissertation exists to put the work in context of other work. The theoretical basis encapsulates the reason for approaching the work in the manner

seen. I would like to take time to explore why people centered public health informatics is different from human-computer interaction, text mining, and public-health informatics. I am interested in changing the unit of knowledge (the document) by making it accessible and able to be used and re-used in different ways. The more accessible the documents are to all groups, the more the unit of knowledge will be utilized in different ways.

People who search the documents are experienced with finding existing knowledge and then creating new knowledge using detective work to put together a story. In general, university faculty and staff who work in tobacco control are knowledge creators since journal articles and talks are their primary product. People in non-profit public health organizations are mainly applying knowledge to create interventions, but are also creating new knowledge by doing new studies. In order to make the tobacco documents accessible to all, different interfaces and tools for searching need to be continually investigated. However, to save time and money, applying principles of people-centered public health informatics may help.

How is a people-centric theory of informatics different from other theories? The people centric theory uses tools from other disciplines to first asses the informatics need in the population. Sometimes the need will be less than what was originally planned. Previous articles have demonstrated how information technology projects fail and why. The reasons are as varied as there are theorists, but the one commonality is a lack of needs assessment.

1. Develop needs assessment tools – The users of a system do not always know what they want, but usually the users have good ideas. The users spend the most time with the system and it is important to develop tools to assess how the users plan to use the

system. The development of surveys is an undervalued skill and often the survey does not elicit what the administrators of the survey really wanted to know. It is crucial to spend time on the development of good questions and pilot test before the survey is released to the public. The same is true for interviews.

2. Assess – The importance of assessing needs of the existing and potential users can be done in a variety of ways, such as survey and interview research. Other ways are by observing the user performing tasks on the existing system.

3. Develop or test new research tools – Based on the analysis of the assessment, tools should be developed for different user groups. It is likely that people have different needs and wants from a website or software.

This process should be repeated to make sure that needs and wishes have not changed for the users. A people-centered public health informatics strategy takes necessities and wishes of the users seriously, but also looks to save resources.

H. REFERENCES

1. Myers B, Hollan J, Cruz I, Bryson S, Bulterman D, Catarci T, et al. Strategic directions in human-computer interaction. ACM Computing Surveys 1996;28(4):794-809.

2. Nielsen J. Medical Usability: How to Kill Patients through Bad Design. 2005 [cited 2005 June 16]; Available from: http://www.useit.com/alertbox/20050411.html

3. Strong G. New directions in human-computer interaction: education, research, and practice. Interactions 1995;2(1):69-81.

4. Sutcliffe A. On the effective use and reuse of HCI knowledge. ACM Transactions on Computer-Human Interaction 2000;7(2):197-221.

5. Wears RL, Berg M. Computer Technology and Clinical Work: Still Waiting for Godot. JAMA 2005;293(10):1261-1263.

6. Hilts PJ. Tobacco Company Was Silent on Hazards. New York Times 1994 May 7th;Sect. 1.

7. Bero L, Barnes DE, Hanauer P, Slade J, Glantz SA. Lawyer Control of the Tobacco Industry's External Research-Program - the Brown-and-Williamson Documents. Jama-Journal of the American Medical Association 1995;274(3):241-247.

8. Glantz SA, Barnes DE, Bero L, Hanauer P, Slade J. Looking through a Keyhole at the Tobacco Industry - the Brown-and-Williamson Documents. Jama-Journal of the American Medical Association 1995;274(3):219-224.

9. Hanauer P, Slade J, Barnes DE, Bero L, Glantz SA. Lawyer control of internal scientific research to protect against products liability lawsuits. The Brown and Williamson documents. Jama 1995;274(3):234-40.

10. Slade J, Bero LA, Hanauer P, Barnes DE, Glantz SA. Nicotine and addiction. The Brown and Williamson documents. Jama 1995;274(3):225-33.

11. Barnes DE, Hanauer P, Slade J, Bero LA, Glantz SA. Environmental tobacco smoke. The Brown and Williamson documents. Jama 1995;274(3):248-53.

12. Glantz SA, Slade J, Bero L, Hanauer P., Barnes DE. The Cigarette Papers [Online]. 1996 [cited 2002 March 4, 2002]; Available from: www.library.ucsf.edu/tobacco/cigpapers/

13. Bero L. Implications of the tobacco industry documents for public health and policy. Annual Rev Public Health 2003;24:267-88.

14. Lee K, Gilmore A, Collin J. Looking inside the tobacco industry: revealing insights from the Guildford Depository. Addiction 2004;99(4):394-397.

15. The British American Tobacco Documents Archive. [cited 2005 September 1]; Available from: http://bat.library.ucsf.edu/index.html

16. Hearst M. Untangling Text Data Mining. In: Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics; 1999 June 20-26; University of Maryland; 1999.

17. Goodwin L, VanDyne M, Lin S, Talbert S. Data mining issues and opportunities for building nursing knowledge. J Biomed Inform 2003;36(4-5):379-88.

18. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag; 2001.

19. Manning C, Schutze H. Foundations of Statistical Natural Language Processing. Boston: Massachusetts Institute of Technology; 1999.

20. Albert S, Gaudan S, Knigge H, Raetsch A, Delgado A, Huhse B, et al. Computer-assisted generation of a protein-interaction database for nuclear receptors. Mol Endocrinol 2003.

21. Arimura H, Abe J, Fujino R, Sakamoto H, Shimozono S, Arikawa S, et al. Text data mining: discovery of important keywords in the cyberspace. In: International Conference on Digital Libraries: Research and Practice; 2000; Los Alamitos, CA, USA: IEEE; 2000. p. 220-226.

22. de Bruijn B, Martin J. Getting to the (c)ore of knowledge: mining biomedical literature. Int J Med Inf 2002;67(1-3):7-18.

23. Ding J, Berleant D, Nettleton D, Wurtele E. Mining MEDLINE: abstracts, sentences, or phrases? Pac Symp Biocomput 2002:326-37.

24. Dugas M, Hoffmann E, Janko S, Hahnewald S, Matis T, Uberla K. XML-based visual data mining in medicine. Medinfo 2001;10(Pt 2):1324-8.

25. Kostoff RN, DeMarco RA. Extracting information from the literature by text mining. Anal Chem 2001;73(13):370A-378A.

26. Mack R, Hehenberger M. Text-based knowledge discovery: search and mining of life-sciences documents. Drug Discov Today 2002;7(11):S89-98.

27. Losiewicz P, Oard DW, Kostoff RN. Textual data mining to support science and technology management. Journal of Intelligent Information Systems 2000;15(2):99-119.

28. Viator JA, Pestorius FM. Investigating trends in acoustics research from 1970-1999. J Acoust Soc Am 2001;109(5 Pt 1):1779-83.

29. Witten IH, Bray Z, Mahoui M, Teahan B. Text mining: a new frontier for lossless compression. Data Compression Conference 1999;29-31:198 -207.

30. Yamanishi K, Li H. Mining open answers in questionnaire data. IEEE Intelligent Systems 2002;17(5):58 - 63.

31. Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques 1999;27(6):1210-4, 1216-7.

32. Srinivasan P. MeSHmap: a text mining tool for MEDLINE. Proc AMIA Symp 2001:642-6.

33. Watts R, Porter A, Zhu D. Factor Analysis Optimization: Applied on Natural Language Knowledge Discovery. [White Paper] [cited 2003 July 15]; Available from: http://www.tpac.gatech.edu/public_papers/FA_Opt2-abs.shtml

34. Dasu T, Johnson T. Exploratory Data Mining and Data Cleaning. New Jersy: Wiley-Interscience; 2003.

CHAPTER 3.
ASSESSING USERS' NEEDS: SURVEY OF THE LEGACY TOBACCO DOCUMENTS AND THE TOBACCO CONTROL ARCHIVES USERS

In their own words:
"Today's teenager is tomorrow's potential regular customer and the overwhelming majority of smokers first begin to smoke while still in their teens."

-Internal document, Philip Morris 1981(1)

A. INTRODUCTION

The objective of this portion of the research project is to study who uses the internal tobacco industry documents and for what purposes they are used. By understanding how and why the documents are used, we can design better search engines and increase strategies for successful searches

In 1994, the release of previously secret internal tobacco industry documents produced several analyses that changed public health (2). Recent litigation and the Master Settlement Agreement of 1998 (3) made millions of tobacco industry internal documents available on the Internet (4). Before the Master Settlement Agreement (MSA), researchers relied on observational studies to examine the behavior of the tobacco industry (2, 5). The internal tobacco industry documents continue to be released, and have given the public health community an unprecedented glimpse into the inner workings and behavior of the industry (2).

As required by the MSA, United States tobacco companies maintain websites where many of the internal company documents can be accessed. Government and private funding sources support the archiving and indexing of these documents on other more stable and permanent Internet sites, such as Tobacco Documents Online

(www.tobaccodocuments.org) and the Legacy Tobacco Documents Library (legacy.library.ucsf.edu) (6-8). Housed at the University of California, San Francisco, the Legacy Tobacco Documents Collection (LTDL) collection, as of June 18, 2004, contained 6.9 million documents with an estimated 41 million pages (6). Funding agencies responsible for the establishment of this online collection include the California Tobacco Related Disease Research program, the National Cancer Institute tobacco document initiative and the American Legacy Foundation.

Table 1 shows characteristics of the different collections. The survey was conducted on users of the Legacy Tobacco Documents Collection and specific collections in the Tobacco Control Archives including: the Brown and Williamson collection, the Joe Camel Collection, the CA documents from the Minnesota depository and the British American Tobacco Company document subset. For more information about the individual collections, see Chapter 2.

Table 1: UCSF Internal Tobacco Document Collections

	Legacy Tobacco Documents Collection	Tobacco Control Archives	British American Tobacco Company	British American Tobacco Document Archive
Years	1950-2003			
Collections	• American Tobacco • Brown & Williamson • Center for Tobacco Research • Lorillard • Philip Morris • RJ Reynolds • Tobacco Institute • Tobacco Depositions and Trial Testimony Archive (DATTA) • Mangini ("Joe Camel") Documents • UCSF Brown & Williamson	• Brown and Williamson collection • Joe Camel Collection • CA documents from MN depository • British American Tobacco Documents	Specialized set from the document depository in Guildford, England.	Will contain all documents from the depository in Guildford, England pertaining to the British American Tobacco Company
Indexing	Provided by individual tobacco companies, given to UCSF via National Association of Attorneys General (NAAG)	Provided by individual tobacco companies, given to UCSF via NAAG.	Indexed manually based on Tobacco Thesaurus (see appendix F)	Full text indexing
Number of documents	7,191,091	17,250 BATCO 4,000 Joe Camel	17,500	718,964
Number of pages	42,325,613	80,000 Joe Camel		2,658,193
Notes		This collection is now contained in Legacy except for BATCo		Currently being obtained from Guildford, and put online

Despite the considerable effort and resources that have been invested into the establishment and maintenance of the online collections of internal tobacco industry documents, there have been no studies of who searches the document collections or why they search. Since the internal tobacco industry documents are freely available, they can be accessed by the public, researchers, advocates, lawyers, tobacco industry personnel,

and others. In addition, the online document sites maintained by the tobacco companies will be available only until 2010. As users will eventually have access only to online tobacco industry document collections maintained by university and non-profit websites, it is important to know how to design the most accessible and sustainable archives for the documents.

Our purpose for this study was to conduct an on-line survey of users of the Legacy Tobacco Documents Library (LTDL) and the Tobacco Control Archives (TCA) to determine: 1) who uses the internal tobacco industry documents, 2) why the documents are used, 3) barriers to searching and 4) suggestions for improvement. With respect to the LTDL, we also determined the characteristics of the users.

B. METHODS

We posted an online survey on the Legacy Tobacco Documents Library and the Tobacco Control Archives using software from Web Surveyor (www.websurveyor.com). In brief, users accessing the archive were offered a link asking them to participate in the on-line survey. Written consent was waived since participation in the survey indicated consent. Our protocol was approved by the UCSF Committee on Human Research (#H2758-18553-02). Personal identifying information (such as names) was not collected, nor was it derived from the electronic access logs. The surveys for both the Legacy Tobacco Documents Library and the Tobacco Control Archives (TCA) took users approximately 5 minutes each to complete.

We mounted the survey on the Legacy and Tobacco Control Archives website for one year (October 2002- September 2003). Many of the same questions were asked for the Tobacco Control Archives and for the Legacy survey, except that the Legacy survey had an additional section on computer experience and demographics (see appendix B for the surveys). Both surveys were developed in collaboration with UCSF library personnel. We decided to make the TCA survey shorter than the Legacy survey to encourage more responses. To develop the surveys, we conducted a pilot test of our draft survey on the site. We eliminated unreliable or invalid survey questions.

The survey asked about three broad areas: searching the internal tobacco industry documents, computer experience and demographics of users of the site. The first set of questions referred to the use of the archive, such as frequency of use, purpose of accessing the archives, and barriers to use. The next section of the survey related to computer experience and usage. The final section of the survey inquired about the demographics of the survey respondent such as their age, gender, geographic location, and occupation. All questions were fixed-choice format, with an option to add comments (See Appendices A and B).

We exported the data from Web Surveyor into SAS 8.0 and Stata 7.0 for analysis (9). We analyzed the survey results by doing frequencies and chi-square texts.

C. RESULTS FROM THE LEGACY TOBACCO DOCUMENTS LIBRARY SURVEY

1. Demographics

Table 1 shows the gender, age, occupation, and country of origin of the respondents. There were 165 respondents to the LTDL survey. Sixty-five percent of respondents were females (see Figure 16). The median age for survey respondents was 32 and the mean was 34. About 20% of respondents were under 20, 24% were 20-29, 16% were 30-39, 20% were 40-49, and 19% were 50 or older (see Figure 17). Most of respondents' ethnicities were non-Hispanic (91%), where 9% identified as Hispanic (see Figure 18). Finally, 72% identified their race as white or Caucasian (see Figure 19). Fourteen percent self-identified as other; 7% as African-American; 4% as Asian, and 2% as American Indian/Alaskan Native, Native Hawaiian or other Pacific Islander.

Most of the respondents were highly educated and had completed some college or greater. Twenty-six percent were college graduates while 23% had some college, 15% had Master's degrees, and 10% had PhDs or MDs. Five percent preferred not to answer this question.

As shown in Table 1, a variety of people searched the documents. A large proportion (25%) of respondents selected the "other" category for occupation. This category included responses from individuals such as a janitor, a psychotherapist, a truck driver, and a federal inmate. Twenty-eight percent selected student as their occupation and 14% were university or college employees. About 6% identified themselves as tobacco-control advocates, 5% were school employees, and 4% each were from industry, journalism, or law-related professions. Finally, 4% were public health officials and 3% were from community non-profit organizations.

People from a variety of countries searched the documents but the majority was from the United States (Table 1). Respondents from sixteen other countries besides the United States answered the survey with the most respondents coming from Canada (7%). Based on a website log analysis, we found that users from 192 countries accessed the documents. While 21% of respondents were from California, there were 29 other states where respondents lived.

Table 2: Demographics of Legacy survey respondents (n=165).

Age	**Age Range (median age = 34)**	**Percent**
	1-9	<1
	10-19	20
	20-29	24
	30-39	16
	40-49	20
	50-59	14
	60+	6
Gender	**Category**	**Percent**
	Male	35
	Female	65
Education	**Category**	**Percent**
	Less than High School	13
	High School grad	7
	Some College	23

	College grad	26
	Master's degree	15
	PhD/MD	11
Occupation	**Category**	**Percent**
	Students	28
	University employees	14
	Tobacco control advocate	6
	School employees (K-12)	5
	Law-related	4
	Government-related	4
	Industry	4
	Journalist	4
	Public health official	4
	Community non-profit	3
	Other	25
Place of origin	Canada	7
	UK	3
	Other (14 countries)	13
	US	77
	California	21
	New York	5
	Washington	5

	Other (25 states)	69
Note: Due to rounding, percents may not add to 100.		

2. Computer Experience

Most of respondents have a computer at home (87%); however, 10% reported that they did not, and 3% were not sure. In addition, most respondents (79%) had a computer at work. Finally, the majority of respondents reported that they are very good or excellent at finding topics of interest on the Internet (74%). Sixteen percent reported that they are good at finding topics, 7% reported they are fair, and 3% reported they are poor or have little or no experience. The people who searched the documents conducted a fair number of searches on the Internet. In fact, 64% reported that they conducted over 20 searches or more monthly, whereas 17% said they conducted 11-20 searches, 9% 6-10 searches, and 11% conducted fewer than 5 searches per month.

3. Experiences with Searching the Documents

Table 2 shows the main reason people reported for searching the Legacy Tobacco Documents Library, the percent of research sponsored by a grant and the problems they identified with searching the documents. Most respondents reported that they searched the documents for personal interest (45%) followed closely by academic research (41%). Using the documents for a teaching tool, public health advocacy and litigation were also important reasons at 18, 7, and 18 percents respectively.

Forty-one percent of respondents said they had problems searching the documents. The most common problems were not being able to find documents (10%) and search capabilities not flexible enough (10%).

People reported that they were more likely to search the documents from home followed by their work locations or a library (Figure 2). The respondents reported that they were most likely to search the LTDL site monthly (63%) (Figure 4).

Another interesting finding was that people were most likely to find the LTDL site by a link from other websites or a search engine (Figure 5). This would suggest that one of the best ways to continue to draw people to searching the site would be to make sure the site comes up in most search engines with common search terms and to submit the site as a link to tobacco control pages. The reasons why the LTDL resource is used were reported as the search capabilities (35%), ease of use (33%), and most complete resource (30%) (Figure 6). The most useful features of searching the documents included searching different tobacco industry collections at the same time (46%), the advanced search features (26%), and the ability to browse by search term (24%) (see Figure 8). Interestingly, the ability to browse by search term is not offered on LTDL but is offered in searching BATCo documents. The BATCo documents were categorized by librarians and organized by topic based on the Tobacco Thesaurus (See Appendix 4). However, this does suggest that if browsing by search term were offered by Legacy, it would be a popular feature.

Finally, 41% of respondents reported that they had problems searching the website (See figure 9 and Table 3). The problems that the users had included that they could not find documents, search capabilities were not flexible, retrieved too many

documents and had a slow connection (see Figure 9 and Table 3). Most users reported that they found what they were looking for (41%) (see Figure 15). Also, most respondents reported that bookbag features, which allow users to save documents by emailing them or adding them to Endnote, were the most difficult features to use (23%) (see Figure 11).

The attributes that users report as being most useful when searching the documents include organization names (41%), cigarette brands (40%), dates (39%), authors (32%), personal names (28%) and Bates number (18%) (see figure 12). The respondents reported that they organized documents most by theme (26%), search terms (25%), and date (23%) and others (11%) (see Figure 14).

Table 3: Reasons and barriers to searching the Legacy Tobacco Documents Library.

	Legacy Survey (n=165)
Reason for searching	**N(%)**
Academic Research	68(41)
Personal Interest	74(45)
Teaching Tool	15(9)
Public Health Advocacy	30(18)
Litigation or legal research	12(7)
Other	30(18)
Do NCI, TRDRP, ACS, or American Legacy Foundation fund any of the research you are doing?	**N(%)**
No	101(61)
Yes	53(32)
Chi square = 0.01 Pr>chi2 = 0.9036	
Did you have any problems searching the website?	**N(%)**
No	89(54)
Yes	68(41)
Chi square = 2.53 Pr>chi2 = 0.1116	

If yes, problem listed:	
Can't find documents	17(10)
Search capabilities not flexible	17(10)
Received too many documents	9(5)
Slow connection	4(7)
Site was down	2(3)
Other	17(10)
Note: Percentages may sum to > 100% because respondents could check more than one item.	

D. RESULTS FROM THE TOBACCO CONTROL ARCHIVES SURVEY

This is a reminder to the reader that the TCA survey was shorter and does not contain sections on demographics or computer experience. There were 99 respondents.

1. Respondents' Purpose for Searching

One of the main questions we were interested in was the purpose for searching. Fifty-three percent responded academic research was their purpose for searching, followed by 31% that responded other, and 25% that answered that they searched the documents for personal interest. To a lesser degree, the documents were utilized for teaching (16%), advocacy (9%) and legal research (6%). Even further behind was for media, testimony at a public hearing, and advertising campaigns.

2. Funding Source

In contrast to Legacy, 28% of those who searched the TCA archives were doing work funded by a granting agency, which included the National Cancer Institute (4%), the American Legacy Association (4%), the American Heart Association (2%), and other (14%). Most of the respondents said they searched from home (43%) and other (40%); although a significant proportion also searched through work (17%), and the public

library (15%). The respondents who selected other reported mainly that they were searching from a school.

Most people reported that in the last six months, this was their first visit to the Tobacco Control Archives (64%). However, 7% reported that they visited the site weekly, and 4 percent responded that they visited daily, 4 percent monthly and 4 percent less than monthly. Respondents reported that they found the Tobacco Control Archives using a search engine (35%), via a link (25%), colleague (9%), citation in an article (8%), or heard of the site through a newspaper article (5%). Other responses (25%) included a school assignment.

3. Experiences with Searching the Documents

We asked respondents why they used the Tobacco Control Archives. The most frequent response was that people did not know any other place to find documents (31%). However, search capabilities (28%), ease of use (16%), most complete tobacco documents resource (14%), speed of searching (11%), helpful index and abstracts (10%), stability of website (9%), and familiarity (2%) were other responses.

Also, we asked respondents what other websites they used to search the documents. They were as follows, 16% replied that they used Tobacco Documents Online, 15% said they used tobacco industry document websites, 13% reported the Center of Disease Control site online (See Chapter 1 for a complete list of internal tobacco documents websites).

Finally, most users searching the Tobacco Control Archives reported that they did not have any problems (67%). However, 19% reported having some problems, which

included not being able to find documents, a slow connection, site was down, and search capabilities not flexible enough. Most respondents reported that the browse feature was one of the most useful features of the Tobacco Control Archives (27%). At the same time, many respondents were not aware of this option (46% each for the Joe Camel collection, the British American Tobacco documents, and the Brown and Williamson documents).

In order to inform the creation and organization of future websites, we were interested in learning how people organized the search results that they found. There were a variety of ways that people organized the documents: by theme (24%), by document title (21%), by search terms and named person (11% each), by named organization and date (9% each) and by document type (e.g., memo, report, etc.) and other method (7% each).

We also asked respondents about the success of their searches. Twenty-six percent of people said that they found what they were looking for, and 36% responded that they were just browsing. However, 24% responded that they did not know if their search was successful and 15% responded "No", they did not find what they were looking for.

Table 4: Reasons and barriers to searching the Tobacco Control Archive (n=99).

Reason for searching	**N(%)**
Academic Research	52(53)
Personal Interest	25(25)
Teaching Tool	16(16)
Public Health Advocacy	9(9)
Litigation or legal research	8(8)
Other	31(31)
Do NCI, TRDRP, ACS, or American Legacy Foundation fund any of the research you are doing?*	**N(%)**
No	65(66)
Yes	33(33)
Chi square = 0.01 Pr>chi2 = 0.9036	
Did you have any problems searching the website?	**N(%)**
No	66(67)
Yes	33(33)
Chi square = 2.53 Pr>chi2 = 0.1116	
If yes, problem listed:	
Can't find documents	5(5)
Search capabilities not flexible	2(2)
Received too many documents	-
Slow connection	3(3)
Site was down	3(3)
Other	6(6)

Note: Percentages may sum to > 100% because respondents could check more than one item.	

E. COMPARISON OF LEGACY AND TCA SURVEYS

To further explore the differences between user responses to the Legacy and TCA sites, we conducted comparisons.

The primary Internet connection and speed from which people searched the documents is shown below in Table 5 and see Figure 3. There was a statistically significant difference between the types of Internet connection the user had and which site they searched. It appears that those who searched the TCA sites were less likely to know the speed of their Internet access point. In addition, Legacy survey respondents are more likely to use DSL than TCA users.

Table 5: Internet access speed for 2 different tobacco document archive sites*

	Tobacco Control Archives N(%)	**Legacy N(%)**
Don't know	38(44)	48(32)
28.8-56K	17(20)	24(16)
14.4K modem	3(3)	2(1)
Cable Modem/DSL	11(13)	49(32)
T1 or faster	12(14)	18(12)
ISDN	2(2)	9(6)
Other	4(5)	1(1)
Total	**100(87)**	**100(151)**

Pearson's Chi square = 18.1839 Pr = 0.006
*Due to rounding, percentages may not sum to 100.

Another interesting difference between the respondents at Legacy compared to TCA is the difference listed in reason for searching (Table 6). Even though it appears that people were more likely to use TCA than Legacy for academic research, there is no difference between those using TCA or Legacy for getting funding from Tobacco Related Disease Program, American Cancer Society, National Cancer Institute or the American Legacy grants (Table 6 and see Figure 1). There was no statistically significant difference between the proportions of users who experienced problems (Table 6).

Table 6: Reasons and barriers to searching the Legacy Tobacco Documents Library compared to the Tobacco Control Archives.

	Legacy Survey (n=165)	TCA survey (n=99)
Reason for searching*	%(N)	%(N)
Academic Research	68(41)	52(53)
Personal Interest	74(45)	25(25)
Teaching Tool	15(9)	16(16)
Public Health Advocacy	30(18)	9(9)
Litigation or legal research	12(7)	8(8)
Other	30(18)	31(31)
Pearson chi square = 18.6685 Pr = 0.002		
Do the National Cancer Institute, Tobacco Related Disease Research Program, American Cancer Society, or American Legacy Foundation fund any of the research you are doing?	N(%)	N(%)
No	101(61)	65(66)
Yes	53(32)	33(33)
Pearson's chi square= 0.01 Pr>chi2 = 0.9036		
Did you have any problems searching the website?	N(%)	N(%)
No	89(54)	66(67)
Yes	68(41)	33(33)
Pearson's chi square= 2.53 Pr>chi2 = 0.1116		
If yes, problem listed:		
Can't find documents	17(10)	5(5)

Search capabilities not flexible	17(10)	2(2)
Received too many documents	9(5)	-
Slow connection	4(7)	3(3)
Site was down	2(3)	3(3)
Other	17(10)	6(6)
Note: Percentages may sum to > 100% because respondents could check more than one item.		

F. CONCLUSION

Our survey results show that many people search the internal tobacco industry documents for academic research and personal interest. However, educators, tobacco control advocates and lawyers appear to be untapped potential user populations.

Some technical barriers may continue to impede people from searching the documents; about 40% said they had some problem searching the documents. These barriers are not necessarily due to limitations of the Legacy or the Tobacco Control archives site. A slow connection, for example, could be on the user's end. However, the user typically does not care where the problems arise. The user is less likely to use the interface if any problems arise, from user's location or the site's part.

However, we found that there was a significant difference in TCA and Legacy respondents in the type of Internet connection users had to access the sites. TCA survey respondents were more likely to have a slower connection or not know what their connection speed was. This finding has a practical application when it comes to

designing sites for slower access speed. The UCSF library may consider using fewer graphics with the TCA related sites.

The internal tobacco industry documents continue to be a useful resource for researchers and lay people but access and searching need improvements to reach to other user populations. Following are various suggestions we propose (in addition, see Appendix C). New website interfaces and search algorithms could incorporate creative ways to attract new users. For example, teachers could establish contests to find the most outrageous quote from a tobacco industry document and use the documents for targeted lessons on topics such as tobacco advertising towards women and children. Tobacco control advocates could search for documents specific to counter-advertising campaigns, as has been done for the Truth campaign (10).

The survey results suggest some techniques that could be useful for improving searching of the documents, including: web-based search interfaces, customizing the documents, and text data mining. Web-based search interfaces that extract and use metadata to display the search results will increase user effectiveness in searching. Additional metadata will allow the user to search through more search fields and eventually allow browsing of the documents (see Appendix C). Customization of the documents will allow user groups that are concerned with subsets of documents to search only in areas of interest to their issues. For example, an advocate working at Americans' for Nonsmokers Rights may be interested in documents to support a smoke-free bars campaign. Finally, text data mining can be used for exploratory text data analysis about a subject using a combination of computational algorithms and user-guided analysis (11).

There are some limitations to this study. First, we could not conduct an online survey of all the internal tobacco industry documents sites, such as Tobacco Documents Online or the sites run by the tobacco industry, like Phillip Morris and RJR. Second, we cannot eliminate the possibility that response bias could affect our results. We were not able to collect extensive baseline data to get an accurate response rate. Recent research on online surveys has suggested that offering monetary or gift incentives, and verbally recruiting interested participants could help increase response rates for online surveys (12). Keeping the survey short and extensively pilot testing the survey, as we did, also helped increase response rates for online surveys (12).

There is always room for improvement; specifically creating targeted resources for advocates and teachers and developing new search technologies. The results of this survey offer an opportunity to tailor the sites to meet the needs of a wider range of user groups.

Figure 1: LTDL Survey - Are you doing research that is funded by a grant?

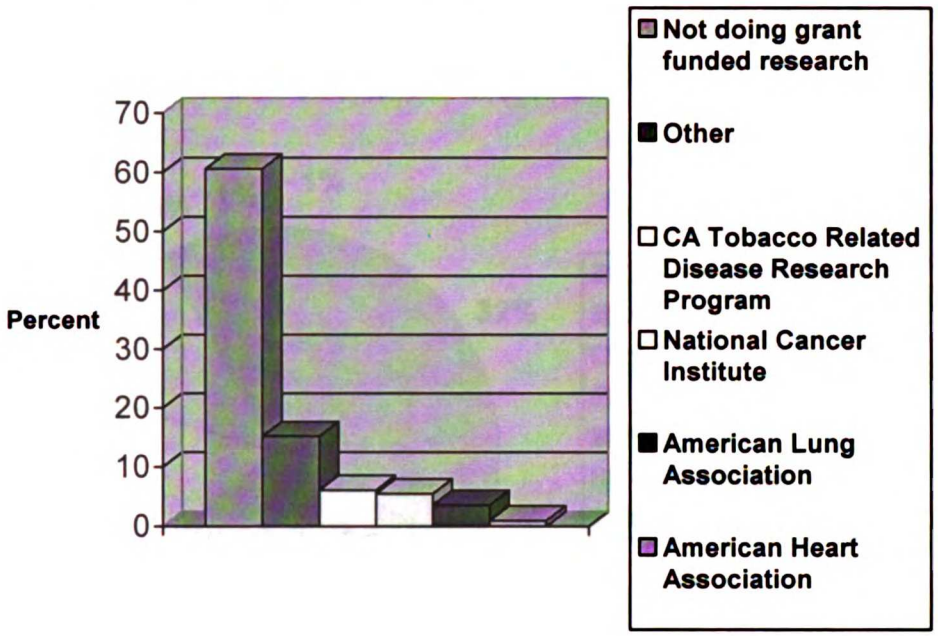


Figure 2: LTDL Survey - Where do you search the documents?

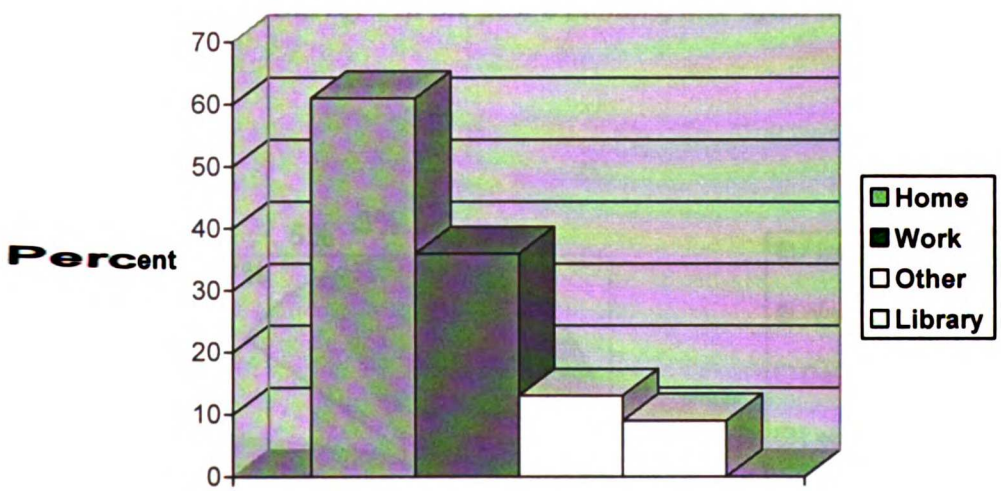


***Figure** 3: LTDL Survey - What speed is the Internet connection that you most **frequ**ently use to access the documents?*

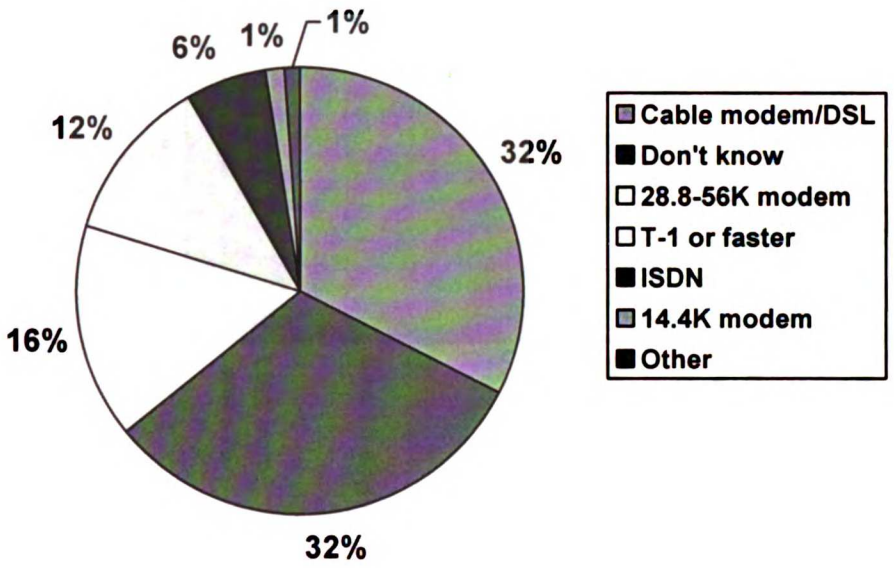


***Figure** 4: LTDL Survey - In the past 6 months, how often have you accessed the **docum**ents?*

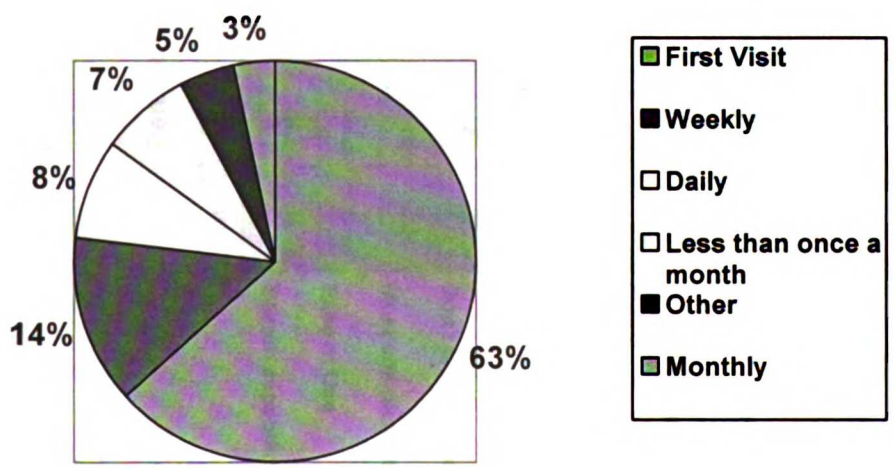


Figure 5: LTDL Survey - How did you find the Legacy Tobacco Documents Library?

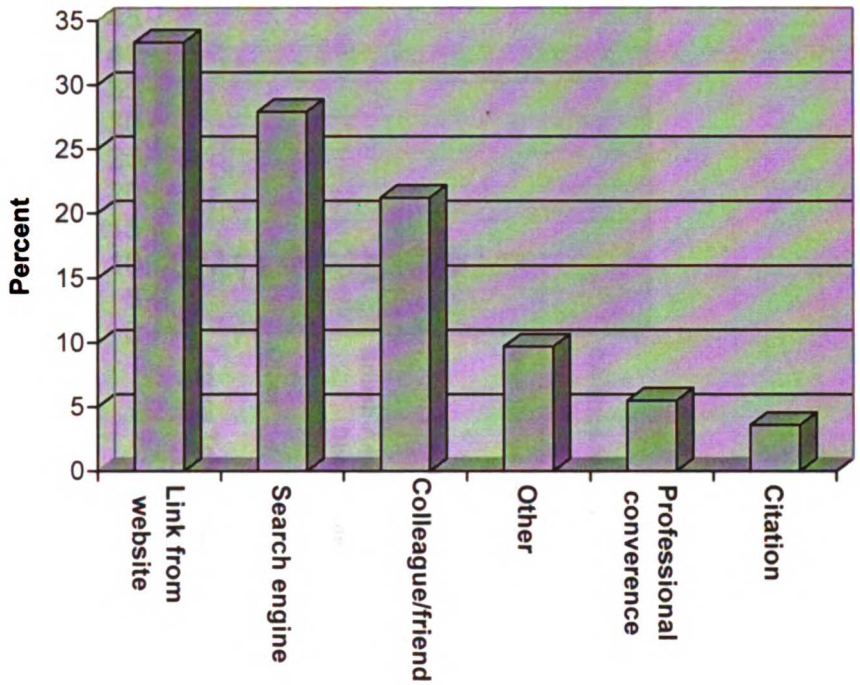


Figure 6: LTDL Survey - Why do you use the Legacy Tobacco Documents Library?

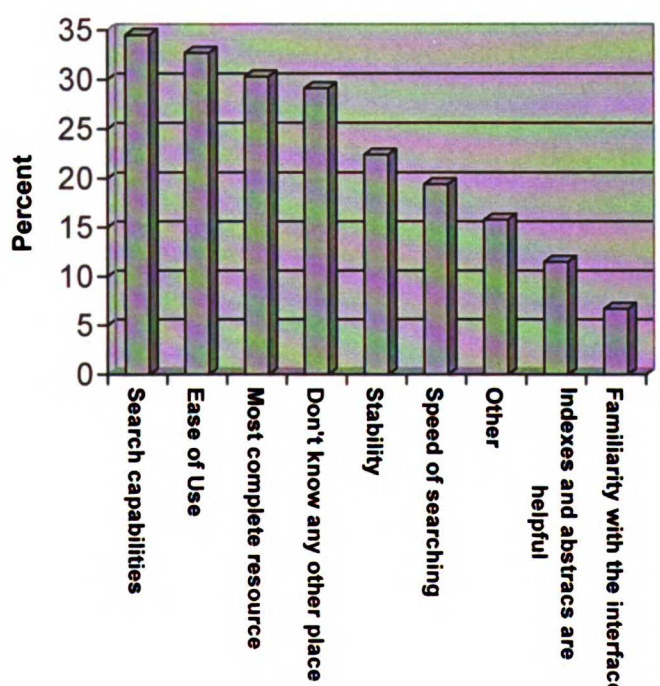


Figure 7: LTDL Survey - Which of the UCSF Tobacco Control Archives do you search most often?

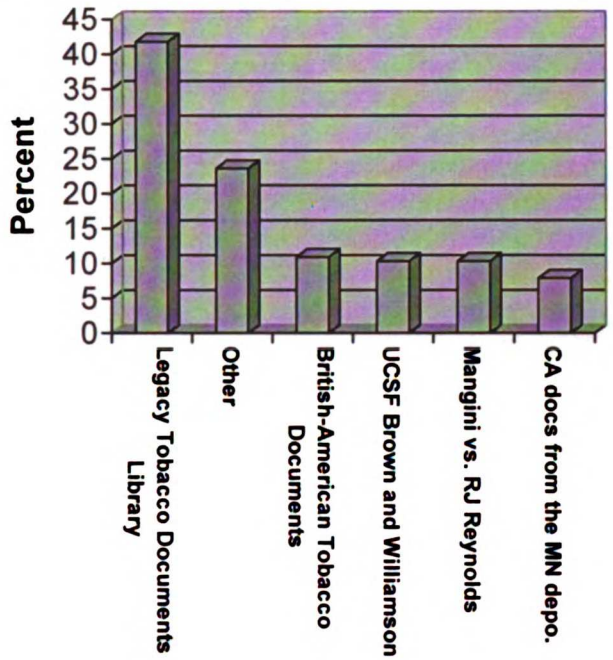


Figure 8: LTDL Survey - What are the most useful features of the Legacy Tobacco Documents Library?

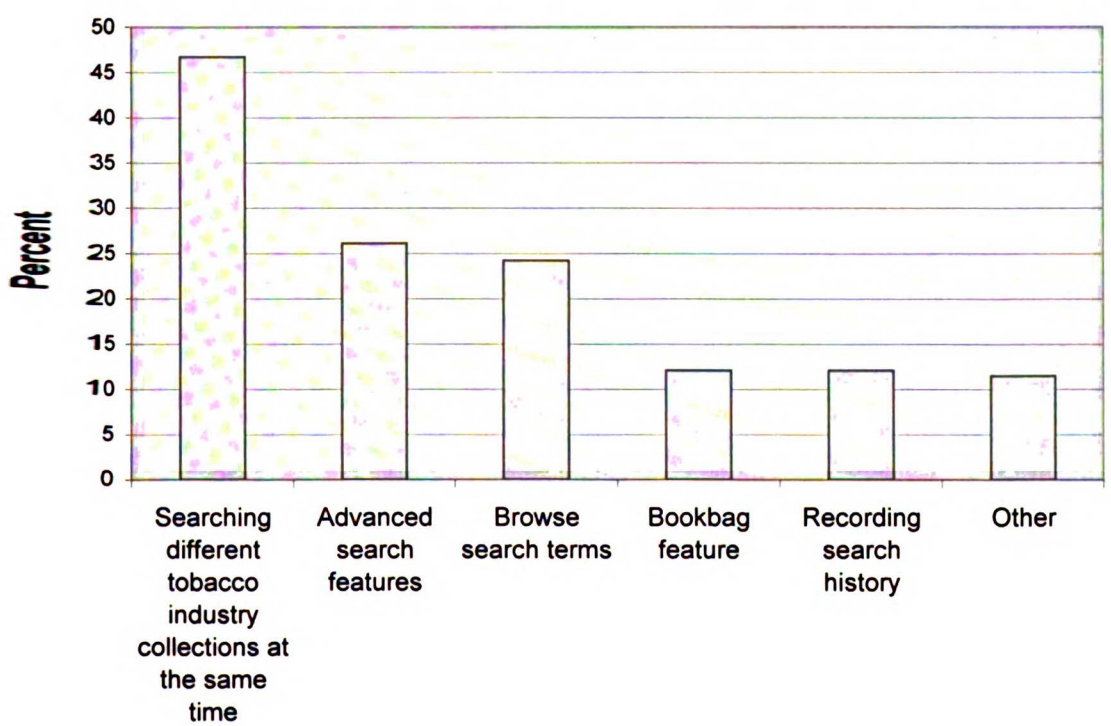


Figure 9: LTDL Survey - Have you had any problems using the website?

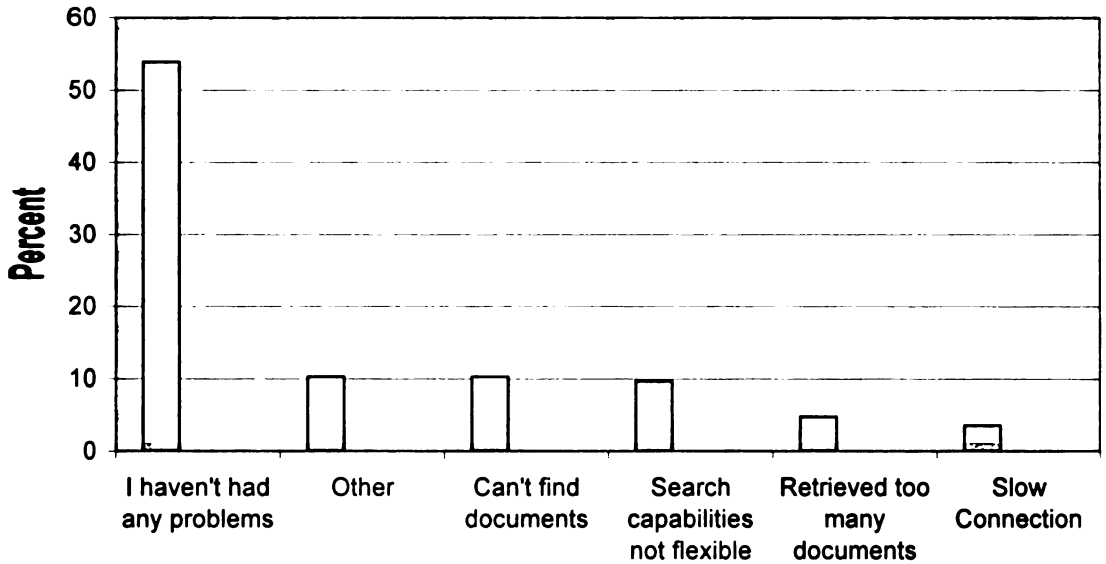


Figure 10: LTDL Survey - Please check other websites you use to search the tobacco industry documents

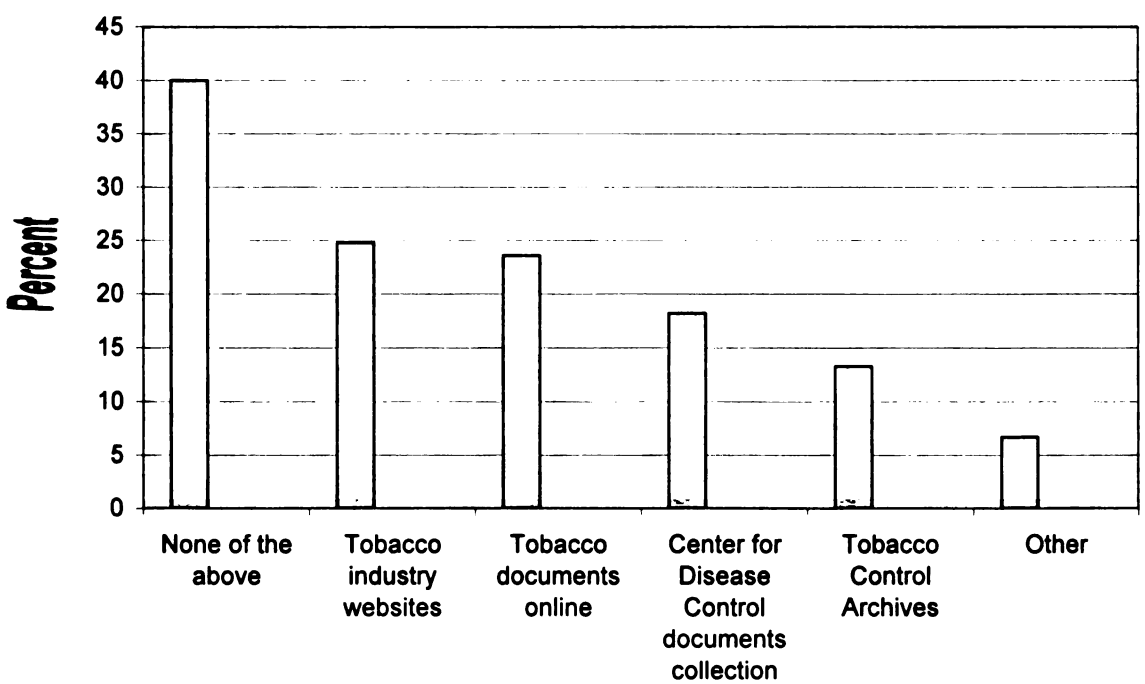


Figure 11: LTDL Survey - What features of the website do you find most difficult to use?

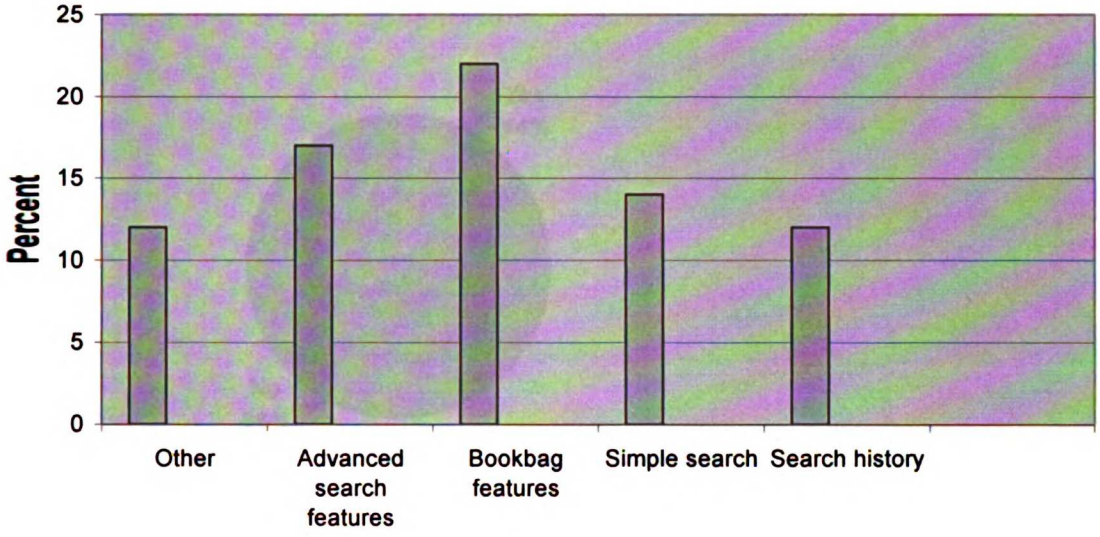


Figure 12: LTDL Survey - What attributes of a document are most important to you when searching online?

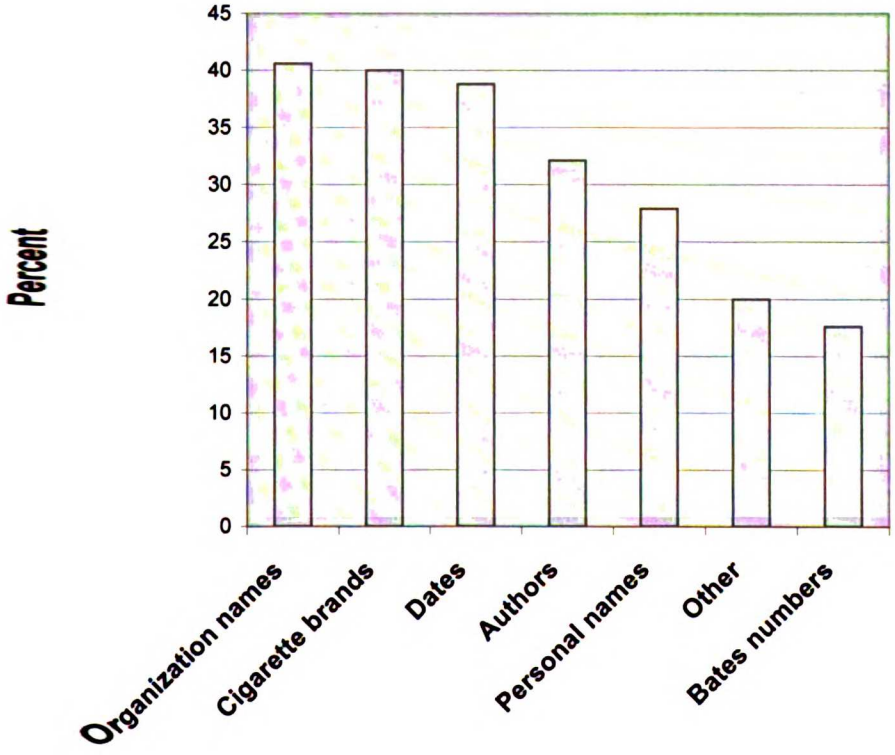


Figure 13: LTDL Survey - Have you had training in how to search the tobacco industry documents?

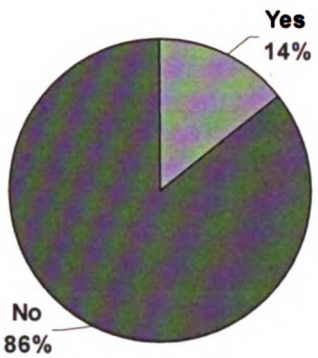


Figure 14: LTDL Survey - How do you organize the documents you search online?

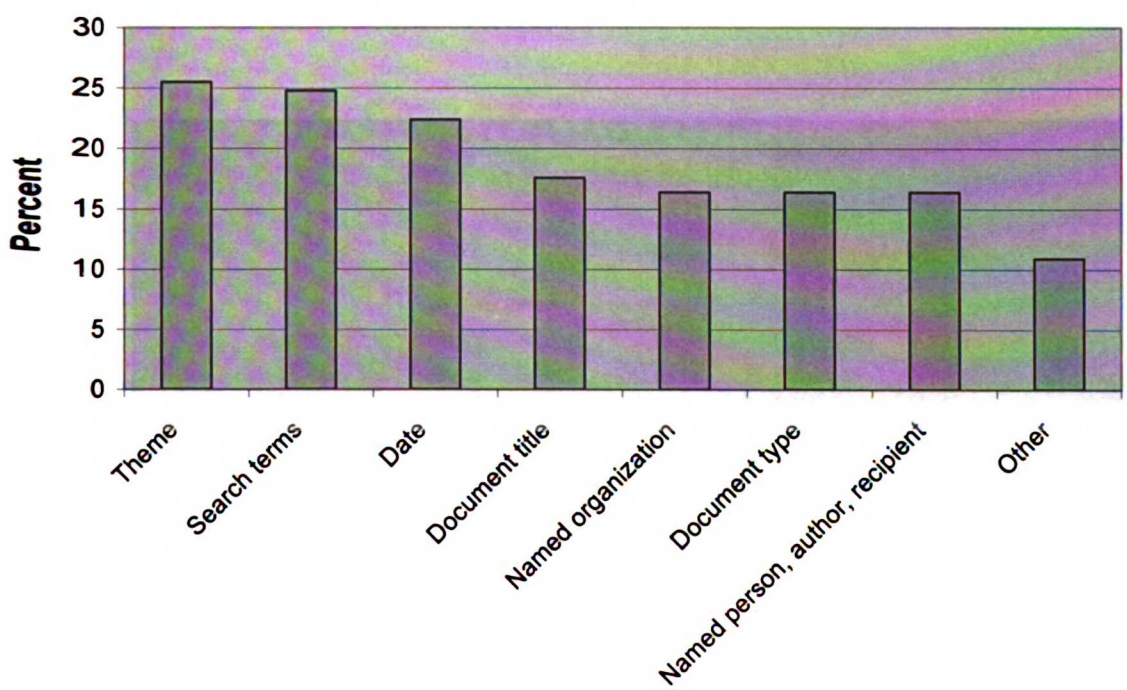


Figure 13: LTDL Survey - Did you find what you were looking for?

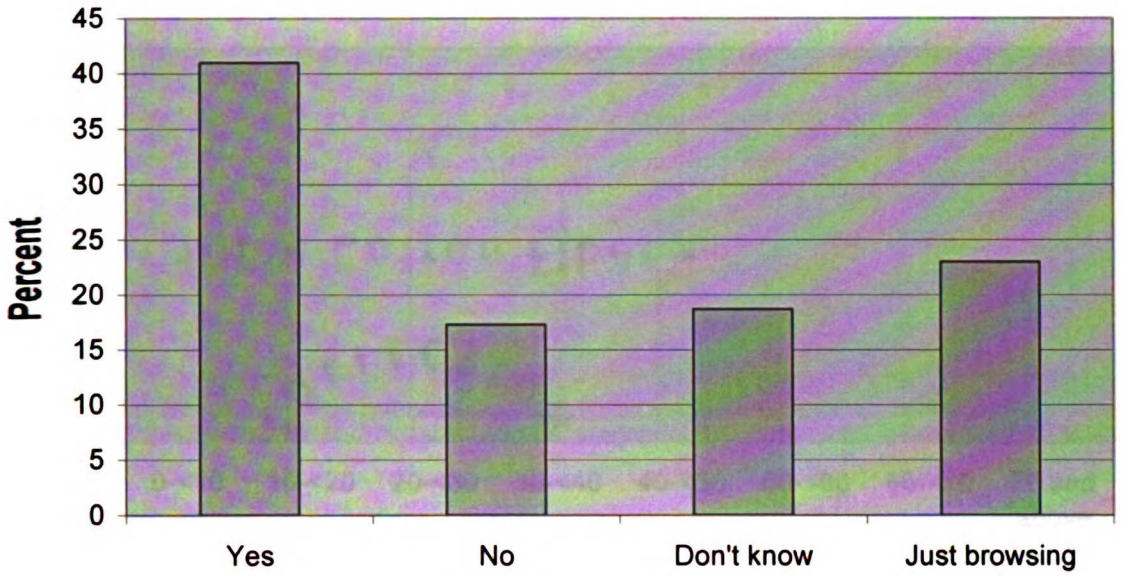


Figure 14: LTDL Survey - What is your gender?

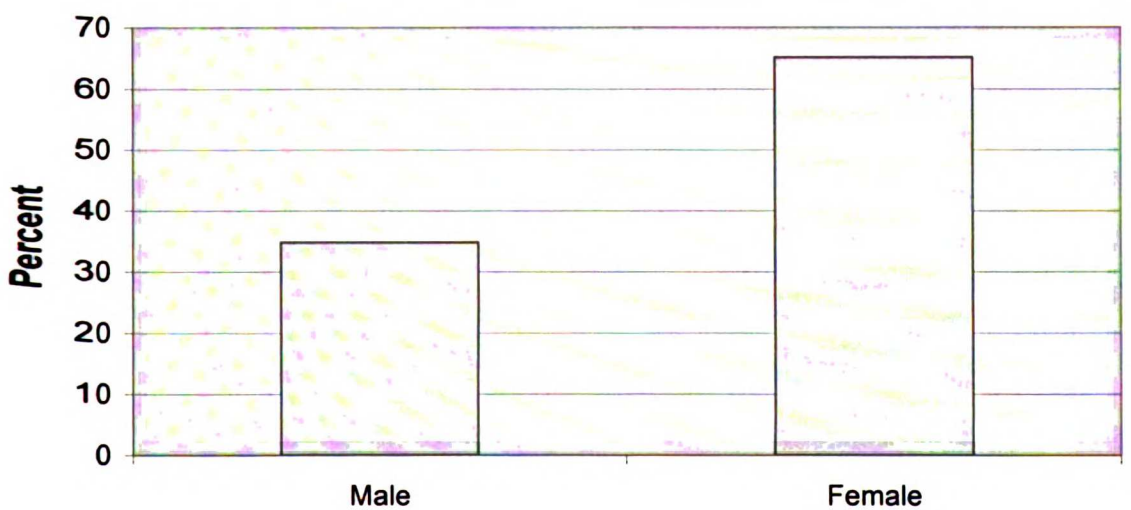


Figure 15: LTDL Survey - How old are you?

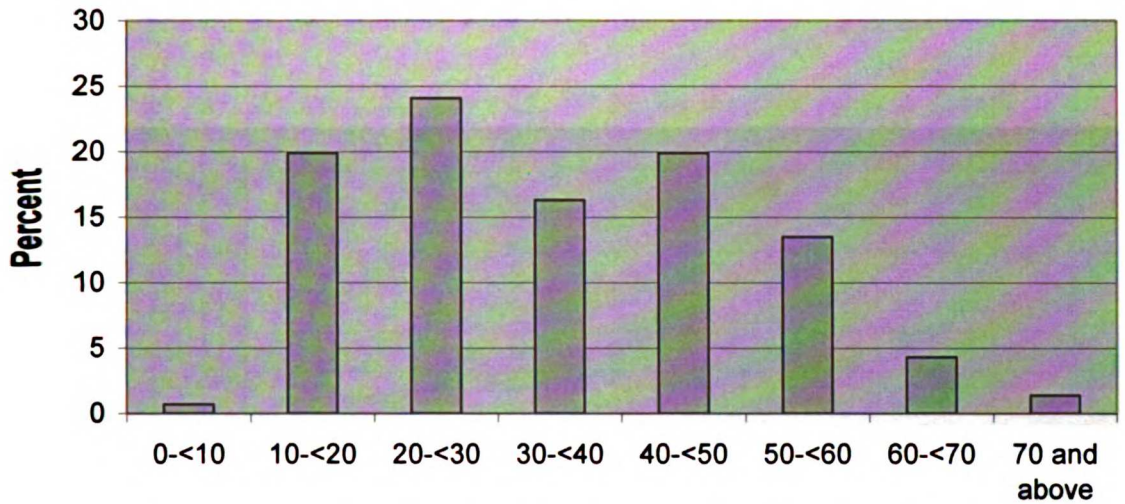


Figure 16: LTDL Survey - What is your ethnicity?

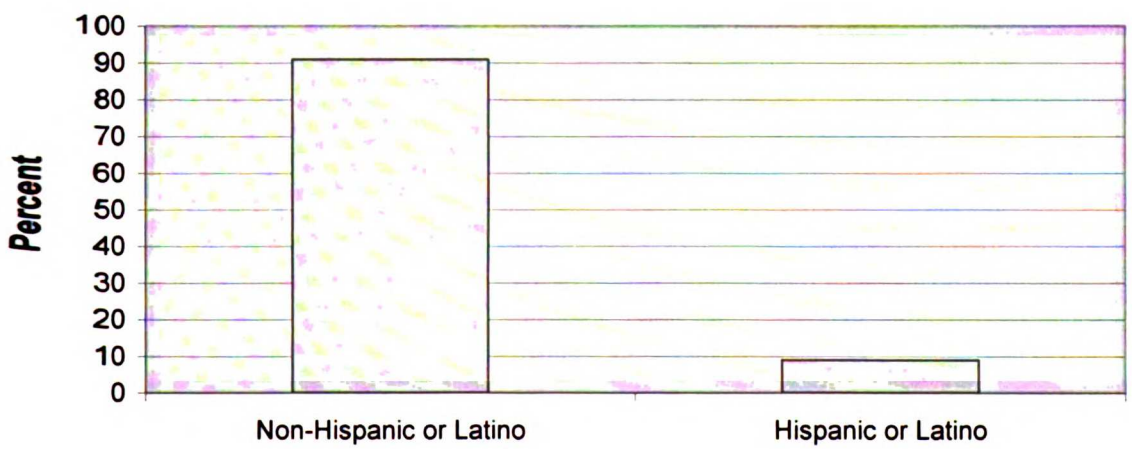
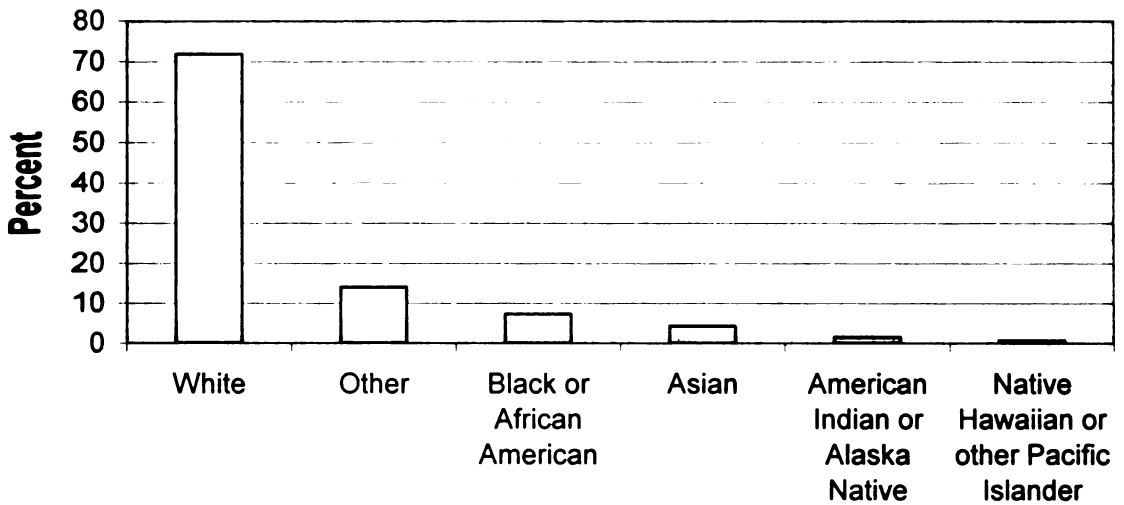


Figure 17: LTDL Survey - What is your race?



REFERENCES

1. Daniel HG, Johnston ME, Levy CJ. Young Smokers Prevalence, Trends, Implications, and Related Demographics. 31 Mar 1981. Bates No. 1000390803/0855. http://legacy.library.ucsf.edu/tid/ftu74e00.

2. Bero L. Implications of the tobacco industry documents for public health and policy. Annual Rev Public Health 2003;24:267-88.

3. The State of Minnesota and Blue Cross/Blue Shield of Minnesota v. Philip Morris Inc., *et al.* 1998 January 13,2004; Available from: http://www.naag.org/issues/issue-tobacco.php

4. Chapman S, Cummings K. Impact of new technologies in tobacco control: call for papers. Tobacco Control 1998;7(3):222.

5. Malone RE, Balbach ED. Tobacco industry documents: treasure trove or quagmire? Tobacco Control 2000;9(3):334-8.

6. University of California San Francisco. The British-American Tobacco Document Collection. 2002 January 28, 2002 [cited 2002 January 13, 2002]; Available from: http://www.library.ucsf.edu/tobacco/batco/

7. Flamenco search of the British-American Tobacco Documents. 2002 [cited 2004 July 9, 2004]; Available from: http://bailando.sims.berkeley.edu/tobacco-interface.html

8. Center for Disease Control. Tobacco Industry Documents. 2002 [cited 2002 March 05, 2002]; Available from: http://www.cdc.gov/tobacco/industrydocs/index.htm

9. SAS Institute Inc. SAS Language Version 8.2. Cary, NC; 2002.

10. Sly DF, Heald GR, Ray S. The Florida "truth" anti-tobacco media evaluation: design, first year results, and implications for planning future state media evaluations. Tob Control 2001;10(1):9-15.

11. Hearst M. Untangling Text Data Mining. In: Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics; 1999 June 20-26; University of Maryland; 1999.

12. Freeman K. Effective recruitment and retention for online surveys. In: IPCC 2002 Reflection on Communication. Proceedings IEEE International Professional Communication Conference; 2002 pp.509-12; Piscataway, NJ, USA: IEEE; 2002.

Chapter 4.
Assessing Users' Needs: Interviews of Tobacco Document Researchers

A. INTRODUCTION

The internal tobacco industry documents represent a tremendous opportunity for public health. The documents have been arguably one of the most useful results of the Master Settlement Agreement in 1998 when the Attorneys' General of 48 states sued the tobacco industry (1). The documents have already proven valuable for advancing public health objectives of reducing tobacco use and exposure(2, 3). However, tobacco control researchers have done most document analyses. Developing new ways of searching could open up analysis to other research disciplines, teachers, and advocates.

The objective of completing the interviews was to discover how novice and expert tobacco document researchers search the documents and to develop new methods of searching the tobacco industry documents by using a combination of quantitative and qualitative studies. From the interviews, we developed a new interface that users could use to search the tobacco industry documents.

There has been little research on text data mining user interface design itself and no research on text data mining using a large public health corpus of documents such as internal tobacco industry documents. The literature on various methodologies of data mining is well developed; however, text data mining is a nascent field (4). There are a number of advantages to develop text data mining for the tobacco industry documents.

While the existing websites are useful search engines for the tobacco industry documents, they are limited in providing any analysis assistance for the user. Using the existing search engine, the user often gets too many search engine results or too few. In

addition, the documents found in the traditional manner have no context, partly because one does not know the history of the tobacco industry. This makes it difficult to assess importance. The objective of the present research is to develop a user-derived interface for a text data-mining engine and test its usefulness for discovering new relations within the tobacco industry documents. Interviews offer the opportunity to examine how people approach searching the documents.

The purpose of this chapter is to gather detailed information about a subgroup of searchers of the internal tobacco industry documents, the researchers. As described in Chapter 3, the researchers are the people who are most familiar with the documents. They spend hours each day building stories out of millions of documents. My major interest was to examine how they conduct their searches and what type of informatics tools would be useful for them. By triangulating the answers from the interviews and the surveys, I was able to develop an interface that could be more useful for the researchers' daily activities.

This chapter includes a publication for IEEE EMB conference (5). It also includes additional analysis of the interviews. This chapter utilizes the interviews to develop a useful interface for researchers; however, in the process of the interviews, rich information surfaced about how internal tobacco industry document research is done.

B. METHODS

We conducted nine in-depth interviews of tobacco documents researchers. The interviews consisted of three components: 1) observations of a 20 minute unstructured

search of the Legacy website, 2) four standardized tasks, and 3) nine open-ended descriptive questions.

The four-structured questions for this part of the evaluation are as follows:

1. Find a document that contains "Ayres", who worked at British American Tobacco and see if you can find out what his job title was.
2. Find a document that was published during 1985-1990 that discusses marketing to young adults.
3. Find a document that says the following "Nicotine is not addictive".
4 Find the document that says that Sylvester Stallone will accept $500,000 from Brown and Williamson to use their products in the movies.

The above questions were standardized tasks to compare how users with different experiences approach the tasks. The first question was devised to test whether the users knew about the Philip Morris Glossary of Names http://www.pmdocs.com/PRIVLOGS/Clog/Clog/Glossary%20Pages%20Index%20rev%201.htm. The second task was designed to test whether the user knew how to do advanced searching. The third task was designed to test whether the user utilized alternative sources for searching such as Tobacco Documents Online. The final task again tested the users' familiarity with advanced searching.

We asked nine open-ended questions about tobacco documents searching:

1. What do you think are the strengths and weaknesses of your own search technique?

2. What is the most important feature of a search engine? (If necessary probe- usability, accuracy, flexibility, sorting capacity, bookmarks, consistency, reversal of actions, shortcuts...)
3. What manner do you use to organize documents that you retrieve: by subject area, date, organization name, theme, etc.?
4. How does the UCSF Legacy collection meet your searching needs?
5. What would you modify to make the searching process more applicable to your specific needs?
6. How long have you been using the UCSF Legacy online collection to perform searches?
7. What additional resources do you use for searching and why?
8. When you do not know the focus of your research, how do you begin to drill down and target specific information?
9. What difficulties have you had using the UCSF Legacy collection?

The interviews were conducted by myself and a student, Tiffany Bright, during the summer of 2003. They were tape-recorded and transcribed by a professional transcriber, Steven Zeluck. Participants were also videotaped during the 20 minute unstructured search and tasks. Only the screen of the computer was videotaped, not the person conducting the searches. The interviews were approximately one hour in length, and conducted face to face. An observer recorded additional observations.

The users were divided into experienced searchers (searched Legacy for six months or greater) and inexperienced searchers (searched for less than six months).

There were four experienced searchers and five inexperienced searchers. They were selected from a convenience sample from the Center of Tobacco Research and Education. From the interviews, we abstracted themes, which were sorted into suggestions that could be addressed by informatics tools or other suggestions about the interface

C. RESULTS

1. Research Themes

To analyze the interviews, we identified recurrent themes within them. Part of the strength of this dissertation is the combination of many methods, including qualitative methods. To develop themes, I read the transcribed interviews thoroughly, organized the responses of individuals and categorized them by themes. The idea behind this framework of working is to focus on a full understanding of the individual case before combining cases and arranging them thematically (6).

The basic themes obtained from the interviews are shown in Table 1. The themes are too many documents, too many duplicate documents, cannot find documents/too few documents on topic, and questions about how to best organize the documents.

The following statement from a new researcher, defined as someone who has searched the documents for less than six months, exemplifies problems when identifying useful documents. "Uh now since I'm inexperienced sometimes I've been checking everything. But I always get back a lot of documents."

Table 1: List of Major Themes Derived from Interviews

Theme
Too many documents/ different people have different limits
Questions about how to best organize the documents
Cannot find documents/too few documents on topic
Hard to explain research methods
Hard to eliminate duplicate documents
Many researchers augment their searches by other websites.

This researcher entered the world of the documents by a simple keyword search for a person's name, Dr. David Kessler, former head of the Food and Drug Administration, and his name retrieved 25,000 documents. The search of Kessler in the documents resulted according to the information provided on the Legacy website, in 32,213 matches over 29,100 records.

This leads to some internal debate from the researcher about where to go next:

"...I will just go ahead and start looking to see what's on top. I could think about narrowing down the dates. Or maybe I would like to sort by day. Isn't there a way to do that? Or haven't I, I did this earlier today. Is that only an advanced search? Oh [I] see now. ..."

The intuition of the researcher is to narrow down the search by date. However, initially the person did not know how to do that. In addition, there is a question about

when the event in which the research is interested occurred. Again it becomes a guessing game about how to obtain documents from a specific year, 1982. The researcher was uncertain about how to represent all documents occurring in 1982 based on the interface.

Another researcher also encountered the problem of her initial search retrieving too many documents. This researcher had already had one year of experience searching and knew how to do an advanced search.

> "...1,767 documents. Now that is a lot of documents. I'm not going to be able to go through them. That's just not efficient for me. So I'm going to subcategorize it (so) another search term, that's 42, help me narrow my search. So – I mean I want to do an advanced search."

The researcher with more experience knows to go to an advanced search right away. She looks for memos with the name of the person she is searching because

> "...We want memos mentioned and having (inaudible) the author, because memos are very informative, they're informal, they're casual, or they're formal, but in any case memos disclose a lot of correspondence between people, you know, whether industry executives or people that they form alliances with, and so memos of the document type, you know, subcategories under a document type is really nice, because the actual correspondence between people is a large part of the story making, okay?"

Researchers become very familiar with the names and players within their particular topic. Often the familiarity is obtained by reading documents over time; however, there is a privileged log list that can help. However some of the quotes below exemplify how a researcher begins to know the people in the documents through their writing.

"...Well, there's a whole list of who people are, who's who in the tobacco industry. And ( ) it was a privileged law file, so if you're wondering who these people are, like R. Carchman, C. Ellis, I know that. And if you do enough documents searching it's the same players over and over again. ... You just become very familiar with who these people are."

The researcher continues to look for documents until a saturation point is reached.

"...And that's – when you keep seeing the same documents come up under certain categories, like Enstrom, Phillip Morris, or Enstrom and memo, Enstrom and confidential and if these documents keep coming up that's when I know I've hit some kind of a saturation point with the whole topic, the whole theme. I'm still getting the same documents whether I hit these, use this search or that search term. I'll know that I'm kind of getting as much mileage out of this story as possible."

In many ways the objective of researchers is to get "as much mileage out of this story as possible". The problems that the researcher runs into while doing this include too many documents to search through and that important details may remain hidden in the documents. Also, in this interview, it becomes evident how important the handwritten notes are in documents.

2. Connecting the Missing Dots

Document researchers are like archeologists, digging for buried treasures and gluing together shards.

"...they're, they're kind of, they're kind of unconditional making these conditions, you know, they really want to cover themselves, right? They want to make sure that it's within certain boundaries, like monetary boundaries and legal boundaries. So that just kind of tells you where they're going with these projects. And (inaudible) can kind of infer a lot. So that's why memos are great, because they're not like reports, and they're not like publications. They really give a people feel. To see -- a feeling of these people. And you know, you want their feelings, because – and personality, because they spell out really the actions, the motivations and actions. And okay. So here's Shook Hardy and Bacon. So this is the memo from Shook Hardy and Bacon, the industry law firm. So this should be a really exciting factor. And okay. "The attached is from Shook Hardy and Bacon" to ...I think they are internal industry lawyers. "The attached is JAMA's instructions to authors forwarded to you regarding the Enstrom discussion. We are trying to obtain a copy of Enstrom's author's statement via RJR." That's Reynolds. "Also for (inaudible) inspiration we are forwarding a paper that may (inaudible)." So you have to figure out that. But the point is this is an industry – these are industry lawyers talking about a publication or possible publication submitted to the Journal of the American Medical Association. So they're kind of brokering something, and that's very not – that's not cool. That's

not really what you know, if you're a scientist and you're submitting an abstract or submitting a possible – a draft or publication to JAMA. That's really between, should be something between you and JAMA. What we see here is kind of a inter – we see these lawyers talking about authorship criteria at JAMA."

3. Strengths and Weaknesses of the Researchers

Within research, everyone has his or her strengths and weaknesses. Some are better at qualitative research, in eliciting responses from their subjects, in determining what to ask next in a semi-structured interview. Likewise, in quantitative analysis, formulating a targeted research question and developing methodology to address that question is crucial. In internal tobacco industry documents research, different researchers approach questions in different manners, depending on their background and situated location.

"...Well I think the strength is that I have a real definitive style like in terms of how to search, or I mean I have a certain style and it's really like off the cuff. There's no real methodology. Like I said I'll just, you know, it's very intuitive, but that's also its flaw. That's also a weakness, because I can't relay to other people specific directions. The history of it is very hard for me to relate to other people if they want it. And that's not a problem because in these papers you write you have to give a message. "

4. Suggestions for searching from the researchers

Some of the following suggestions were obtained from the interviewees about searching using LTDL (Table 2). The suggestions have been utilized in a training developed for advocates and researchers.

Table 2: Suggestions from Interviewers on how to better search the documents

- Bookmark page advanced page in Netscape or IE
- Using date limits
- Most people work with the PDFs
- Expand PDFs to ~120%
- Most people print out documents
- Start with a general search (simple search and get more specific)

D. CONCLUSION

Rather than creating a product and presenting it to the users, we asked the users what they currently do and adapted the technology to their current workflow.

We found that text mining could be useful for researchers of the tobacco documents in a number of areas. First researchers could use classification of the documents to gather data into one of several discrete concepts. For example, the term "young adults" could be sub-classified into documents pertaining to marketing, advertising or health effects. We also found that the users would benefit from the suggestions of search terms; for example: young adults AND marketing: –Y.A. AND marketing - Young Adult* AND marketing

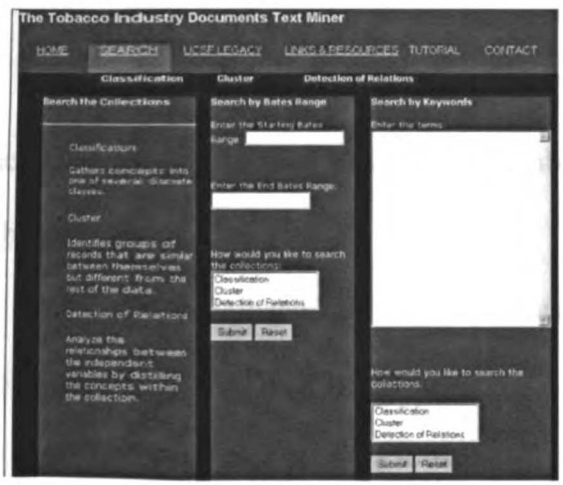
An additional benefit would be detection of relations, which would enable the user to analyze the relationships between independent variables by distilling the concepts within the collection.

Other benefits include factors that are unique to documents research. Often, researchers will find a particular document of interest, but are unable to recall the search process of how the document was found.

Another need identified from interviewing the researchers was the ability to track the stages of revisions in an internal tobacco document and identify the changes made to the document over time. Following the path of a document from conception to dissemination would be useful for a researcher. The tobacco industry lawyers often edited scientific documents for content(7, 8). The researchers would like to know who edited the document and when.

When applied research groups ask their users what products they want, they are more likely to find that the needs can be met with less resources than originally thought. Creative solutions are necessary in public health during these times of budgetary crisis. Text mining has the potential to be useful in many different public health settings from outbreak detection to breast cancer prevention. However, we need to go to the users in the state and local health departments to find out how they use data and what text-mining algorithms suit their needs.

Figure 1- A screenshot from text data miner interface



Finally, the clustering of documents based on automatic categorization assignments would be useful for both experienced and novice users. The novice users could use clustering to get an overview of a particular area of interest. Experienced users could use clustering to help them draw connections that they previously did not think of or use it to narrow down a search with many hits.

ACKNOWLEDGMENTS

We thank the researchers in the Center for Tobacco Control Research and Education who participated in the interviews. We also thank Kirsten Neilsen, the tobacco

documents librarian at UCSF and Annamaria Baba for their helpful comments. This work was supported by Tobacco Related Diseases Research # 12DT-0186.

REFERENCES

1. The State of Minnesota and Blue Cross/Blue Shield of Minnesota v. Philip Morris Inc., *et al.* 1998 January 13,2004]; Available from: http://www.naag.org/issues/issue-tobacco.php

2. Malone RE, Balbach ED. Tobacco industry documents: treasure trove or quagmire? Tobacco Control 2000;9(3):334-8.

3. Bero L. Implications of the tobacco industry documents for public health and policy. Annual Rev Public Health 2003;24:267-88.

4. Hearst M. Untangling Text Data Mining. In: Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics; 1999 June 20-26; University of Maryland; 1999.

5. Michel M, Bright T, Bero L. Creating a text data-mining application for use in public health informatics. In: IEEE-EMBS Annual International Conference. San Francisco, CA: IEEE EMB; 2004.

6. Patton M. Qualitative Research and Evaluation Methods. 3rd ed. Thousand Oaks, CA: Sage Publications; 2002.

7. Bero L, Barnes DE, Hanauer P, Slade J, Glantz SA. Lawyer Control of the Tobacco Industry's External Research-Program - the Brown-and-Williamson Documents. Jama-Journal of the American Medical Association 1995;274(3):241-247.

8. Hanauer P, Slade J, Barnes DE, Bero L, Glantz SA. Lawyer control of internal scientific research to protect against products liability lawsuits. The Brown and Williamson documents. Jama 1995;274(3):234-40.

CHAPTER V.

EXPLORATIONS IN TEXT MINING: FOCUSING ON CLUSTERING AND BAYESIAN LEARNING

A.) PURPOSE

The purpose of this chapter is to describe the methods and algorithms used to mine text, i.e., using statistical algorithms to make novel connections in text. The chapter examines existing programs that could be used for text mining, with a focus on exploring how clustering and Bayesian learning, specific types of text mining, could be of use to internal tobacco document researchers.

This chapter is not a comprehensive look at all software for text mining; however, it does have a practical approach towards the evaluation.

B.) INTRODUCTION

The definition of text data mining is widely contested in the literature. Some define text mining as abstracting meaningful categories automatically from text or text summarization. Others define text mining as finding patterns or doing exploratory analysis in text (1). It differs from information retrieval, which involves identifying and ranking documents that match the information needs of the users. With information retrieval, the user has already perceived an information need and issues a query in order to obtain information. In text data mining, the user has the potential to obtain information they had not anticipated.

Artificial intelligence has two major categories of machine learning: unsupervised learning and supervised learning. One of the main distinctions between text mining and

statistical language processing is the concept of unsupervised learning. Statistical language processing uses frequencies of words to examine a corpus of documents, while text mining is a tool for the user to discover information not previously known about the corpus. In supervised learning the answer is predetermined. Unsupervised learning is machine learning, or learning via a computer, that has not been predetermined. In other words, the answer is not already known.

For example, imagine one had 1000 documents and already knew that 200 were related to animal testing and second hand smoke, 200 were on testing benzopyrenes in rats, and the remaining 600 were on nicotine metabolism in mice. One could use various algorithms on a computer to put the documents into vectors and run machine learning algorithms on a training set, then on a test set, however the classification is already known. The test set is used to determine the accuracy of the classification (2).

Unsupervised learning in text mining is more difficult since there is no gold standard, or correct answer to which to compare the results, since the results are not already categorized. In the prior example if one obtains the same classification or cluster of results, one would have no way to know if the results were accurate except by reading through each document and categorizing it. While it is possible, although not necessarily pleasant, to read and manually classify 1000 documents, it becomes much less so when one is dealing with a dataset of millions of documents.

Although text data mining technology is not completely developed, it holds great promise for business, industry and the government in dealing with information overload. Software companies such as SAS, IBM, and SPSS, have embraced text mining and recognized it to be a major priority for the 21st century (3). Business could use text

mining for defining marketing segments, determining purchasing patterns, or other things. Some fields in which data mining has been applied include retail and marketing, banking, health insurance, transportation, product manufacturing, healthcare and pharmaceutical industries, and governmental agencies such as the Federal Bureau of Investigation (4). However the technology requires further development and appropriate application, and the available products have yet to be stringently evaluated. The following table presents the current state of affairs.

Table 1: Pros and Cons of the Current State of Text Mining

Pros	**Cons**
Potentially make unknown connections and lead down a useful train of thought	Pursue lots of connections that do not lead anywhere
Determine patterns in data that were inaccessible without text mining	May end up with as many connections as in information retrieval leading to more information overload
Define new categories and categorize millions of documents quickly for more metadata, or data about data.	Cost can be immense
	Not user accessible
	Amount of knowledge to effectively use these systems is inaccessible to most people.

In summary, however, text mining appears to offer the best strategy for dealing with information overload, and with regard to the present study the possibilities of rendering the internal tobacco industry documents more usable and accessible.

C. NEED FOR TEXT MINING TO ADDRESS INFORMATION OVERLOAD

Information overload is common vernacular which refers to the overwhelming amount of information a person in society must process. Since the dawn of the digital

age, we have had more sources of information available to us with less filtering. While the information age has improved communication and access has increased, questions remain about the impact of information anxiety on people and on society.

Many researchers have attempted to address information overload in a variety of ways: by describing the problem, providing qualitative and quantitative solutions for the problem, and developing tools to solve the problem. Text mining is one of the possible solutions.

1.) Description of the information overload problem

One of the major descriptive studies on the quantity of information was done at the School of Information Management Sciences at UC Berkeley by Professor Lyman(5). The report documented the information flow through 4 different media: print, film, magnetic, and optical. It was found that in 2002 these four media accounted for 5 exabytes of new information and that 92% of that information was stored on magnetic media such as hard disks (6). Most people do not have a concept of how big five exabytes is. It is equivalent to 1,000 terabytes, where a terabyte is 1,000 gigabytes.

Gigabyte (GB) = 1,000 million bytes = 10^9 bytes

Terabyte (TB) = 1,000 gigabytes = 10^{12} bytes

Petabyte (PB) = 1,000 terabytes = 10^{15} bytes

Exabyte (EB) = 1,000 petabytes = 10^{18} bytes

The report attempted to help readers conceptualize this huge amount of data by comparing it to analogies readers are familiar with,

"...If digitized, the nineteen million books and other print collections in the Library of Congress would contain about ten terabytes of information; five exabytes of

information is equivalent in size to the information contained in half a million new libraries the size of the Library of congress print collections."(6)

Other findings of the report are that we are in the midst of an information explosion. For example, the report estimated that new stored information grew about 30% a year between 1999 and 2002. Americans spend significant time gathering information. Studies on media usage document that adults use the telephone 16.17 hours a month, listen to radio 90 hours a month, and watch TV 131 hours a month (5). In addition, average internet usage is 25 hours and 25 minutes a month at home and 74 hours and 26 minutes at work (5). The explosion of information in general life parallels the experience of searching the internal tobacco industry document. The number of documents continues to increase, which makes finding particular documents increasingly difficult.

2. General Societal and Technological Implications of Information Overload

Information overload affects people across disciplines, such as law, medicine, and social science in similar ways, by causing information anxiety. Information anxiety is the response to information overload of "I can never know enough" or "This is an interesting tangent". Everything becomes interrelated and boundaries between topics are fuzzy.

This feeling can lead to cognitive overload. Kirsh acknowledges that the workplace is a complicated knowledge driven environment, filled with multitasking, and shifting teams of people. In fact, one documented effect of cognitive overload is tension with colleagues, loss of job satisfaction and strained personal relationships (7). Information anxiety, which was defined by Richard Wurman, author of Information Anxiety, as being produced by the gap between what ..."we understand and what we

think we understand", is pervasive in the workplace and at home. People increasingly have a hard time not checking their email and there are more distractions than ever. Attention has been turning to management strategies from the software perspective.

Kirsh lists four systems that contribute to information overload: too much information supplied, too much information demanded, multitasking, and inadequate infrastructure to help reduce metacognition(7). These causes of information overload are present in many fields, including tobacco documents research. He refers to both pushed information, information that enters our space of being and requires action, and pulled information, that is information we want. There is an oversupply of both; however, they differ for different knowledge workers. For example, one knowledge worker, that is people who work with information and knowledge, in business management might receive 70% pushed information and 30% pulled information, whereas a tobacco documents researcher might have 25% pushed information and 75% pulled information(7). Since the 75% represents time spent in searching, clearly tobacco knowledge workers would benefit from tools that modify this ratio.

Another important problem in searching is a consequence of information supply; the belief that there are always higher quality facts out in the ether that we do not know about (7). This is partially justified because in the past few years the amount of information has increased exponentially, while quality has only risen linearly (7).

In order to deal with information overload, people adapt different strategies such as: accumulating information blindly, performing just-in case learning, or massive surface clutter(7). The blind accumulation strategy is if any information might right now or at some point in the future become relevant, it is to the benefit of the knowledge

worker to keep it. This strategy results in overstocking and potentially not being able to find the necessary piece of information in the future, or spending too much time filing. The just-in case learning strategy knows enough to be prepared for anything. It is rewarded in the school system and valued in our society. An example of the surface clutter strategy is a knowledge worker that tries to keep all information accessible at all times.

These strategies apply to different tobacco knowledge workers and could be utilized in interface design. For example, a person who likes to keep all information near by at all times may prefer a history bar, which is a listing of their recent or past searches open on the screen. Workflow analysis can also help in a situation like this. Finding out what people need when and how they complete a task is vital to interface design. For example, while writing in a word processor we are constantly interrupted by other tasks to complete, such as references to lookup online, or in the filing cabinet. We are faced with choices about what to do when and when it is important to complete the task at hand. By better understanding workflow in a particular environment, programs can be designed to better fit the user, rather than asking the user to fit the program. It is predicted that in the future, word processors will contain additional tools to help users conduct more of their work within one program instead of many (7). Rather than having to switch to an Internet browser to look up an article, you could look up the article from within the word processor. Especially in utilizing the immense resource of the internal tobacco industry documents, text mining offers invaluable strategies for researchers and laypeople.

D. TEXT MINING FOR THE TOBACCO INDUSTRY DOCUMENT RESEARCHERS

Many academic communities, outside of statistics, computer science, and biomedical informatics, have not adopted text data mining as a method of research. This fact may be due to the lack of applicable software, the startup price of commercial products, or the lack of a perceived need. The usefulness of text data mining in practice is also a point of contention (see Table 1). Although literature on various methodologies of data mining is well developed; text data mining in a research context is a nascent field (1). In particular, there has been no research on text data mining using the tobacco industry documents.

The present study undertakes to solicit suggestions from users of the internal tobacco industry documents for product features that they thought would be of benefit to them. Based on interviews conducted in the summer of 2003, (Chapter 4) it was possible to determine key themes that users were interested in for future document interfaces (Table 2). An important note is that these items would most likely be different from other groups of document searchers such as teachers, advocates, or lawyers. Most groups would contain specific suggestions useful to their field.

Table 2: Suggestions of Researchers to Improve Searching the Tobacco Industry Documents

Suggestion	Comments
Include Optical Character Recognition (OCR)	OCR is included in BATda but not in LTDL.
Include other collections, such as a full OCR version of LexusNexus, Bilely, and the BATda within Legacy	BATda and LTDL will eventually be merged. However, Lexus Nexus is a separate interface and will remain so.
Add the ability to perform complex Boolean searches	Completed for BATda and LTDL
Eliminate redundant documents	Complicated and useful to some researchers in that redundant documents occur in separate files and it can be useful to see which person has which document.
Find similar documents	See Appendix D

Based on the suggestion from users that they would like a feature that finds similar but not redundant documents, we undertook a systematic evaluation of open-source and commercial clustering products that could be used for clustering the internal tobacco industry documents.

There were other reasons to pursue clustering besides users demands. First, LTDL is based on retrieval of information that has been indexed on limited metadata. Although a clustering program would not have additional metadata to utilize unless automatic categorization and indexing were performed, clustering would enable the discovery of "new knowledge" and new connections that perhaps would not have been made based on the search and retrieval method of document searching.

Clustering could help by gathering similar project topics together. For example, Project Alpha and Project Premier were both projects having to do with environmental

E. USING TEXT MINING PROGRAMS AND ALGORITHMS

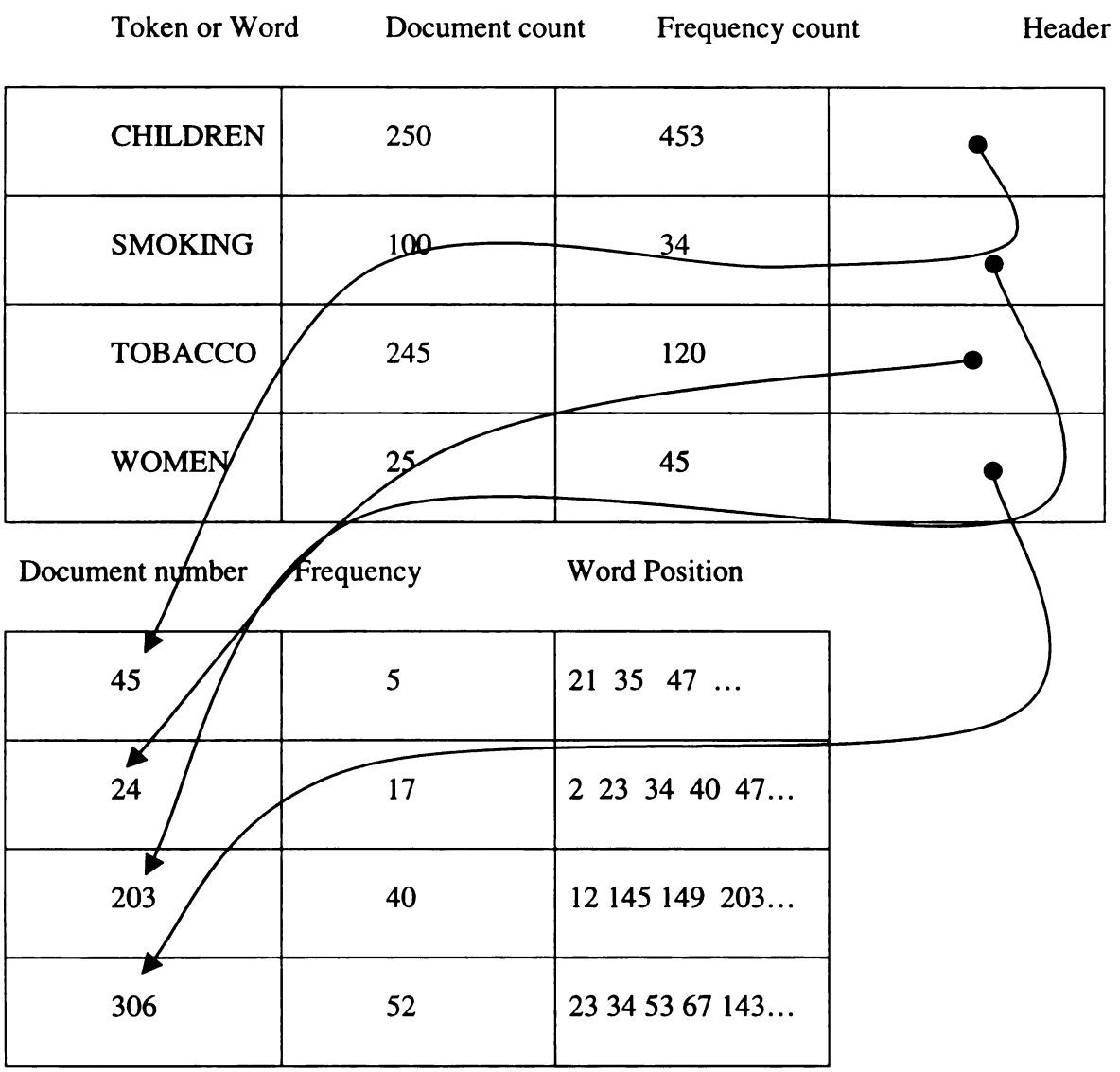
1. Preprocessing

In order to perform any sort of text mining algorithm, the corpus must be preprocessed. There are a number of steps in preprocessing. First, each word must be stemmed, or broken down into its common root (8). For example, the word 'child' might be used as a rood word for variants such as 'children', 'childish', and 'child's'. Second, an inverted file or dictionary must be created (8). The inverted file lists the all the words or tokens in a file. From each token two very useful statistics can be obtained:

1. Document count: The document count displays how many documents in which the word or token appears.

2. Total Frequency count: This is the number of times the word occurs across all the documents.

An example of an inverted dictionary is shown below:



The figure above is derived from Jackson and Moulinier (7).

In addition, to perform text mining algorithms of any sort documents must be added to vectors to perform numerical calculations. This vectorization is commonly referred to as term frequency vs. inverse document frequency or tf*idf. An example of

a simple vector model is shown below. Term frequency is a count of how many times a term t occurs in a document. Inverse document frequency is the relative presence of a term where idf = log (N/n_t). N is the total number of documents in the collection or corpus and n_t is the number of documents in which term t appears.

2. Definition of Clustering

Clustering, also known as segmentation analysis and taxonomy analysis, is defined as partitioning groups of similar items together(9).

There are 2 basic approaches to hierarchical partitioning, or nested clusters, where smaller clusters occur with in larger clusters, called agglomerative and divisive (9). Agglomerative approaches clustering from the bottom up, that is it starts at the bottom and at each level merges a selected pair of clusters into a single cluster(10). In comparison, top down clustering or divisive clustering starts at the top and at each level recursively splits one cluster into 2 new clusters depending on achieving the largest between-group dissimilarity(10).

3. General Difficulties with Clustering Algorithms

Clustering algorithms can be broken up into several categories. These algorithms are considered an unsupervised learning problem, which is much more difficult than supervised learning because of two major problems. First, we do not know when creating categories for clusters how many clusters there should be. Second, we do not know if the given classification we have for the data is "correct". Keep in mind that the data may

have more than one correct answer so evaluation of the results is very difficult (2). With supervised learning there is a clear measure of success (10).

Unsupervised learning is defined as learning without the help of a supervisor or teacher. There is no gold standard and the results cannot be verified. The question remains as to what use unsupervised learning would be if you cannot measure the success of the assignment.

As mentioned above, we do not know how many categories the documents could be divided into (2). We also do not know if the given classification we have for the data is correct. Keep in mind that the data may have more than one correct answer so evaluation of the results is very difficult.

4. Description of Algorithms

I will now describe two common text mining algorithms that are used in the products evaluated: k-means and naïve Bayesian learning.

The following symbols will be employed in the descriptions of algorithms.

Table 3- Symbolic Notation for Algorithms

Notation	Meaning
$X = \{x_1 ..., x_n\}$	is the set of n objects to cluster
$C = \{c_1, ...c_j, ...c_k\}$	is the set of clusters
P(X)	is a powerset (set of subsets) of X
$\mathrm{Sim}(\cdot,\cdot)$	is the similarity function
$S(\cdot)$	is the group average of the similarity function
m	is the dimensionality of vector space $\mathbb{R}^m$
M_j	is the number of points in cluster c_j
$\overrightarrow{s(c_j)}$	is the vector sum of vectors in cluster c_j
N	is the number of word tokens in training corpus
$w_{1,...,j}$	are the number of word tokens in the training corpus
$\pi(\cdot)$	function assigning words to clusters
$C(w^1 w^2)$	number of occurrences of string $w^1 w^2$
$C(c_1 c_2)$ s.t. $\pi(w^1) = c_1 \pi(w^2) = c_2$	number of occurrences of string $w^1 w^2$ $X = \{\overrightarrow{x_1}, ... \overrightarrow{x_n}\} \supseteq \mathbb{R}^m$
$\overrightarrow{\mu_j}$	is the centroid for cluster c_j
$\sum_j$	is the covariance matrix for cluster c_j

From Manning and Shutze(11)

a) K-means - K-means is a very common clustering algorithm. It can be thought of as an ANOVA procedure in reverse(12). With ANOVA you see how similar

the variance is within a group and between groups and do hypothesis testing that tells if one of the groups is significantly different from the others. In k-means you minimize within group variance and maximize between group variance.

It is an iterative algorithm where cluster centers, or centroids, are chosen at random and each object is assigned to the cluster to which it is closest (11). Then the centers are recomputed and the process is repeated. In Figure 1 the algorithm is described in detail.

As mentioned above, first you select K points as the centroid, then assign all points to their closest centroid, recompute the centroid, and continue the iteration until the centroid does not change (13).

Given: $X = \{\vec{x_1}, \ldots \vec{x_n}\} \supseteq \mathbb{R}^m$
a distance measure d: $\mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$
a function for computing the mean $\mu : P(\mathbb{R}) \rightarrow \mathbb{R}^m$

From (10):

1. For any given cluster assignment, C, the total cluster variance is minimized with respect to $\{m_1 \ldots, m_K\}$.
2. Given a current set of means $\{m_1 \ldots, m_K\}$ the result is minimized by assigning each observation to the closest cluster mean. Shown in equation 2

$$C(i) = \arg\min_{1 \le k \le K} \| x_i - m_k \|^2$$

3. Iterate step 1 and 2 until assignment to cluster does not change which means the algorithm converges.

b) *Bayesian Learning* - Bayesian learning is based on Bayes' Rule which is expressed as follows: P(B|A) = P(A|B)P(B)/P(A){Manning, 1999 #102}. It is a very useful technique and often performs better than more sophisticated methods(12). Naïve Bayesian classifiers use prior probabilities to classify new objects.

F. TEXT MINING CASES

1. Comparison of Leximancer and SAS Text miner

I compared two programs: SAS Text Miner and Leximancer. SAS Text Miner performs a variety of algorithms and I evaluated the K-means algorithm. Leximancer uses Naïve Bayesian learning for text mining. I evaluated 100 tobacco industry documents from the British American Tobacco Documents Collection (BATCo).

Evaluating SAS Text Miner was made more challenging by the poor OCR. OCR, or optical character recognition, is used to translate a scanned document into text. Figure 1 shows the OCR score for 1,520 BATDA documents, with an average OCR score of 68.9. The OCR score range is 0 to 100, where 0 is no character recognition and 100 is perfect character recognition. Since then, the number of available text documents has exploded, however, at the time of the evaluation I used 100 documents to evaluate SAS text miner. Unfortunately, the results were skewed by the poor OCR and the results were not very meaningful (See Table 4). The results could have been improved by running the documents through a newer OCR program, which might improve the accuracy of the OCR. Also the results could be improved by further refining the dictionary that SAS text miner uses to cluster programs. Adding more stop words or words that are not indexed and clustered could help refine the results.

Table 4: Example of results from clustering 100 documents in SAS Text Miner

Description Terms	Frequency	Proportion of documents
With, have, will, would, test	25	.20
Animal, issue, company, do test	13	.10
Additive, state, smoke, tobacco test	21	.16
Chemical, follow, other, animal, do	20	.16
Most, public, concern, issue, increase	20	.16
Report, smoke, one, more, increase	29	.23

Table 5: Example of results from training 100 documents in Leximancer using Naïve Bayesian learning

Concept	**Absolute count**	**Relative Count**
Testing	121	.24
Cancer	83	.16
Animals	81	.16
Human	80	.16
Studies	77	.15
Risk	45	.08
Effects	30	.06
Chemical	30	.06
Products	30	.06

The comparison shows that while there are similar concepts found (Chemical, Test, and Animal); many of the terms found in the same documents were different. This could be due to the use of different algorithms by the products.

2. Comparison of Results within Leximancer

In order to test the results obtained from Leximancer, the results from a gold standard were compared to a naïve search. The gold standard used in the comparison was internal tobacco industry documents used to write a journal article on child labor policy in the British American Tobacco Documents. The assumption is that humans searching through the documents will be able to detect the most salient document for a research article. There were 17 documents found in British American Tobacco Document Archives that were obtained from the paper. The paper is currently in submission. The concepts from these documents are shown in Table 6. Documents with a Legacy URL were excluded from the comparison because they have not been OCRed, or converted into text.

The naïve search submitted into BATDa was "child labour", which retrieved 965 documents. The concepts derived from these documents are shown in Table 7. These documents were obtained by creating a python script to submit a query to BATDa and retrieve the PDF version of the documents. These documents are in a searchable PDF form.

Table 6: Results from training 17 documents from the British American Tobacco Documents Archives

Concept	Absolute count	Relative Count
Tobacco	38	.97
Labour	33	.85
Work	20	.51
Issue	19	.49
IUF	18	.46
Issues	14	.36
Meeting	14	.36
ITGA	14	.36
Industry	12	.31

Figure 1: Gold standard of 'child labour' search on Leximancer (N=17)

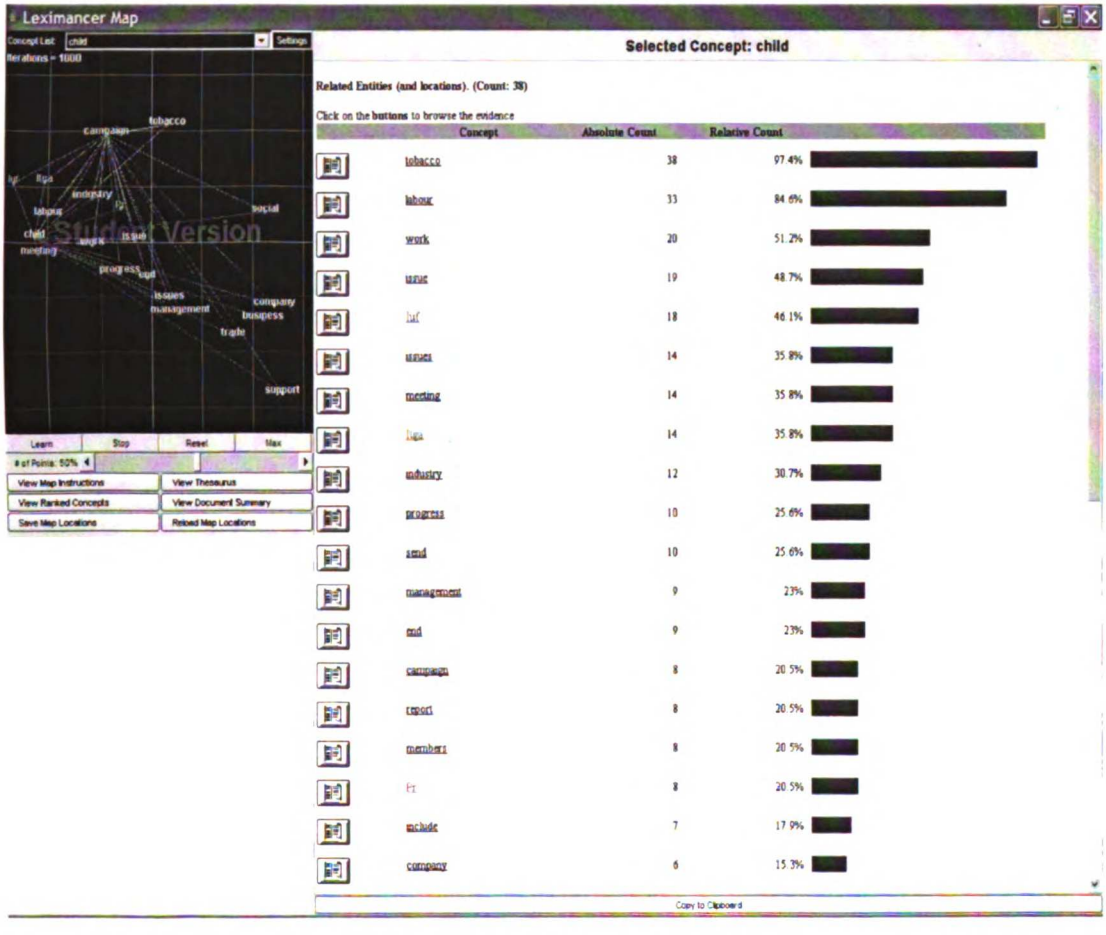
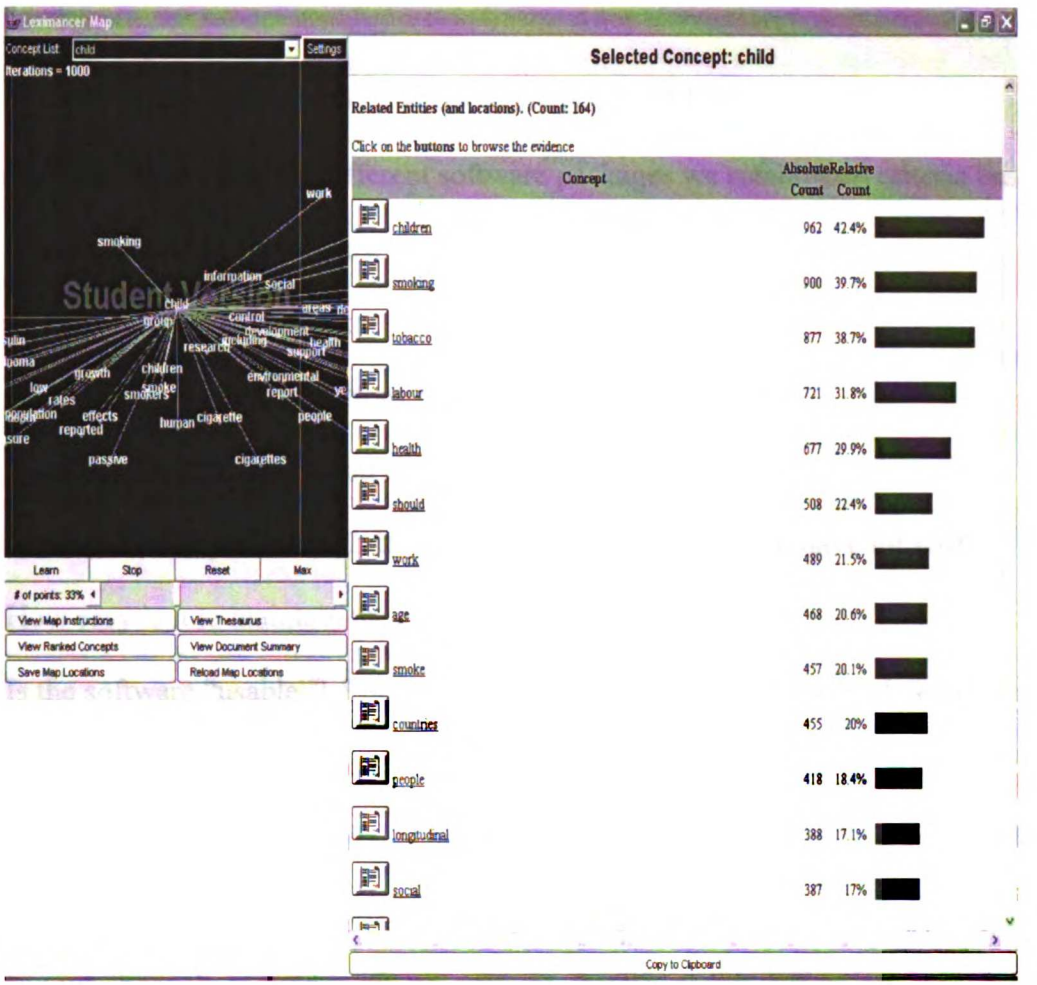


Table 7: Results from training 965 documents from the British American

Tobacco Documents Archives

Concept	Absolute count	Relative Count
Children	962	.42
Smoking	902	.40
Tobacco	857	.39
Labour	721	.31
Health	677	.30
Should	505	.22
Work	489	.22
Age	465	.21
Smoke	457	.20

Figure 2: Naïve search of 'child labour' search on Leximancer (N=965)



The two sets of results show the derivation of similar concepts. However there are only a few concepts in common. The results from the gold standard (see Figure 1) are focused around the researcher's paper topic. In addition the gold standard results show specific relevant acronyms such as the International Union of Food, Agricultural, Hotel, Restaurant, Catering, Tobacco and Allied Workers' Associations (IUF) and International Tobacco Growers' Association (IGTA). The results from the naïve search are broad (see Figure 2). There are many more concepts derived from 965 documents and in general the relative count of the concepts is lower in the naïve search. This indicates that the

documents are less focused and about a variety of topics even though they all contain the term 'child labour'.

G. EVALUATION CRITERIA

In order to evaluate the two different software packages we established criteria based on practical and academic standards.

These criteria are listed below:

1. What is the cost of the software?
2. What are the installation and equipment requirements?
3. Is the product open source? If so what implications does that have for cost?
4. How easy is it to maintain the software?
5. Is the software "usable"? For which groups and under what circumstances?
6. Is there an available user community and support in utilizing the software?

To examine how clustering could help researchers address problems in searching the documents, we evaluated 2 programs 1. SAS text miner and 2. Leximancer
The reason we chose these products is because they represent two different cost models of searching the documents. SAS text miner is a high end product for text mining algorithms in general, not just clustering. Leximancer is a product designed specifically for clustering and is a lower end product that is easily expandable into a research environment.

SAS text miner claims a number of benefits from its product: 1) a reduced time-to-decisions and a more accurate organizational view 2) improved organizational performance and 3) recognize trends and predict business opportunities(14). The impetus

for the product appears to be information overload. There are many promises when it comes to text mining and the usages remain undefined particularly for researchers.

Leximancer is a program that can be used to explore large text collections, make automatic taxonomy discovery, index documents, code open ended surveys and statistically compare documents(15). It is targeted to academic researchers and produces a conceptual map allowing an overview of a large number of documents.

The two software programs are compared in Table 8. Overall, SAS Text Miner is more comprehensive and has extensively tested their data mining algorithms. They are also responsive to their users. However, there are significant barriers to using SAS Text miner, including the additional cost and the high learning curve.

Leximancer is easily accessible and usable without much of a learning curve. It costs less than SAS text miner. However, it only does context analysis and it can take an extensive amount of computer time. One example tested on a computer running XP Pro with a Pentium 4 processor, 1.7 Gigabytes and 512 Mg of RAM ran for at least 12 hours. This lengthy run time can be considered a barrier, given that most researchers would want to run the results on many documents. SAS, on the other hand, is optimized for large datasets and has a substantial advantage.

Table 8: Evaluation criteria for two text mining programs

Evaluation Criteria	SAS text miner	Leximancer
Cost	Academic license is $650 per year plus additional licensing fees ($300)	$50 for students $1,000-3,000 for a site use of up to 19 people
Installation	Complicated – requires multiple CDs on top of SAS base	Easy – one CD – no complicated installation instructions
Equipment requirements	PC or Linux based – requires a fast computer with lots of memory	Minimal requirements
Open source	No	No
Maintenance	SAS fixes the software bugs but it may take sometime	Leximancer fixes the software bugs. It is a small group in Australia but fairly responsive
Usable	Not easy without extensive training	Easy to use and has had usability tests published
User community and support	Extensive user community and support on newsgroups is helpful	Small user community

H. CONCLUSIONS

The increase of internal tobacco documents parallels the information overload in society. Searchers of the documents would benefit from development of text mining research tools such as Leximancer and SAS text miner. These programs use different

methods but assist users in similar activities such as identifying themes, linking concepts and summarizing documents. All of these activities would be useful to researchers of the internal tobacco industry documents.

However, the results show that the two programs provide different results and do not necessarily match a gold standard. The first test compared the same number of documents text mined in two different programs. The results were not similar. They had a different number of terms and only 3 concepts matched. There could be many reasons for this dissimilarity. First the programs were not optimized. Stop words used were basic and appeared as a concept in the results. The stop words could be tailored for the tobacco documents and this would improve the results. Second, the programs use two different algorithms and it is possible that one algorithm performs better than the other. There are articles that suggest that Naïve Bayesian classifiers perform just as well or better than clustering algorithms (stat soft ref and data clustering review).

The second test compared two sets of documents on child labour. The first set was derived from a submitted research paper. The second was obtained from a naïve search. The results obtained show an overlap of some concepts; however most are unique to the test.

This suggests that text mining documents do not obtain the same documents as a human does. However, it also shows that there may be some unexplored concepts within the topic. Leximancer provides a useful overview of a large query and a potential starting point for those who are overwhelmed by their search. It could also be useful for

researchers who want one final look at the concepts derived from the documents. In addition it provides useful summaries of the documents.

For a more sophisticated researcher, SAS Text Miner offers a variety of stable algorithms that are optimized and verified. There are some distinct advantages to working with a program like SAS Text miner which include shorter run times and a wider user community. However, many of the algorithms used are not easily accessible by the casual user. In such a case, Leximancer would be more useful. Overall text mining is beginning to cross over from a solely computer science discipline to more popular use. The internal tobacco industry documents would benefit from adapting either SAS Text miner or Leximancer as an additional research tool.

I. REFERENCES

1. Hearst M. Untangling Text Data Mining. In: Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics; 1999 June 20-26; University of Maryland; 1999.
2. Hudson D, Cohen M. Neural Networks and Artificial Intelligence for Biomedical Engineering. New York: IEEE Press Marketing; 2000.
3. Merrill GH. The Babylon project: toward an extensible text-mining platform. IT Professional 2003;5(2):23-30.
4. Fernandez G. Data Mining Using SAS Applications. Boca Raton: Chapman & Hall/CRC; 2003.
5. Lyman P, Varian H. How Much Information. 2003 November 23, 2003 [cited 2004 February 17]; Available from: http://www.sims.berkeley.edu/how-much-info-2003
6. Lyman P, Varian H. How Much Information. Berkeley: UC Berkeley; 2003 October 27.
7. Kirsh D. A Few Thoughts on Cognitive Overload. Intellectica 2000;1(30):19-51.
8. Jackson P, Moulinier I. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. Amsterdam/Philadelphia: John Benjamins Publishing Company; 2002.
9. Kachigan. Mulitvariate Statistical Analysis. NY: Radius Press; 1982.
10. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag; 2001.

11. Manning C, Schutze H. Foundations of Statistical Natural Language Processing. Boston: Massachusetts Institute of Technology; 1999.

12. Stat Soft Inc. Electronic Statistics Textbook-Naive Bayes Classifier Introductory Overview. 2004 [cited 2005 September 8]; Available from: http://www.statsoft.com/textbook/stnaiveb.html

13. Steinbach M, Karypis G, Kumar V. A Comparison of Document Clustering Techniques. Minneapolis: University of Minnesota; 2000 May 23. Report No.: TR 00-034.

14. SAS Text Miner. 2005 [cited 2005 August 31]; Available from: http://www.sas.com/technologies/analytics/datamining/textminer/

15. Smith A. Leximancer. [cited August 31 2005]; Available from: http://www.leximancer.com/overview.html

PART III: FUTURE DIRECTIONS

Chapter 7.

Conclusions

A. WHAT THIS STUDY ADDS

This research presents a case study of how to approach a large public health informatics problem through a new method called people-centered public health informatics.

Increasingly we face datasets that are massive, complex, and content-rich, and they are not always easy from which to extract meaningful pieces of information. We know that we can produce research papers, which are one of the knowledge units of interest for researchers, but it takes much time and energy for even a single paper to be produced. The internal tobacco industry documents pose problems representative of those that we will face in the future in public health.

From the single paper, we move towards an amalgamation of knowledge about a particular content field, in this case, tobacco control. From content knowledge, comes dissemination to teachers, to researchers, and to advocates as well as others. Diffusion of innovation can be a slow process and sometimes we battle diffusion of incorrect knowledge, or results from implementing technologies that are inappropriate for the situation (1). These diffusions can be seen as set backs in progress. To help prevent these setbacks, we start simply with the needs of people, and then ask how technology can help.

One major finding from this dissertation is that there are some sub groups of potential document users who are not using the documents as much as they could be. These groups include advocates, lawyers, and teachers. These groups could potentially be targeted by new interfaces such as those explored in Appendix D.

B. TEXT MINING

Text mining remains in the form of a black box, which few can understand. This dissertation attempted to use one form of text mining, clustering, on internal tobacco industry documents. The conclusion is that clustering would be useful for researchers using the internal tobacco industry documents. My evaluation of 2 programs suggests that either SAS text miner or Leximancer, which performs Bayesian learning, would be helpful to the researchers. A formal usability test would be the next step to take in implementation.

While text mining is still a nascent field, it is becoming useful for applied researchers to explore. This dissertation was the first attempt to implement a text mining system on a corpus for applied research purposes. In order to further the field of text mining, more ventures of this sort should be tested and for different user groups.

Another finding is that in the case of the internal tobacco industry documents, tools such as those implemented in Leximancer or SAS text miner are of help. In order to make further progress with the internal tobacco industry documents, we need to increase the metadata tags for each document to include subject. Also we need to improve many existing metadata tags.

Such a situation seems ideal for computerized solution, as a computer never tires of reading, summarizing, or categorizing data. If the proper algorithms were set in place, a computer can crunch away without needing to eat, sleep or take breaks. It obviously has none of the cognitive limitations as humans do. However, the state of the art in research is not quite adequate and each topic of computerized reading, summarizing, and categorizing is a large research topic within itself.

Public health informatics practioners need to develop plans in how to approach a large informatics problem that focus on multi-disciplinary solutions drawing from human-computer interaction, cognitive science, bioinformatics, computer science, and information and library science.

As we face exponential growth of data, we face the challenge of how to turn that data into information and then later into knowledge which we can use to inform policy and ultimately improve public health. We are far from a gold standard where we can efficiently turn a dataset into a knowledge production machine.

C. RECOMMENDATIONS

Based on the research completed in this dissertation, I would recommend a new approach to projects based on large datasets of either textual data or numerical data. First I would recommend that a complete needs assessment be done before the information technology is developed. The needs assessment should involve knowledgeable social scientists that are familiar with techniques such as focus groups, surveys, and interviews. The needs of the people who are expected to utilize or who could utilize the information

should be considered first and foremost. This may seem to be an obvious observation; however, it has been neglected time after time.

Another approach that should be taken in the near future is knowledge modeling of tobacco control in software. Fortunately, the tobacco control community has a first step at doing this through the UCSF Tobacco Thesaurus. I modified it for use in another interface called Flamenco (see Appendix D). This knowledge modeling would facilitate document categorization and add to the possibilities of what could be done with the documents.

In 1989, Mark Musen developed a tool to facilitate knowledge-acquisition called Protege (2). For the purposes of Protégé, ontology is defined as the "explicit formal specifications of terms in the domain and relations among them". (3)

The following reasons for developing an ontology have been elucidated in the paper *Ontology Development 101: A Guide to Creating Your First Ontology"(3):*

- To share a common understanding of the structure of information
- To enable reusable domain knowledge
- To make assumptions within a domain explicitly represented
- To separate domain knowledge from operational knowledge
- To analyze domain knowledge

Understanding of knowledge and the structure of information is one of the most common reasons to develop an ontology (3). Already in the example of internal tobacco documents we know of a number of projects that use different XML (Extensible Markup language) schema to represent documents. Tobacco Documents Online explicitly

describes their XML schema, called Tobacco Citation Markup Language or TCML (4). The addition of the thesaurus terms into Protégé would advance the informatics of tobacco control.

Given this research, one of the most useful strategies to improve accessibility and usability of the documents would be to allow them to be indexed by Google, or a similar company. As shown from the respondents to the LTDL survey, most people use Google as their primary search engine (68%) and it is a usable system. Because most people who have access to a computer know about and use Google, it would be an excellent baseline website to display and use the documents.

There are a few reasons why the documents are not yet accessible by Google. First, most of the documents are images stored as .tif or .gif files. Google does not index these files. In order to implement the indexing by Google, we would need to scan and convert all documents into text. This process is currently underway.

However there are some potential problems with using Google. First there are many junk documents, such as documents that are receipt or brief emails, with no attachments. It could be frustrating for users to retrieve too many documents. In addition, there is concern that Google is a proprietary system. However, there are application program interfaces that could be used to write user interfaces that would plug into Google that would be specific for the documents (5).

The tobacco industry documents will remain a useful database for analyzing industry behavior for years to come. It is to the benefit of everyone to continue to investigate new methods for searching this remarkable resource.

C. REFERENCES

1. Rogers EM. Diffusion of innovations. New York: Free Press of Glencoe; 1962.

2. Musen MA. Automated Generation of Model-Based Knowledge-Acquisition Tools. London: Pitman; 1989.

3. Noy N, McGuinness D. Ontology Development 101: A Guide to Creating Your First Ontology. [cited 2004 10/8]; Available from: http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

4. Tobacco documents online. [Web site] 2002 [cited 2002 March 5]; Available from: http://tobaccodocuments.org/

5. Google. [cited 2005 September 7]; Available from: http://www.google.com/apis/

Appendix 1.
Survey of the Use of the Legacy Tobacco Documents Library

The purpose of this questionnaire is to explore why users search the Legacy Tobacco Documents library. We also want to find out what you like and dislike about searching the tobacco industry documents. Your response will help us develop new methods for searching the documents.

Your responses are strictly confidential. We are concerned about your privacy and statistics will only be reported on the group level. No information will be collected that can identify you with your responses.

Thank you for your response!

Part I: Demographics

1. What is your gender?
   - ❑ Female
   - ❑ Male

2. How old are you?

   _____

3. What is your race / ethnicity?
   - ❑ Black
   - ❑ White
   - ❑ Latino/a
   - ❑ Asian / Pacific Islander
   - ❑ Native American
   - ❑ Other _____
   - ❑ Prefer not to answer

4. What is the highest education level you have completed?
   - ❑ Less than high school
   - ❑ High school graduate
   - ❑ Some college
   - ❑ College graduate
   - ❑ Master's degree
   - ❑ Ph.D./ MD
   - ❑ Prefer not to answer

5. What is your occupation?
   - ❑ Academic researcher
   - ❑ Librarian
   - ❑ Lawyer
   - ❑ Law clerk

- ❑ Tobacco control advocate
- ❑ Teacher
- ❑ Public health official / employee
- ❑ Voluntary health organization employee
- ❑ Congressional aide
- ❑ Journalist
- ❑ Policy maker
- ❑ Student (Please specify level) _____
- ❑ Other (Please specify) _____

6. Where do you live? (*Pull down menu for City, State and Country*)
   City _____ State ____ Country _____

Part II: Usage

7. What is your purpose for searching the Legacy Tobacco Documents Library? (*Please check all that apply*)
   - ❑ Academic research
   - ❑ Personal interest
   - ❑ Public health advocacy
   - ❑ Litigation
   - ❑ Teaching tool
   - ❑ Media story
   - ❑ Legal research
   - ❑ Advertising campaign
   - ❑ Testimony at public hearing
   - ❑ Other_____

7. Where do you search the Legacy Tobacco Documents Library? (*Please check all that apply*)
   - ❑ Work
   - ❑ Home
   - ❑ Library
   - ❑ Other_____

8. If this is not your first visit, what was the date that you first used the Legacy Tobacco Documents Library?
   Month _____ Year _____ (*Pull down menu*)
   - ❑ Don't know

9. How did you find the Legacy Tobacco Documents Library?
   - ❑ Link from another website
   - ❑ Colleague/friend told me about it
   - ❑ Search engine
   - ❑ Citation in journal, book, or newspaper article

- ❑ Heard about it at a professional conference
- ❑ Other _____

10. Why do you use the UCSF Legacy Tobacco Documents Library? (*Please check all that apply*)
    - ❑ Search capabilities
    - ❑ Ease of use
    - ❑ Speed of searching
    - ❑ Don't know any other place to find tobacco industry documents on-line
    - ❑ Familiarity with the interface
    - ❑ Indexes and abstracts are helpful resources
    - ❑ Other_____

11. Did you find what you were looking for?
    - ❑ Yes
    - ❑ No
    - ❑ Don't know
    - ❑ Just browsing

12. Which of the UCSF Tobacco Control Archives do you routinely search?
    - ❑ UCSF Brown and Williamson/Mr. Butts Documents
    - ❑ Mangini vs. RJ Reynolds
    - ❑ California documents from the State of Minnesota Depository
    - ❑ Legacy Tobacco Documents Library
    - ❑ British-American Tobacco Documents (BATCO)

13. Please check other websites you use to search the tobacco documents:
    - ❑ Philip Morris Documents website
    - ❑ RJ Reynolds Tobacco Documents website
    - ❑ Tobacco Documents online
    - ❑ Brown and Williamson
    - ❑ CDC documents collection
    - ❑ Lorillard
    - ❑ Tobacco Archives
    - ❑ Council for Tobacco Research Documents website
    - ❑ Other (Please specify) _____
    - ❑ None of the above

14. Have you had any problems using the UCSF Legacy Tobacco Documents Library website?
    - ❑ Yes
    - ❑ No

15. If yes, what problems have you had using the website? (*Please check all that apply*)

- ❑ Slow connection
- ❑ Site was down
- ❑ Too many documents retrieved during searches
- ❑ Search capabilities not flexible
- ❑ Other technical barriers (please specify)

  _____

- ❑ Other (please specify)

  _____

16a. What are the most useful features of the UCSF Legacy Documents Library?

_____
_____
_____
_____
_____

b. What features of the website do you find most difficult to use?

_____
_____
_____
_____
_____

17. What additional features on the Legacy Tobacco Documents Library would facilitate your searches?

_____
_____
_____
_____

18. In the past six months, how often have you visited the Legacy Tobacco Documents Library?

- ❑ First visit
- ❑ Daily
- ❑ Weekly
- ❑ Twice a month
- ❑ Once a month
- ❑ Less than once a month

Please answer questions 19-21 if you search the documents frequently (greater than once a month - this will be a skip pattern), otherwise skip to Part III.

19. What attributes of a document are most important to you when searching online:
    - ❑ Personal Names

- ❑ Authors
- ❑ Organization Names
- ❑ Dates
- ❑ Cigarette brands
- ❑ Thesaurus terms
- ❑ Bates Numbers
- ❑ Other _____

20. What are your methods for organizing the documents that you find during a search?

_____
_____
_____
_____

21. How do you like to organize *(or sort?)* the documents you search online? (*Check all that apply*):

- ❑ By document title
- ❑ By document type
- ❑ By date
- ❑ By named persons, authors, recipients
- ❑ By named organizations, corporate authors, recipients
- ❑ By search terms
- ❑ By theme
- ❑ Other _____

22. Have you ever had training in how to search the tobacco industry documents?

- ❑ Yes
- ❑ No

23. Please feel free to add any additional comments about searching the tobacco industry documents using the Legacy website.

_____
_____
_____
_____

(*We could cut the survey here)*

Part III: Computer access and experience

24. Do you have access to a computer at home?

- ❑ Yes
- ❑ No

25. Do you have Internet access to a computer at home?

- ❑ Yes

- ❑ No

26. If so, what speed is your Internet access at home?
    - ❑ 28.8K
    - ❑ 56K
    - ❑ 128K
    - ❑ DSL
    - ❑ T1
    - ❑ Don't know

27. Do you have access to a computer at work?
    - ❑ Yes
    - ❑ No

28. Do you have Internet access to a computer at work?
    - ❑ Yes
    - ❑ No

29. If so, what speed is your Internet access at work?
    - ❑ 28.8K
    - ❑ 56K
    - ❑ 128K
    - ❑ DSL
    - ❑ T1
    - ❑ Don't know

30. Which search sites do you use?
    - ❑ Google
    - ❑ MSN
    - ❑ Yahoo
    - ❑ Alta-Vista
    - ❑ Copernic
    - ❑ Lycos
    - ❑ Other _____

31. Rate your ability to find topics that you are interested in on the Internet:
    - ❑ Excellent
    - ❑ Very good
    - ❑ Good
    - ❑ Fair
    - ❑ Poor

32. How many web searches do you conduct per month?
    - ❑ Less than 1 search
    - ❑ 1-5 searches
    - ❑ 5-10 searches

- ❑ 10-20 searches
- ❑ Over 20 searches

Thank you very much for your participation!!!
If you have any questions about the survey, please contact:

Martha Michel, M.S.
martham@itsa.ucsf.edu
Department of Clinical Pharmacy
3333 California St., Suite 420
Box 0613, Laurel Heights
San Francisco, CA 94143

Appendix 2.
Survey of the Use of the UCSF Tobacco Control Archives/ BATCO collection

The purpose of this questionnaire is to explore why users search the UCSF Tobacco Control Archives/ BATCO collection. The following document collections are referred to in this survey: Brown & Williamson Collection, Joe Camel Campaign: Mangini v. R. J. Reynolds Tobacco Company Collection, California Documents from the State of Minnesota Depository, British American Tobacco, and Tobacco Litigation Documents. Also, please see our survey (***link to survey***) on the Legacy Tobacco Documents Collection. We want to find out what you like and dislike about searching the tobacco industry documents. Your response will help us develop new methods for searching the documents.

Your responses are strictly confidential. We are concerned about your privacy and statistics. No information will be collected that can identify you with your responses.

Thank you for your response!

Part I: Demographics

1. What is your occupation?
   - ❑ Academic researcher
   - ❑ Librarian
   - ❑ Lawyer
   - ❑ Law clerk
   - ❑ Tobacco control advocate
   - ❑ Teacher
   - ❑ Public health official / employee
   - ❑ Voluntary health organization employee
   - ❑ Congressional aide
   - ❑ Journalist
   - ❑ Policy maker
   - ❑ Student (Please specify level) _____
   - ❑ Other (Please specify) _____

2. Where do you live? (*Pull down menu for City, State and Country*)
   City _____ State ____ Country _____

Part II: Usage

3. What is your purpose for searching the UCSF Tobacco Control Archives? (*Please check all that apply*)
   - ❑ Academic research
   - ❑ Personal interest
   - ❑ Public health advocacy
   - ❑ Litigation

- ❑ Teaching tool
- ❑ Media story
- ❑ Legal research
- ❑ Advertising campaign
- ❑ Testimony at public hearing
- ❑ Other_____

4. Where do you search the UCSF Tobacco Control Archives? (*Please check all that apply)*
   - ❑ Work
   - ❑ Home
   - ❑ Library
   - ❑ Other_____

5. In the past six months, how often have you visited the UCSF Tobacco Control Archives?
   - ❑ First visit
   - ❑ Daily
   - ❑ Weekly
   - ❑ Twice a month
   - ❑ Once a month
   - ❑ Less than once a month

6. If this is not your first visit, what was the date that you first used the UCSF Tobacco Control Archives?

   Month _____ Year _____ (*Pull down menu*)
   - ❑ Don't know

7. How did you find the UCSF Tobacco Control Archives?
   - ❑ Link from another website
   - ❑ Colleague/friend told me about it
   - ❑ Search engine
   - ❑ Citation in journal, book, or newspaper article
   - ❑ Heard about it at a professional conference
   - ❑ Other _____

8. Which of the UCSF Tobacco Control Archives collections do you use?
   - ❑ UCSF Brown and Williamson/Mr. Butts Documents
   - ❑ Mangini vs. RJ Reynolds
   - ❑ California documents from the state of Minnesota Depository
   - ❑ Legacy Tobacco Documents Library
   - ❑ British-American Tobacco Documents (BATCO)

9. How frequently do you use the each of the following UCSF Tobacco Control Archives collections?

Website	Frequency of searching			
	Daily	Weekly	Occasionally	Never
Legacy website				
Brown and Williamson Collection				
Joe Camel Campaign: Mangini v. R. J. Reynolds Tobacco Company Collection				
California Documents from the State of Minnesota Depository				
Documents from the British-American Tobacco Company				
Tobacco Litigation Documents				

10. Do you browse any of the following collections in the UCSF Tobacco Control Archives?

Collection	Yes	No
Joe Camel Campaign: Mangini v. R. J. Reynolds Tobacco Company Collection		
Documents from the British-American Tobacco Company		
Brown and Williamson Collection		

11. Why do you use the UCSF Tobacco Control Archives? (Please check all that apply)
    - ❑ Search capabilities
    - ❑ Ease of use
    - ❑ Speed of searching
    - ❑ Don't know any other place to find tobacco industry documents on-line
    - ❑ Familiarity with the interface
    - ❑ Indexes and abstracts are helpful resources
    - ❑ Other_____

12. Did you find what you were looking for?
    - ❑ Yes
    - ❑ No
    - ❑ Don't know
    - ❑ Just browsing

13. Please check the other websites you use to search the tobacco documents:

- ❑ Legacy Tobacco Documents Library
- ❑ Philip Morris Documents website
- ❑ RJ Reynolds Tobacco Documents website
- ❑ Tobacco Documents online
- ❑ Brown and Williamson
- ❑ CDC documents collection
- ❑ Lorillard
- ❑ Tobacco Archives
- ❑ Council for Tobacco Research Documents website
- ❑ Other (Please specify) _____
- ❑ None of the above

14. Have you had any problems using UCSF Tobacco Control Archives website?
    - ❑ Yes
    - ❑ No

15. If yes, what problems have you had using the website? (Please check all that apply)
    - ❑ Slow connection
    - ❑ Site was down
    - ❑ Too many documents retrieved during searches
    - ❑ Search capabilities not flexible
    - ❑ Other technical barriers (please specify)

      _____
    - ❑ Other (please specify)

      _____

19a. What are the most useful features of the UCSF Tobacco Control Archives?

_____
_____
_____
_____
_____

b. What features of the website do you find most difficult to use?

_____
_____
_____
_____
_____

20. What additional features on the Tobacco Control Archives would facilitate your searches?

21. Have you ever had training in how to search the tobacco industry documents?
    - ❑ Yes
    - ❑ No

22. Any comments?

(We could cut the survey here)

Thank you very much for your participation!!!
If you have any questions about the survey, please contact:

Martha Michel, M.S.
martham@itsa.ucsf.edu
Department of Clinical Pharmacy
3333 California St., Suite 420
Box 0613, Laurel Heights
San Francisco, CA 94143

Appendix 3.
Flamenco and Future Possibilities

Specific Aims

The objective of this research is to evaluate new methods of searching the tobacco industry documents using three informatics programs. We would like to see if supplemental tools to searching the internal tobacco industry documents are helpful for users. We are investigating whether these programs will enable people to examine the industry in a different way and perhaps make connections that were not easily made using a traditional search engine.

While access to information has increased, information overload has become a barrier to effectively using the data. For example, it is estimated that there are about 7 million internal tobacco industry documents in the UCSF Legacy Library and the Tobacco Control Archives and this number is steadily growing (1). In addition to the Legacy Library, the British American Tobacco documents are in the process of being scanned and converted into text using optical character recognition software. It is estimated that there are 8 million documents in the British American Tobacco document collection, located in Guildford England.

In order for the users of the Legacy Tobacco Documents Library to sort through all this information, we need to investigate new methods of searching and displaying electronic text. The internal tobacco industry documents can be used to find out the tactics of the tobacco industry and help improve public health (2).

My specific aim for this proposal is:

1. To compare a new web-based interface, Flamenco and a commercial program, Leximancer, to the web based interface currently used to search the British American Tobacco Documents Archive (BATDa). We will compare the programs' usability for complex searching. The Flamenco interface uses hierarchical metadata to facilitate searching and browsing the document collection. Leximancer is a commercial program used for clustering documents. BATDa is a collection of internal tobacco industry documents produced from the British American Tobacco Company. The paper depository in Guildford, England is estimated to contain over 8 million documents.

Despite the considerable effort and resources that have gone into establishing and maintaining the tobacco industry documents, there have been no studies of methods to search the documents in a more effective manner. The documents offer an opportunity to learn about how people search through vast amounts of information. Thus, our findings could help people search more effectively in other large text databases such as PubMed or the Cochrane Library, a large electronic database of systematic reviews.

Background

In 1994, previously secret, internal tobacco industry documents were at the center of several analyses and research studies that changed public health (3). Recent litigation and the Master Settlement Agreement of 1998 (4) have made millions of tobacco industry internal documents available on the (5). As required by the Master Settlement Agreement, tobacco companies maintain websites where many of the documents are

located. Government and private funding sources have begun to support the archiving and indexing of these documents on other Internet sites(1, 6).

Multiple locations on the Internet contain different types of internal tobacco industry documents. However, the University of California at San Francisco (UCSF) library is attempting to create an Internet archive of all the tobacco industry documents. The UCSF library archives the Legacy collection, which as of December 9, 2004, is 7.2 million documents with an estimated 41.8 million pages; approximately 1.5 terabytes of data (1). In addition, the UCSF library contains four separate online collections, called the Tobacco Control Archives (the Brown and Williamson Collection, Joe Camel Campaign: Mangini v. R.J. Reynolds Tobacco Company Collection, California Documents from the State of Minnesota Depository, and the British American Tobacco Company (BATCo) documents).

The Flamenco project (FLexible Access to MEtadata in Novel COmbinations) is one example of a method that uses metadata to aid in both browsing and searching large document collections (7). Metadata is information or data about the data being studied. The browser dynamically generates query previews to guide the user about how to further narrow their search (8). The Flamenco project was established specifically to deal with complex information collections. With the assistance of a computer program, documents from the British American Tobacco Company have been manually indexed with rich metadata terms. The index terms that were used are based on the UCSF Tobacco Thesaurus (9). Thus, using these indexed terms, we will develop a new multifaceted search engine based on Flamenco for the tobacco industry documents.

The preliminary studies of the Flamenco Interface have been encouraging. The first study interviewed nine subjects about their preferences in searching using either the Flamenco interface or a traditional tree interface. All of the subjects preferred the Flamenco interface matrix view to the single tree view, which is similar to a standard search engine interface (10). Studies of Flamenco suggest that people who are motivated searchers are more likely to prefer Flamenco for searching. They are also more likely to be experienced searchers. In this project, we will also test additional methods that might be more attractive to less experienced searchers such as BatDa.

Previous Studies

Survey of Legacy and TCA:

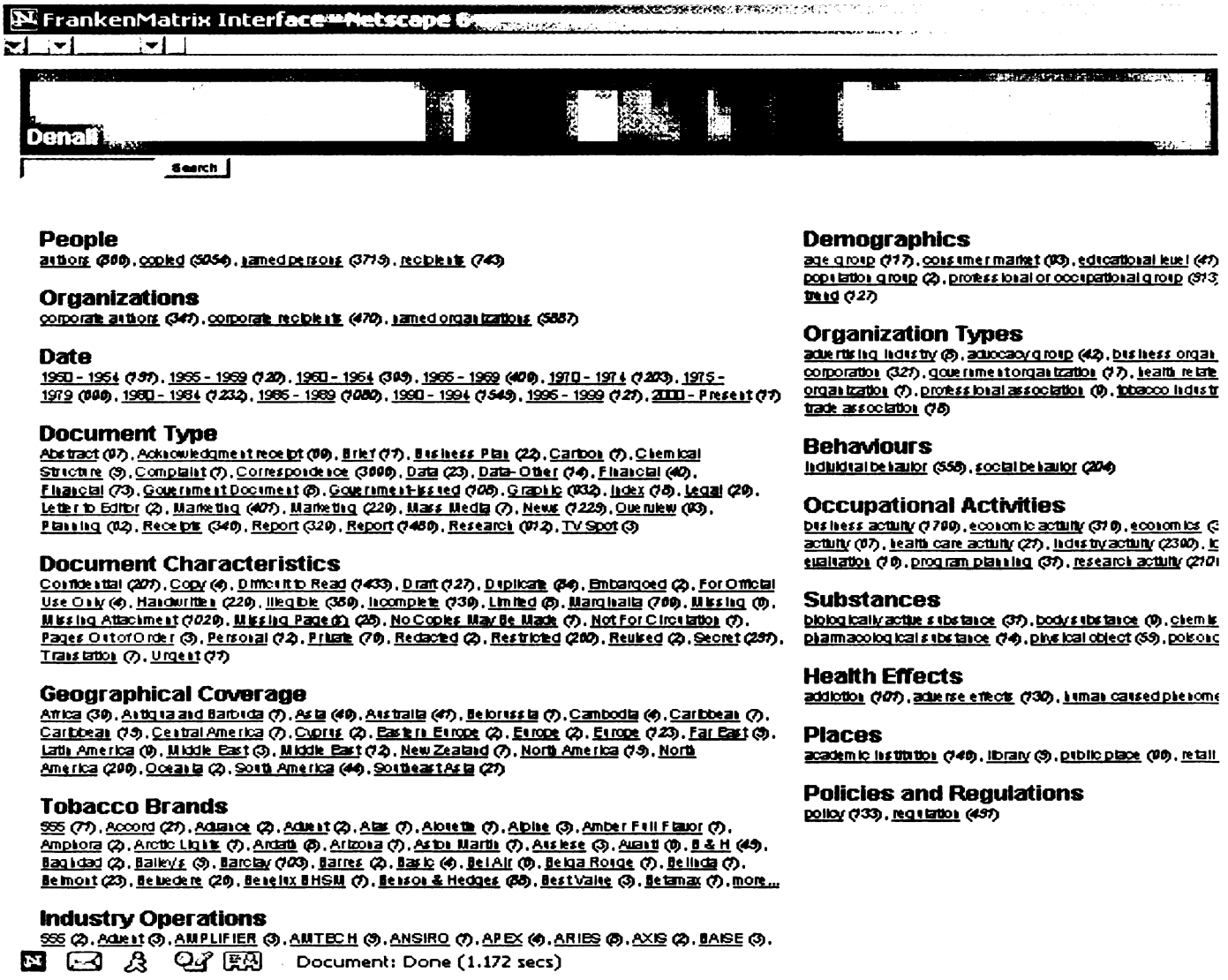
We have conducted a survey to determine: 1) who uses the internal tobacco industry documents, 2) why the documents are used, 3) barriers to searching and 4) suggestions for improvement. Among 165 respondents, personal interest was the primary reason reported for searching the Legacy website (45%, 74/165) followed by research (41%, 67/165), advocacy (18%, 29/165), and other (18%, 29/165).

We also conducted an online survey of users of the Tobacco Control Archives with the same objectives. We determined among 99 respondents that 53% searched the documents for academic research (52/99), followed by other at 31% (31/99) and personal interest at 25% (25/99). Sixteen percent used TCA as a teaching tool (16/99). Our results offer an opportunity to modify the site for a wider range of users.

Flamenco

We have mounted an example of the Flamenco interface on (flamenco.berkeley.edu/tobacco). First, we had to collapse terms to make sure each category had an adequate number of documents. We have added 13,000 documents to the Flamenco interface.

Figure 1:



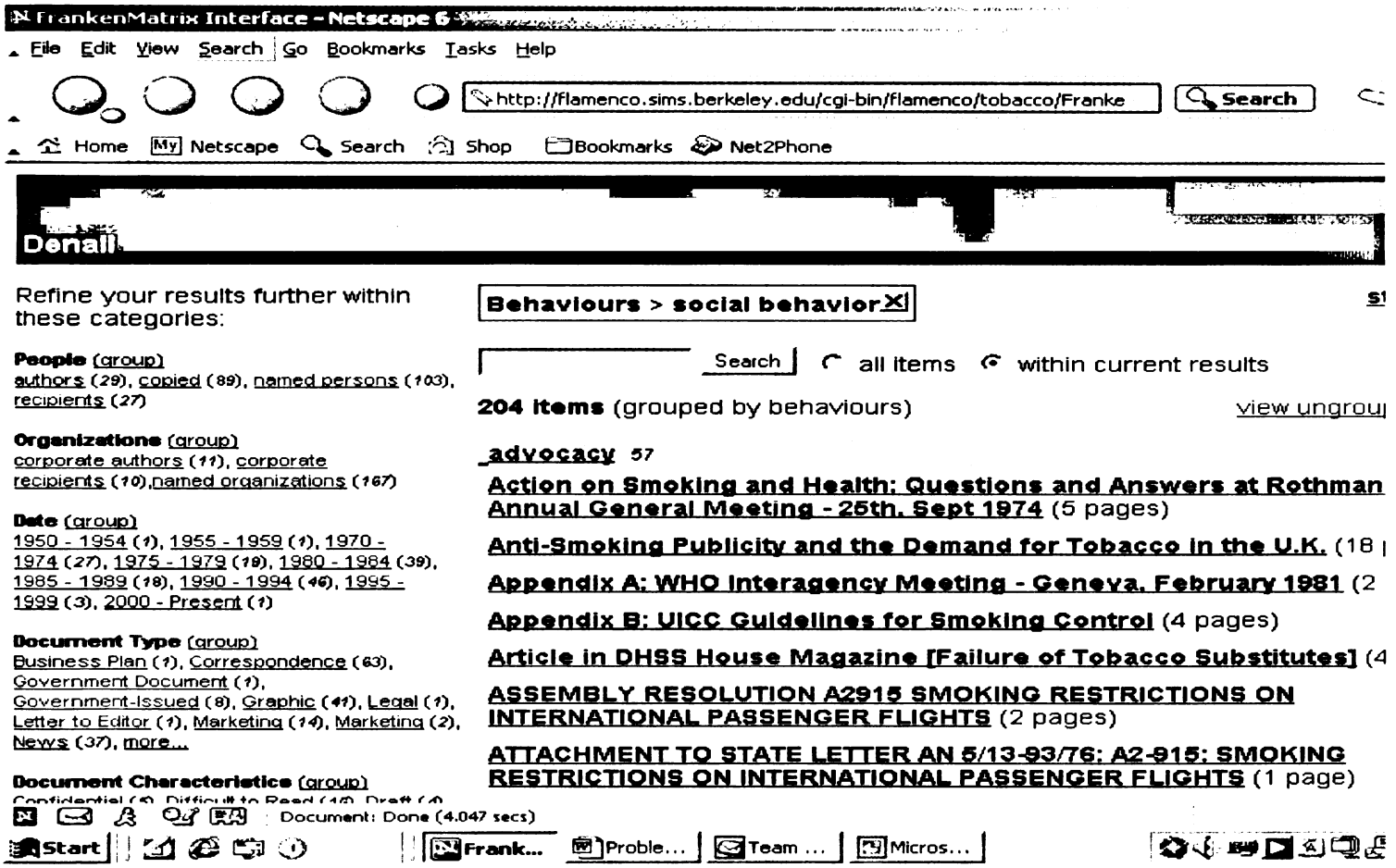


Figure 2: Screenshot of Tobacco Flamenco

Comparison of various text data mining programs

We conducted a survey of text data mining programs to determine the advantages and disadvantages of various programs and found that text mining could be useful for researchers of the tobacco documents in a number of areas. Clustering of documents based on automatic categorization assignments would be useful for both experienced and novice users. The novice users could use clustering to get an overview of a particular area of interest. Experienced users could use clustering to help them draw connections that they previously did not think of or use it to narrow down a search with many hits.

Methods

AIM 1-- To compare a new web-based interface, Flamenco and a commercial program, Leximancer, to the web based interface currently used to search the British American Tobacco Documents Archive (BATDa). We will compare the programs' usability for complex searching.

1a) Methods

There are 3 products that we are proposing to test, Leximancer, Flamenco and BatDa for usability. Each participant will test all three programs.

We will recruite 47 subjects from the UCSF community to compare 3 methods of searching using the attached questionairre. We determined the number of subjects to recruit by using a sample size calculation assuming that the power is equal to 0.8 and the alpha, or probability of making a type I error is $p<0.05$. We will add 5 additional subjects to account for attrition for a total of 52 subjects.

The study will first randomize the methods of searching and ask the users to search for the answers to 5 questions per computer program. When they are done searching for the questions, the users will be asked to fill out a brief questionaire about each program that they used. The study will take about one hour.

For the evaluation, users will be asked to find documents to answer 5 open-ended questions by using the standard Legacy interface and the Tobacco Flamenco interface. We will compare ease of use, interest, and time to find document for the three interfaces(11). When a person finds the document with which they are satisfied, they will save it on the laptop for evaluation. The document retrieval test will be timed but the

user will not be aware of the time to avoid the stress of a testing situation. The usability study conducted after the document retrieval tests will not be timed.

After the study is completed the subjects will receive an Amazon.com gift certificate for $10.

1b) Recruitment

We will recruit people for this study by asking students in a Health Policy class, tobacco control class, and a study design class. In addition we will ask people affiliated with the Center for Tobacco Control Research and Education and the Center for Research Integrity and Science Policy to participate. If we do not have enough subjects from that population, we will ask the general public by posting flyers around campus.

1c) Content of Questionnaire

The surveys for evaluating the different programs will take approximately 10-15 minutes to complete. Questions will include background information about the survey respondent such as their age, gender, and education level. We will ask the user about ease of use, intuitiveness of the interface, and difficulty finding specific answers to the questions. The five questions will be as follows.

1. The British American Tobacco company documents have been made available through a website at UCSF called the British American Tobacco company Documents Archives. There is also a warehouse that contains the original paper documents. What is the name of the location of the paper depository in England?

2. What does the FCTC stand for and can you find a few relevant documents on British American Tobacco's stance on the FCTC?
3. Name two consultants that have worked in Latin America for British American Tobacco.

irviews will be tape recorded and transcribed. Ms. Michel will analyze the qualitative data using content analysis and will categorize the data by subject

Aggregated quantitative statistics and qualitative findings from the surveys will be submitted for publication in a peer-reviewed journal.

We will enter the numerical data into Stata 8.0 for analysis (12). First, all variables will be explored in a bivariate table. Nominal data will be analyzed using the Chi square test testing for significance level of $p<.05$ (13).

We will look at the number of questions a user answered correctly by the type of interface they were using to search the documents.

References

1. University of California San Francisco. The British-American Tobacco Document Collection. Access Date: January 13, 2002. URL: http://www.library.ucsf.edu/tobacco/batco/.

2. Bero L. Implications of the tobacco industry documents for public health and policy. Annual Rev Public Health 2003;24:267-88.

3. Glantz SA, Slade J, Bero L, Hanauer P., Barnes DE. The Cigarette Papers [Online]. Access Date: March 4, 2002. URL: www.library.ucsf.edu/tobacco/cigpapers/.

4. The State of Minnesota and Blue Cross/Blue Shield of Minnesota v. Philip Morris Inc., et al. Access Date: January 13,2004. URL: http://www.naag.org/issues/issue-tobacco.php.

5. Chapman S, Cummings K. Impact of new technologies in tobacco control: call for papers. Tobacco Control 1998;7(3):222.

6. Center for Disease Control. Tobacco Industry Documents. Access Date: March 05, 2002. URL: http://www.cdc.gov/tobacco/industrydocs/index.htm.

7. Hearst M, Elliott A, English J, Sinha R, Swearingen K, Yee K. Finding the Flow in Website Search. Communications of the ACM 2002;45(9):42-49.

8. Elliott A. Flamenco Image Browser: Using Metadata to Improve Image Search During Architectural Design. In: Proceedings of the ACM CHI; 2001; Seattle, WA; 2001. p. 2.

9. University of California San Francisco. UCSF/ANRF Tobacco Documents Thesaurus. Access Date: December 11, 2002. URL: http://www.library.ucsf.edu/tobacco/thesaurus.html.

10. English J, Hearst M, Sinha R, Swearingen K, Yee K. Flexible Search and Navigation using Faceted Metadata. Submitted for Publication.

11. Dumas JS, Redish JC. A Practical Guide to Usability Testing. Portland, OR: Cromwell Press; 1999.

12. StataCorp. Stata Statistical Software V. 8. College Station, TX. 2004.

13. Hosmer D, Lemeshow S. Applied Logistic Regression. New York: Wiley; 1989.

Appendix 4.
The Tobacco Thesaurus

This appendix shows an adoption of the Tobacco Thesaurus for the Flamenco program. The thesaurus is contained at http://www.library.ucsf.edu/tobacco/thesaurus.html. It was developed by Americans for Nonsmokers Rights Foundation and used by the UCSF TCA in the indexing of the Joe Camel Campaign: Mangini vs. R. J. Reynolds Tobacco Company Collection and the British-American Tobacco Collection.

The thesaurus is a form of a controlled vocabulary and attempts to encompass all information about tobacco and tobacco control. It is broken up into broader terms, narrower terms and related terms, each of which relate one term to another in a controlled manner.

The controlled vocabulary is used for indexing and searching. Advantages of it include normalization of indexing concepts; identify index terms with a clear semantic meaning. Problems can occur when retrieving documents if people who are not familiar with the controlled vocabulary or can not find where a concept is located. Another problem may arise if important terms are missing from the thesaurus.

To adapt the tobacco thesaurus for the Flamenco program, I and my colleague Kirsten Swearingen, edited and created facets for the thesaurus. That is we strictly applied a hierarchical structure into a format that the Flamenco program could parse. We created a hierarchy with 834 terms based on these relations.

	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7
BEHAVIORS							
4	behavior						
5		individual behavior					
6			alcohol consumption				
7			brand loyalty				
8			brand switching				
9			drug use				
10				gateway theory			
159			faith influence				
11			health belief				
839			former smoker				
160			smoker				
161				heavy smoker			
173				light smoker			
189			non-smoker				
12			outdoor smoking				
450			patient				
451				psychiatric patient			
13			relapse				
14			risk taking behavior				
15			smoker satisfaction				
16			smoking attitude				
17			smoking cessation				
18				nicotine gum			
19				nicotine patch			
20				nicotine replacement therapy			
22					nicotine inhaler		
23					nicotine nasal spray		
25				smoking cessation method			
26					acupuncture		
27					cognitive behavioral therapy		
28					drug therapy		
29					health education program		
30						church based program	
31						community based program	
32						family based program	
33						school based program	
34						self help method	
35						smoking campaign	
36						tobacco education program	
37						workplace based program	
38					hypnosis		
39					public health program		
40						smoking prevention	
41					smoking intervention		
42					stress management		
43					treatment program		
44						telephone counseling	

	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
45						treatment outcome
46			smoking history			
47			smoking reduction			
48			tobacco abstinence			
49			tobacco use			
50				smoking initiation		
51				snuff dipping		
52				tobacco chewing		
78			tourism			
53			voting			
54		social behavior				
55			activist strategy			
56			advocacy			
57				ally		
58				anti-smoking advocacy		
59				media advocacy		
60					editorial	
61					letter to the editor	
158			ethics			
62			peer influence			
63			petition drive			
64			public awareness			
65			social cost			
66			social influence			
67				boycott		
68				civil disobedience		
69				corporate responsibility		
70				divestment of funds		
71				shareholder resolution		
72			societal attitude			
73			sociocultural norm			
74			underage smoking			
OCCUPATIONAL ACTIVITIES						
190	occupational activity					
191		criminal investigation				
192		economic activity				
193			fund raising			
194			funding			
195				funding alternative		
196				funding source		
197			sales pricing			
199		economics				
200			economic analysis			
201				economic forecast		
202				economic impact		
203				market trend		
204			economic cost			
205				absenteeism		

	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
673					political expenditure	
674						campaign contribution
675						soft money
227				health care cost		
228					hospital length of stay	
229					insurance	
230						health insurance
231						life insurance
208		educational activity				
209			public service announcement			
210			tobacco control program			
211				demand reduction		
212			training program			
154			educational material			
155				tobacco education material		
156					bilingual tobacco education material	
213		entrepreneur				
214		government activity				
215			city planning			
216			government sponsored conference			
217			governmental spending			
219			law enforcement			
220				sting operation		
221			tobacco subsidy			
222		health care activity				
223			diagnostic procedure			
224				autopsy		
225				respiratory function test		
226			health care			
232			health care provider influence			
233		lobbying				
271			grass roots lobbying			
272				letter writing campaigns		
273				phone banks		
234		industry activity				
235			advertising			
2				advertising activity		
3					issue advertising	
236				advertising campaign		
237				advertising effectiveness		
238				advertising expenditure		
239				advertising message		
241				bilingual advertising		
242				counter advertising		
243			brand awareness study			
			budget			
244			charitable donation			
245			cigarette design			

246 cigarette adhesive
247 cigarette paper
248 cigarette perforation
249 filter design
250 cigarette ventilation hole
251 corporate intelligence
252 corporate merger
253 diversion of funds
254 industry strategy
255 industry sponsored conference
256 economic development
257 event sponsorship
258 sports sponsorship
259 motor sport sponsorship
260 financial investment
261 health claim
262 industry funding of education
263 industry recommendation
264 industry response
265 industry sponsored prevention program
266 accommodation
268 industry terminology
269 international trade
274 preemption
275 product development
276 product liability
299 tobacco industry internal policy
300 tobacco industry policy
301 tobacco industry sponsorship
302 tobacco lobby
303 voluntary agreement
305 hearing
306 repeal
307 lawsuit
308 class action suit
309 cost-recovery lawsuit
310 individual lawsuit
311 international lawsuit
312 Medicaid lawsuit
313 passive smoking lawsuit
314 unfair business practice
135 court decision
840 Supreme Court decision
315 legal appeal
316 legal precedent
317 litigation
318 patent
852 trademark
319 settlement
320 dispute settlement

321 global settlement
322 master settlement agreement
323 monetary damage
324 settlement distribution
325 testimony
362 whistleblower
327 program evaluation
328 program planning
329 research activity
330 animal research
331 animal model
332 animal behavior
333 animal smoke inhalation
334 animal subject
91 animal
174 mammal
175 laboratory rat
92 biotechnology
335 chemosensory research
336 chromatography
337 cigarette analysis
338 combustion study
339 data analysis
340 statistical reliability
341 statistical validity
342 fractionation experiment
343 genetic engineering
344 human subject
345 industry sponsored research
346 inhalation study
348 mouse skin painting
349 pharmacology
350 toxicology
351 risk assessment
352 screening test
353 spectrometry
354 youth risk behavior surveillance
95 business activity
96 marketing
brand image
97 cost sharing
99 mail order sale
100 mall intercept
101 mall sample
102 market forecast
176 market segment
103 market segmentation
104 geographic market
105 target market
106 market testing

107 marketing research
108 marketing strategy
109 opinion poll
277 promotions
278 cigarette promotion code
point of purchase
279 promotional campaign
280 promotional merchandise
281 cigarette sample
282 public relations
283 sales
284 distribution
285 export
286 import
287 profit
288 revenue
289 market share
290 taxation
291
292 royalty
293 sales rate
294 tobacco sales
295 cigarette sale
296
297 self service displa
111 sports marketing
112 wholesale trade
705 pricing
706 price elasticity
114 business meeting
326 media advertising
118 advertising medium
851 tobacco art
119 billboard
180 mass media
181 media campaign
183 newspaper
184 newspaper advertising
185 radio
187 television
120 broadcast advertising
121 broadcast restriction
123 radio advertising
124 television advertising
125 celebrity endorsement
117 coupon
126 print advertising
127 product distribution
128 product placement
129 paid product placement

130 slotting fee
131 unpaid product placement
132 sign
134 tombstone advertising
707 production
355 tobacco processing
356 ammonia processing
357 freon
358 ozone treatment
359 tobacco aging
360 tobacco farming
647 plant
649 tobacco leaf
650 burley tobacco
651 oriental tobacco
652 tobacco leaf constituent
653 turkish tobacco
654 virginia tobacco
361 tobacco storage
708 manufactured product
709 death certificate
710 tobacco product
711 bidi
712 cigar
713 cigarette
714 discount brand
715 fire safe cigarette
716 generic brand
718 low yield cigarette
719 menthol cigarette
720 nicotine free cigarette
721 premium brand
722 roll your own
723 safer cigarette
724 single cigarette
725 slow glow cigarette
726 unfiltered cigarette
727 cigarette packaging
728 generic packaging
729 warning label
730 tar and nicotine la
731 cigarillo
732 pipe tobacco
733 smokeless tobacco
734 tobacco product accessory
735 actron filter
736 carbon filter
737 charcoal filter
738 dumbbell filter
739 nox filter

740 strickman filter
741 tobacco product attribute
742 cigarette additive
743 acetaldehyde
744 acrolein
745 ammonia
746
747
748 cadmium
749 camphene
750 chemosol
751 cocoa
752 coumarin
753 cytrel
754 eugenol
755 glycerin
756 glycerol
757 lead
758 menthol
759 nitrate
760 polonium
761 sugar
762 gas phase
763 liquid phase
764 particulate matter
765 particulate phase
766 smoke constituent
767 benzene
768 benzopyrene
769 carbon dioxide
770 carbon monoxide
771 chromium
772 formaldehyde
773 hydrogen cyanide
774 lead
775 nickel
776 nitric oxide
777 nitrosamine
778 palladium
779 phenol
780 tar
781
782 tobacco odor
783 vapor phase
570 vending machine
599 cigarette vending machine

PEOPLE

136 demographics

- 79 age group
  - 80 adult
  - 81 child
  - 82 elderly
  - 83 fetus
  - 84 infant
  - 85 middle aged adult
  - 86 preschool children
  - 677 adolescent female
  - 678 adolescent male
  - 87 young adult female smoker
  - 88 young adult male smoker
  - 89 young adult smoker
  - 90 youth
- 137 consumer market
- 138 disabled person
- 139 purchasing pattern
- 153 educational level
  - 145 educational group
  - 146 student
    - 147 college student
    - 148 elementary school student
    - 149 high school student
    - 150 junior high school student
    - 151 medical student
    - 152 middle school student
- 140 socioeconomic status
  - 141 lower class
  - 142 middle class
  - 143 upper class
- 676 population group
  - ethnicity/nationality
    - 679 African American
    - 680 Arab
    - 681 Asian
    - 683 Caucasian
    - 684 Chinese American
    - 691 Hispanic American
    - 692 Japanese American
    - 696 Mexican American
    - 697 Native American
    - 698 Puerto Rican
- 685 family
  - 686 family influence
  - 687 parent
- gender
  - 695 male

688 female
sexual orientation
682 bisexual
689 gay man
694 lesbian
690 heterosexual
693 religion
188 Jew
Muslim
699 sociographic segment
700 poverty
701 rural area
702 suburbanite
703 urbanite
704 wealth
545 trend
816 public health statistics
817 epidemiology
172 life expectancy
818 morbidity
819 mortality
820 infant mortality
821 sudden infant death syndrome
822 tobacco related death
823 quality of life
824 statistical prevalence
784 professional or occupational group
785 athlete
786 attorney general
787 blue collar worker
788 bus driver
789 college administrator
790 flight attendant
791 government employee
792 employee
793 Surgeon General
794 health care provider
795 doctor
796 nurse
797 hospitality industry employee
798 bartender
799 waiter
800 waitress
801 lawyer
802 plaintiff lawyer
803 tobacco industry lawyer
804 legislator
805 member of Congress
806 state representative
807 state senator

808 US senator
809 military personnel
810 pilot
811 restaurant worker
812 tobacco farmer
813 tobacco industry employee
814 tobacco industry scientist
815 white collar worker

ORGANIZATIONS

388 organization
389 advertising industry
390 advocacy group
391 health advocacy group
392 prevention task force
393 industry front group
394 smokers' rights group
395 business organization
396 advertising agency
397 agriculture organization
398 black market
399 gray market
400 entertainment organization
401 entertainment industry
402 Political Action Committee
403 tobacco organization
404 tobacco subsidiary
405 corporate structure
406 corporate officer
407 corporation
408 publicly held corporation
409 educational organization
410 government organization
411 Congress
412 Food and Drug Administration
413 government agency
116 county government
414 local government agency

415
416
417 health care related organization
418 health maintenance organization
419 health related organization
420 hospitality industry
421 legal system
168 legal concept
169 legal right

170

171

422 court
423 Supreme Court
424 criminal justice system
439 law firm
440 plaintiff law firm
441 tobacco industry law firm
442 nonprofit organization
443 pharmaceutical industry
444 philanthropic foundation
445 professional association
446 labor union
447 tobacco industry structure
448 tobacco manufacturer
449 trade association

SUBSTANCES

chemistry

157 element, ion, or isotope
162 inorganic chemical
163 asbestos
164 nitrogen
165 nitrogen compound
166 nitrogen dioxide
167 nitrogen oxide
363 organic chemical
364 alcohol
365 alicyclic hydrocarbon
366 alkaloid
367 carbohydrate
368 heterocyclic compound
369 hydrocarbon
370 alicyclic hydrocarbon
371 aliphatic hydrocarbon

372

373 lipid
374 naphthalene
375 nicotine
376 cotinine
377 nicotine level
378 nitroalkane
379 nucleic acid, nucleoside or nucleotide
380 organophosphorous
381 peptide
382 nicotine receptor
383 polycyclic aromatic hydrocarbons
384 polypropylene
385 solanesol
386 steroid
387 uric acid

452 pharmacological substance
453 appetite suppressant
454 bupropion hydrochloride
455 buspirone
456 gateway drug
457 narcotic
458 oral contraceptive
459 recreational drug
648 marijuana
460 sedative
461 stimulant
462 caffeine
655 poisonous substance
656 carcinogen
657 cocarcinogen
658 hydrogen cyanide
659 pesticide
547 physical object
548 anatomical structure
549 abnormality
550 anatomical abnormality
551 acquired abnormality
552 body part
553 larynx
554
555 body system
556 cardiovascular system
557 central nervous system
558 endocrine system
559 gastrointestinal system
560 musculoskeletal system
561 reproductive system
562 respiratory system
563 urogenital system
564 cell or cell component
565 gene or genome
566 organ or organ component
567 endothelium
568 epithelium
569 salivary gland
571 substance
572 biologically active substance
573 hormone
574 immunologic factor
575 receptor
576 vitamin
577 vitamin a
578 vitamin b
579 vitamin c
580 vitamin d

581 vitamin e
582 body substance
583 cotinine
584 enzyme
585 cigarette ingredient
586 expanded tobacco
587 freeze dried tobacco
588 reconstituted tobacco
589 tobacco substitute
590 food
591 alcoholic beverage
592 candy cigarette
593 tobacco smoke
594 bidi smoke
595 cigar smoke
596 cigarette smoke
597 mainstream smoke
598 pipe smoke

HEALTH EFFECTS

463 phenomenon
464 addiction
465 adverse effects
466 human caused phenomenon
467 environmental effect of humans
468 fire
469 indoor air quality
470
473 pollution
475 secondhand smoke
476 sick building syndrome
477 poisoning
478 natural phenomenon
479 biological function
480 body weight
481
482 breast feeding
483 metabolism
484
485 pregnancy
486
487
488 respiration
489 weight gain
490 weight loss
491 pathological function
492 allergy

493			
494			birth defect
495			disease
517			alcoholism
518			asbestosis
519			cancer
520			cardiovascular disease
521			central nervous system disease
522			
523			
524			
525			cerebrovascular disorder
526			chronic obstructive pulmonary disease
527			depression
528			Gastrointestinal disease
529			immune system disease
530			
531			impotence
532			mouth disease
533			musculoskeletal disease
534			Periodontal disease
535			respiratory disease
536			skin disease
537			thrush
538			tobacco amblyopia
539			ulcer
540			Urogenital disease
541			hyperplasia
542			low birth weight
841			symptom
543		poisoning	
544			nicotine poisoning
PLACES			
601	place		
602	academic institution		
603		college	
604		elementary school	
605		high school	
606		junior high school	
607		middle school	

608 library
609 public place
610 airport
611 bar
612 billiards room
613 bingo parlor
614 bowling alley
615 card room
616 casino
617 cruise ship
618 factory
619 building
620 health care facility
621 nursing home
622 hospital
623 hotel
624 museum
625 pool hall
626 prison
627 public transportation
628 airplane
629 bus
630 taxicab
631 train
632 recreational facility
633 restaurant
635 smoking section
636 sports event
637 stadium
638 train station
639 workplace
853 workplace liability
854 workplace productivity
640 residence
641 apartment
642 dormitory
643 retail outlet
634 shopping mall
644 convenience store
645 pharmacy
646 tobacco store

POLICIES / REGULATIONS

660 policy
661 economic policy
662 health policy
663 public health policy
664 public policy
665 employee rights
666 product restriction
667 public smoking law

668 social policy
669 tobacco policy
670 youth access
671 smoke free policy
672 smoking restriction
826 regulation
827 administrative regulation
828 advertising restriction
829 Americans with Disabilities Act
830 cigarette tax
832 Freedom of Information Act
825 public record
833 impact of regulations
834 legislation
835 antitrust legislation
836 legislative event
837 referendum
429 clean indoor air act
430 constitutional amendment
431 federal legislation
433 local ordinance
435 zoning
436 minimum purchase age
437 proposed legislation
438 state legislation
838 licensing
425 law
426 case law

Appendix 5.
Creating a text data-mining application for use in public health informatics

M.C. Michel, M.S.[1], L.A. Bero Ph.D.[2] T. Bright[3]
[1]Graduate Group in Biological and Medical Informatics, UCSF, San Francisco, CA, USA
[2]Department of Clinical Pharmacy and Institute for Health Policy Studies, UCSF, San Francisco, CA, USA
[3]Department of Information Systems, University of Maryland Baltimore County, Baltimore, MD, USA

Abstract— **Recent litigation and the Master Settlement Agreement of 1998 have made millions of tobacco industry internal documents available on the Internet (http://legacy.library.ucsf.edu). The Legacy interface, housed at the University of California, San Francisco, is based on a traditional information retrieval model in which documents are indexed and retrieved based on user-specified queries.**
One problem with the Legacy interface is information overload. In an attempt to ease this problem, we are developing a text-mining interface to enable exploratory analysis and discovery of information from collections of data. Users could uncover new patterns and concepts and thus text mining could result in searches that are targeted and specific, which would decrease information overload.
In order to determine information needs, nine in-depth interviews with regular users of the Legacy interface were conducted. Results show that participants identified clustering as a useful tool in identifying and extracting key concepts and identified the need to recognize relationships between terms and concepts within the data. We encourage researchers who are developing text-mining interfaces to survey the users to learn what particular aspects of their research could be enhanced by text mining.

Keywords—**Text data mining, public health, user-interface design**

I. Introduction

The internal tobacco industry documents represent a tremendous opportunity for public health. The documents have been arguably one of the most useful results of the Master Settlement Agreement in 1998 when the Attorneys' General of 48 states sued the tobacco industry [1]. The documents have already proven valuable for advancing public health objectives of reducing tobacco use and exposure[2, 3]. However, tobacco control researchers have done most document analyses. Developing new ways of searching could open up analysis to other research disciplines, teachers, and advocates.

Our objective is to develop new methods of searching the tobacco industry documents by using a combination of quantitative and qualitative studies to inform the development of informatics tools to discover new information about the tobacco industry. Using these methods will enable people to examine the industry in a different way and facilitate connections that were not easily made using a linear search engine.

There has been little research on text data mining itself and no research on text data mining using a large public health corpus of documents such as internal tobacco industry documents. The literature on various methodologies of data mining is well developed; however, text data mining is a nascent field[4]. There are a number of advantages to develop text data mining for the tobacco industry documents.

While the existing websites are useful search engines for the tobacco industry documents, they do not provide any analysis assistance for the user. Using the existing search engine, the user often gets too many search engine results or too few. In addition, the documents found in the traditional manner have no context. The objective of our research is to develop a user-derived interface for a text data-mining engine and test its usefulness for discovering new relations within the tobacco industry documents.

II. METHODOLOGY

Two methods were employed to abstract the potential text mining needs of users of the internal tobacco industry documents. First, an online survey of users was conducted on two websites (http://legacy.library.ucsf.edu/ and http://www.library.ucsf.edu/tobacco/batco/). The surveys were analyzed for themes.

We also conducted nine in-depth interviews of tobacco documents researchers. The interviews consisted of three components: a 20 minute

unstructured search of the Legacy website, four structured questions, and nine open-ended descriptive questions. The questions are listed below:
The questions for this part of the evaluation are as follows:

1. Find a document that contains "Ayres", who worked at British American Tobacco and see if you can find out what his title was.
2. Find a document that was published during 1985-1990 and discusses marketing to young adults.
3. Find a document that says the following "Nicotine is not addictive".
4 Find the document that says that Sylvester Stallone will accept $500,000 from Brown and Williamson to use their products in the movies.

Then we asked ten open-ended questions about tobacco documents searching:

1. What do you think are the strengths and weaknesses of your own search technique?
2. What is the most important feature of a search engine? (If necessary probe- usability, accuracy, flexibility, sorting capacity, bookmarks, consistency, reversal of actions, shortcuts...)
3. What manner do you use to organize documents that you retrieve; by subject area, date, organization name, theme, etc.?
4. How does the UCSF Legacy collection meet your searching needs?
5. What would you modify to make the searching process more applicable to your specific needs?
6. How long have you been using the UCSF Legacy online collection to perform searches?
7. What additional resources do you use for searching and why?
8. When you do not know the focus of your research, how do you begin to drill down and target specific information?
9. What difficulties using the UCSF Legacy collection?

The interviews were 1 hour in length.

The users were divided into experienced searchers (searched Legacy for 6 months or greater) and inexperienced searchers (searched for less than 6 months).

From these two sources, we abstracted themes, which were sorted into suggestions that could be addressed by informatics tools or other suggestions about the interface

III. Results

We found that text mining could be useful for researchers of the tobacco documents in a number of areas. First researchers could use classification of the documents to gather data into one of several discrete concepts. For example, the term "young adults" could be sub-classified into documents pertaining to marketing, advertising or health effects. We also found that the users would benefit from the suggestions of search terms; for example: young adults AND marketing: –Y.A. AND marketing - Young Adult* AND marketing

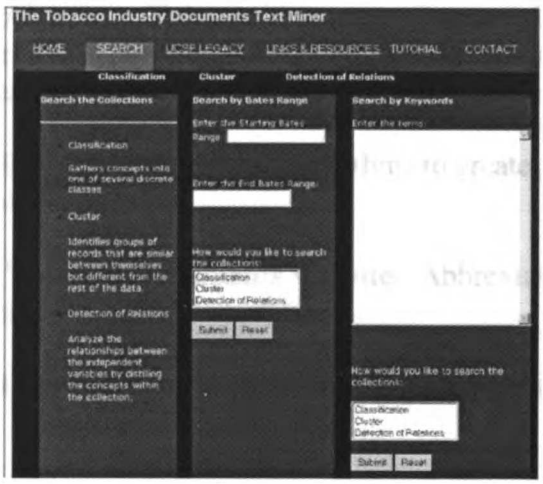
An additional benefit would be detection of relations, which would enable the user to analyze the relationships between independent variables by distilling the concepts within the collection.

Other benefits include factors that are unique to documents research. Often, researchers will find a particular document of interest, but are unable to recall the search process of how the document was found.

Another need identified from interviewing the researchers were the ability to track the stages of revisions in an internal tobacco document and identify the changes made to the document over time. Following the path of a document from conception to dissemination would be useful for a researcher. The tobacco industry lawyers often edited scientific documents for content[5, 6]. The researchers would like to know who edited the

document and when.

Figure 1: A screenshot from text data miner interface



Finally the clustering of documents based on automatic categorization assignments would be useful for both experienced and novice users. The novice users could use clustering to get an overview of a particular area of interest. Experienced users could use clustering to help them draw connections that they previously did not think of or use it to narrow down a search with many hits.

IV. Conclusion

Rather than creating a product and presenting it to the users, we asked the users what they currently do and adapted the technology to their current workflow.

When applied research groups ask their users what products they want, they are more likely to find that the needs can be met with less resources than originally thought. Creative solutions are necessary in public health during these times of budgetary crisis. Text mining has the potential to be useful in many different public health settings from outbreak detection to breast cancer prevention. However, we need to go to the users in the state and local health departments to find out how they use data and what text-mining algorithms suit their needs.

Acknowledgment

We thank the researchers in the Center for Tobacco Control Research and Education who participated in the interviews. We also thank Kirsten Neilsen, the tobacco documents librarian at UCSF and Annamaria Baba for their helpful comments. This work was supported by Tobacco Related Diseases Research # 12DT-0186.

References

[1] "The State of Minnesota and Blue Cross/Blue Shield of Minnesota v. Philip Morris Inc., *et al.*," vol. 2002, 1998.

[2] R. E. Malone and E. D. Balbach, "Tobacco industry documents: treasure trove or quagmire?," *Tobacco Control*, vol. 9, pp. 334-8., 2000.

[3] L. Bero, "Implications of the tobacco industry documents for public health and policy," *Annu Rev Public Health*, vol. 24, pp. 267-88, 2003.

[4] M. Hearst, "Untangling Text Data Mining," presented at Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, 1999.

[5] L. Bero, D. E. Barnes, P. Hanauer, J. Slade, and S. A. Glantz, "Lawyer Control of the Tobacco Industrys External Research-Program - the Brown-and-Williamson Documents," *Jama-Journal of the American Medical Association*, vol. 274, pp. 241-247, 1995.

[6] P. Hanauer, J. Slade, D. E. Barnes, L. Bero, and S. A. Glantz, "Lawyer control of internal scientific research to protect against products liability lawsuits. The Brown and Williamson documents," *Jama*, vol. 274, pp. 234-40, 1995.

Appendix 6.
Glossary and Abbreviations

American Legacy Foundation: Abbreviated as ALF

British American Tobacco Company Documents: Abbreviated as BATco. These documents are approximately 17,500 documents which were obtained from the British American Tobacco company and were manually categorized and indexed.

Clustering: Using statistical algorithms to create groups that are alike and distinct from other groups.

Legacy Tobacco Documents website: Abbreviated as LTDL or Legacy and funded by the American Legacy Foundation

Medical informatics: A multi-disciplinary field that studies how medical related data is utilized, stored and retrieved.

National Association of Attorneys General: Abbreviated as NAAG

National Cancer Institute: Abbreviated as NCI

Phillip Morris: Abbreviated as PM- also knows as Altria

People-centered public health informatics: A multi-disciplinary field that uses qualitative and quantitative methods to determine the needs of the population and implements those needs through informatics related solutions.

Public health informatics: A multi-disciplinary field that studies how public health related data is utilized, stored and retrieved.

R.J. Reynolds – Abbreviated as RJR – the tobacco company.

Support vector machines -

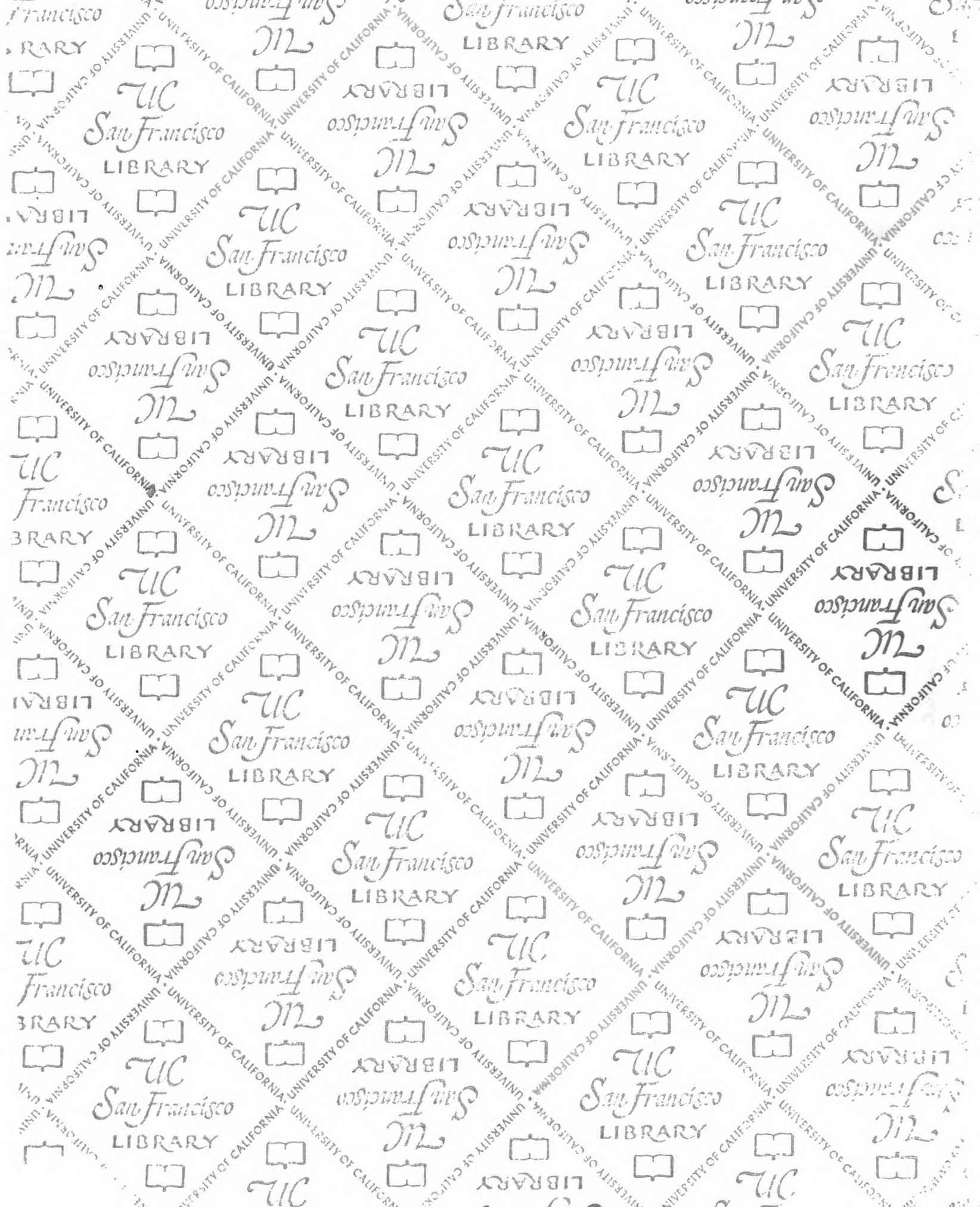
Text data mining (also text mining) – Using statistical algorithms on text to derive unknown information.

Tobacco Control Archives: Abbreviated as TCA

Tobacco Depositions and Trial Testimony Archive: Abbreviated as Tobacco DATTA

Tobacco Documents Online: Abbreviated as TDO

Tobacco Related Disease Research Program: Abbreviated as TRDRP


UC
San Francisco
LIBRARY
UNIVERSITY OF CALIFORNIA · UNIVERSITY OF CALIFORNIA

For reference

Not to be taken from the room.