A Knowledge-Network Model Of Scientific Communities

by

Jose Maria Gonzalez Pinto

Ingeniero en Sistemas Computacionales, Instituto Tecnologico de Merida. Merida, Yucatan, Mexico 1999

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 2002

Author .....
Program in Media Arts and Sciences
May 13, 2002

Certified by .....
Walter Bender
Senior Research Scientist, MIT Program in Media Arts and Sciences
Thesis Supervisor

Accepted by .....
Andrew Lippman
Chair, Departmental Committee on Graduate Studies
Program in Media Arts and Sciences

A Knowledge-Network Model of Scientific Communities

by

Jose Maria Gonzalez Pinto

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on May 13, 2002. In partial fulfillment of the requirements for the degree of Master of Science in Media Arts and Sciences

Abstract

The amount of information organizations possess now days is overwhelming and the need of being capable of extracting valuable knowledge from such large amount of information is imperative. This thesis presents a software tool capable of extracting valuable knowledge (e.g. expertise) of a scientific community, generating relationships among community members automatically and revealing these relationships through a visualization tool. The types of relationships that this tool reveals are of the form of "who knows what" and "who can collaborate with whom" (both based on areas of expertise). The work presented was conducted and evaluated within the context of research institutions.

Thesis Supervisor: Walter Bender
Title: Senior Research Scientist, MIT Program in Media Arts and Sciences

A Knowledge-Network Model Of Scientific Communities

by

Jose Maria Gonzalez Pinto

The following people served as readers for this thesis:

Reader.....

Brian K. Smith
Assistant Professor
MIT Media Laboratory

Reader.....

Steven Whittaker
Research Visitor
British Telecommunications (BT)

Acknowledgements

To my family, the most important part of my life...the best source of my inspiration and desires of being better in each aspect of my life... to my mother for being an excellent outstanding example of perseverance and optimism... to my brothers Pedro and Jaime for being the best friends I could ever have, for helping me maintain a good sense of humor and for their unconditional support....thanks!!!...to my uncle Javier Omar for being such a wonderful example, for his support and advice, i cannot find the words to express my gratitude for all the things you have done for me.....thank you!!!!!!

To my advisor Walter for helping me re-think about my thinking....for pushing always to make me go beyond of what i thought it was good enough... thanks coach!!!.... to my readers Brian and Steven: your critiques and comments were really helpful in this work, thanks for your time!!!

To my community of colleagues and friends: our conversations and discussions helped me a lot during this two wonderful years: Selene, Juan, Alex, Ramesh, Cameron, Sunil, Vadim, Marcos, Jorges, Ernestititin, Elvis, Pedro, Jennifer, Erika, Anna, Claudia, Catherine, Cindy and the list continues..... it has been a truly amazing experience...thanks to all of you.....I will miss you guys!!!! and i guess i will see you when i see you...perhaps in another dimension.... we'll see....

Special thanks to the Telmex Fellows Program, without its support this work could have never been done.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The problem

A growing concern for organizations and groups has been to extract valuable knowledge (e.g. areas of expertise) from the large amount of accessible textual data they possess and use such data to facilitate collaboration and group formation among its members. Organizations are aware that collaboration between its members is the only way to achieve tasks larger than any one person alone could accomplish. Furthermore, due to increase in available information, as pointed out by Yimam, "organizations are giving more emphasis to capitalization of their knowledge. To this end, the utilization of various automated techniques for capitalization on the increasing mass of digitized knowledge that organizations generate in the conduct of their business is receiving a great deal of attention" [Yimam 1999]. "The real value of information systems is in connecting people to people, so they can share what expertise and knowledge they have at the moment, given that the cutting edge is always changing. Therefore, it remains evident that if technology is to foster the effective utilization of knowledge in organizations, it has to be able to support not only access to documented knowledge but, most importantly, knowledge held by individuals" [Stewart, 1997].

This thesis proposes the development of a software tool that will enable members of the community to reinforce collaboration by providing them with the information required to learn about other members areas of expertise (e.g. Artificial Intelligence, Knowledge Representation, Physics, etc.). Moreover, our software tool will draw relationships among them automatically by providing a visualization

tool capable of capturing the "sense and substance" of the data [Tufte 1990]. It is true that for an organization is important to manage their knowledge and expertise. Furthermore, it is through a better understanding and management of an organization's intellectual resources that enables the organization to prosper. As organizations grow in size, geographical scope and complexity, the need of solving more involvement tasks and cross-collaboration among organizations increases; in order to face this need, it has become important to be able to put together the right people with the right people. Our tool will be capable of contrasting two organizations in terms of their areas of expertise among their members. The idea is to provide an answer to the question outline above: given two scientific organizations, who are the people that should get in contact to collaborate in a project? Our goal is not to express in which areas a given scientific institution is weak, but to provide a software tool where they can look at and identify those areas where two organization can be benefit if they would like to work together.

1.2 Overview of this thesis

Chapter 2, "Example", presents three different scenarios where the software tool proposed in this thesis could be used.

Chapter 3, "Theory", outlines the theoretical foundations for this research. In particular we present some work done in the area of knowledge management and we introduced the theoretical issues which guided us in the design of the visualization module of the tool.

Chapter 4, "Design and implementation", describes the design and development of the two main parts of our software tool: the text-mining and the visualization.

Chapter 5, "Evaluation", presents the evaluation of the tool developed. We

describe the results obtained from tests where users were able to evaluate our software tool within the context of the scenarios they were presented.

Chapter 6, "Conclusion", describes the contributions and conclusions of this research as well as future work.

Chapter 2

Example

In this chapter we present an example of use of the software tool developed. Let us start by remembering how this idea came in the first place. At the Media Lab, the interaction with sponsors help students and faculty show their work and engage in useful conversations. These interactions occur at least twice a year during the "Sponsors Meeting Week". When I arrived at the Lab, two years ago, I still remember my first interaction with one sponsor. After talking for a while, he just asked me if I knew about people working in Wireless Technology in the Lab. At that point, I did not know about that. And my first thought was: what if a software tool can actually provide such information? What if we can do better than the typical sources to find that kind of information such as the Media Lab public web site? What if we could use the information already available and distill it to provide a useful tool for this kind of need? So, let us outline some details of what are the possible scenarios of use of the software tool proposed in three different scenarios: first, as a provider of the information necessary to get a better understanding of what a given research institution is doing, second, as a facilitator of team formation, and third, contrasting two scientific institutions.

2.1 Someone unfamiliar with a scientific community

The idea expressed by this example is to try to answer the question that many people ask themselves when they are introduced for the first time to a laboratory, such as the Media Lab. Presumably, one would like to get an understanding of what is going on in the laboratory, what are the projects currently being done, what is the laboratory about in terms of the different research areas, who is work-

ing in what and so forth. Furthermore, if the person interested in understanding these issues has particular interests in some research areas, to whom should he/she refer to talk about x, y and z areas? Our tool attempts to provide answers to those questions. The user will be able to see the research areas where the laboratory is doing some work. And from there, a user will be able to have access to those people's names working in those areas. Let us think about another possible scenario, where our tool is used as an augmenting feature to what a user reads online: Sara is reading on the web a research paper published in an important journal. The author of the paper is a member of a given scientific institution, let us pretend that he works in the Media Laboratory. Sara would like to know more about the research of the author. She probably will go the Media Lab web site and look for the personal web page of the author. And from there she may be able to have access to other papers. However, those papers may or may not have relation to the one she was reading. With our tool, she would go beyond that. She would be able to see relationships between the author of the paper and the rest of the community in terms of research areas they have in common. This experience would give her immediate access to those areas that she is interested plus the fact that she would know about some other researchers working in those areas.

2.2 Team formation and collaboration

This is another valuable use of our tool. Imagine a person within the scientific institution who would like to start a collaborative project. The question is: who are the people that, according to the research they have done, would be ideal for the project? The software tool proposed here, will facilitate this issue of team formation. A user will be able to navigate through our visual representation and find those who are the experts in the areas that the project will be dealing with.

2.3 Contrasting two scientific institutions

As collaboration facilitates the solution of complex problems, it is an important issue to put the "right people" of a given laboratory with the "right people" of another. By this we mean that our tool will illustrate a given user with a comparison in terms of common areas of research between two scientific organizations. Let us think again about Sara to get a better idea of another way of seeing this issue of contrasting two scientific institutions and why is this important. Sara has identified a researcher who works at the Media Laboratory doing something that she is very interested in. Moreover, she was able by using our tool, of finding some other people working on similar areas, so now she has good resources of information of one of her areas of interests. What if, by using our tool, she could contrast the Media Laboratory with another scientific institution to find out more information about the topic she likes? What if she can get access to this information by realizing of these similarities in research areas through a visual representation? Definitely, she will be in a position of being more productive by accessing relevant information without the need of dealing with the classical problem of search engines: thousands of results from a typical query and some of them with not real relevance to what the user is looking for.

Chapter 3

Theory

This thesis depicts a software tool, which aims to provide a medium to facilitate and motivate collaboration and team formation in a scientific community, as well as contrasting two of these organizations in terms of their areas of research interests. In this chapter, the theoretical foundations that motivated the development of this application is presented. There are two main parts: first is the issue of knowledge management and second, the visualization.

3.1 Knowledge Management

It has been argued that knowledge (and expertise) is created, used and disseminated in ways that are inseparable from social factors [Erickson and Kellog]. Erickson continues his argument by pointing out that Knowledge management is not only an information problem but also a social problem. It is worthy here to cite an example to clarify why knowledge management is not only about information:

> "One of us once interviewed accountants at a large accounting and consulting firm about their information usage practices. The goal was to find out how they thought they would use a proposed database of their company's internal documents. In the course of the investigation, an unexpected theme emerged: the accountants said that one of the ways in which they wanted to use the documents was as a means of locating people. The accountants' claim —that they wanted to use a document retrieval system to find people— was, at the time, quite surprising. However, in the course of further interviews, it came to make sense: It was only through the people that the accountants could get some of the knowledge they needed. As one accountant explained, 'Well, if I'm putting together a proposal for Exxon, I really want to talk to people who've already worked with them: they'll know the politics and the history, and they can introduce me to their contacts. None of that gets into reports!'"

We can extract, from the five points that Erickson considered as important, two that applied to our research: first, some type of knowledge tend not to get written down: comments, opinions, or conjectures. Our software tool does not attempt to substitute human-to-human interaction. It claims to facilitate those who should get in contact due to the overlap in research areas. But of course, it is only through real interaction that it will be possible to know if the recommendations of the software tool are correct. Second, the importance of getting access to social resources such as contacts and referrals. Our software tool will be able to discover and represent those social resources within the context of a given scientific institution. We are not claiming to build a community but to discover those relationships who arise from the similarity of research interests.

As an example of the relevance of having a system capable of presenting the relationships between the members of a community in terms of areas of expertise we cite Harold "Doc" Edgerton, the inventor of the strobe light and one of the century's most prominent engineers, who once explained, when he wanted to find something out, first he would ask around to see whether anybody knew the answer, then he would try it out in the lab himself, and only then would he try looking the information up in a book or library (Edgerton, personal communication, 1989).

Some work related to one of the issues addressed here ("who knows what" in an organization setting) is ReferralWeb by Kautz and his group. They contend that the best way of finding an expert is through what is called "referral chaining" whereby a seeker find the needed expert through referral by colleagues (1996). They used the co-occurrence of names in close proximity in any documents publicly available on the World Wide Web as evidence of direct relationship. In partic-

ular they used: links found on home pages, lists of co-authors in technical papers and citation of papers, exchanges between individuals recorded in news archives and organization charts. The work proposed here differs from ReferralWeb in that we are not only using the notion of co-occurrence of names but also using a more powerful mining process as part of the analysis. Second, our work does not depend of a any search engine as ReferralWeb and since we are more interested in the characterization of a scientific community, the web is not the only resource that we are going to be using. Furthermore, the notion of social radius is not used in this thesis neither the typical social network visualization of connected nodes.

Other system that addresses the "who knows what" task with referrals is *ContactFinder*, proposed by Krulwich and Burkey. *ContactFinder* is an agent that reads messages posted on bulletin boards, extracting topic areas using heuristics. ContactFinder post a referral to a person when it encounters a question that matches that person's previous postings. There are two reasons why our tool is not considering this as a useful input: one is that posting messages are not always a good indicator of expertise, but rather of interest, and two, the fact the messages posting as well as emails are subject of serious privacy issues.

Another relevant work in the area of expert finding is *Answer Garden* [Ackerman and McDonald,1996]. This system is basically a question-answering and routing system that answer questions for technical help by retrieving stored question-answer pairs, but also provide facilities to route un-answered questions to a defined group of experts.

Vivacqua [Vivacqua, 1999] describes an expert finder agent in a spirit of ideas expressed by Kautz in that it suggests a personal agent that both profiles ones expertise and seeks for another expert when help is needed. Their work differs

ours in the fact that their work aimed to build communities of experts exchanging information taking as source of information Java source code. They depend on having users (not necessarily from the same organization) willing to share their profiles. Our work is not taking source code as a resource (we take publicly available information) and it focus in an already existing community.

Another work relevant in our context is Yenta [Foner 1997]. Yenta is a decentralized matchmaking system designed to find people with similar interests and introduce them to each other. In its core is a multi-agent strategy and a completely decentralized, peer-to-peer architecture. Yenta uses any kind of text: electronic mail messages, the contents of various newsgroup articles, the contents of the user's files in a filesystem, etc. In order to establish similarity, keyword vectors were used. Our work differs from Yenta, mainly because we are using a more powerful text-processing algorithm to perform clustering (see Chapter 4). Second, Yenta aims to perform matchmaking in the form individual-to-individual, we provide a mechanism to perform at the organization-to-organization level. Third, Yenta uses content that does not provide valid clues of expertise (e.g. emails, newsgroup articles). Electronic messages and newsgroup articles are very noisy in the sense that they tend not to be evaluated by an authority capable of validating whatever is being asserted. Our approach is arguably more reliable since we are using research papers, which are a more formal and valid way of expressing expertise.

3.2 Visualization

In some ways, as pointed out by [Ware 1999], a visualization can be considered an internal interface in a problem-solving system that has both human and computer components. A visualization can be the interface to a complex computer-based information system that supports data gathering and data analsys. On the human side, the visualization can act as an extension of cognitive processes, augmenting working memory by providing visual markers for concepts and by revealing structural relationships between problem components. Visualization is therefore, the process of transforming data, information, and knowledge into visual form making use of human's natural visual capabilities.

3.2.1 Interactive Visualization

Interactive visualization is a process made up of a number of interlocking feedback loops that fall into three broad classes. At the lowest level is the data manipulation loop, through which objects are selected and moved using the basic skills of eye-hand coordination. Delays of even a fraction of a second in this interaction cycle can seriously disrupt the performance of higher-level tasks. At an intermediate level is an exploration and navigation loop, through which an analyst finds his or her way in a large visual data space. As people explore a new town, they build a cognitive spatial model using key landmarks and paths between them, and something similar occurs when they explore data spaces. But exploration can be generalized to more abstract searching operations. Kirsh and Maglio (1994) define a class of epistemic actions as activities whereby someone hopes to better understand or perceive a problem. At the highest level is a problem-solving loop through which the analyst forms hypotheses about the data and refines them through an augmented visualization process. The process may be repeated

through multiple visualization cycles as new data is added, the problem is reformulated, possible solutions are identified, and the visualization is revised or replaced. [Ware 1999]. Following [Gershon et al.], it is true that with effective visual interfaces we can interact with large volumes of data rapidly and effectively to discover hidden characteristics, pattern, and trends. In our increasingly information-rich society, research and development in visualization has fundamentally changed the way we present and understand large complex data sets. As an example, the data landscape idea which has been applied to data related to the terrestrial environment, has also been applied to abstract data spaces such as the World Wide Web [Bray, 1996].

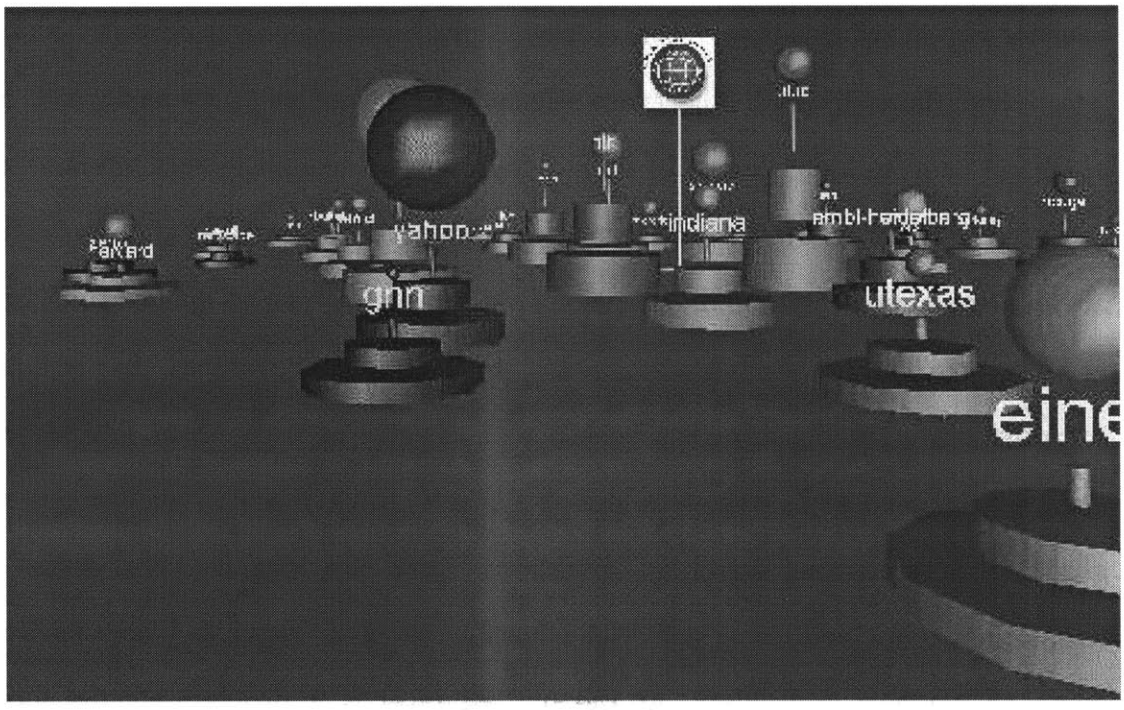

Figure 3.1: Data landscape applied to the World Wide Web

Statistical analysis of Internet hyperlinks formed the dataset underlying the structure of Tim Bray's map of the Internet as it existed in 1995. These flat images

were captured from a 3-dimensional mapscape. Each totem pole like structure represents a web site. The size of the basal disks graphically represents the quantity of content on the site, its height denotes the site's visibility on the net measured by a count of hyperlinks coming into the site from other sites and the spherical headpiece represents the site's hyperlinks out to other sites —a quality Bray calls luminosity[ed.]. Another work that uses the idea of a landscapes is the one developed by David Small [Small 1996]. In his work, he developed an application to visualize the complete plays of William Shakespeare (about one million words). He used a 3D space to organize complex relationships among different information elements. His approach for visualizing such amount of data includes a rendering model that was optimized for rapid navigation and changes in scale. He used a technique called *greeking* to maintain the overall shape of each line of the text presented as well as visual filtering to highlight the text the user is focus. See Figure 3.2.

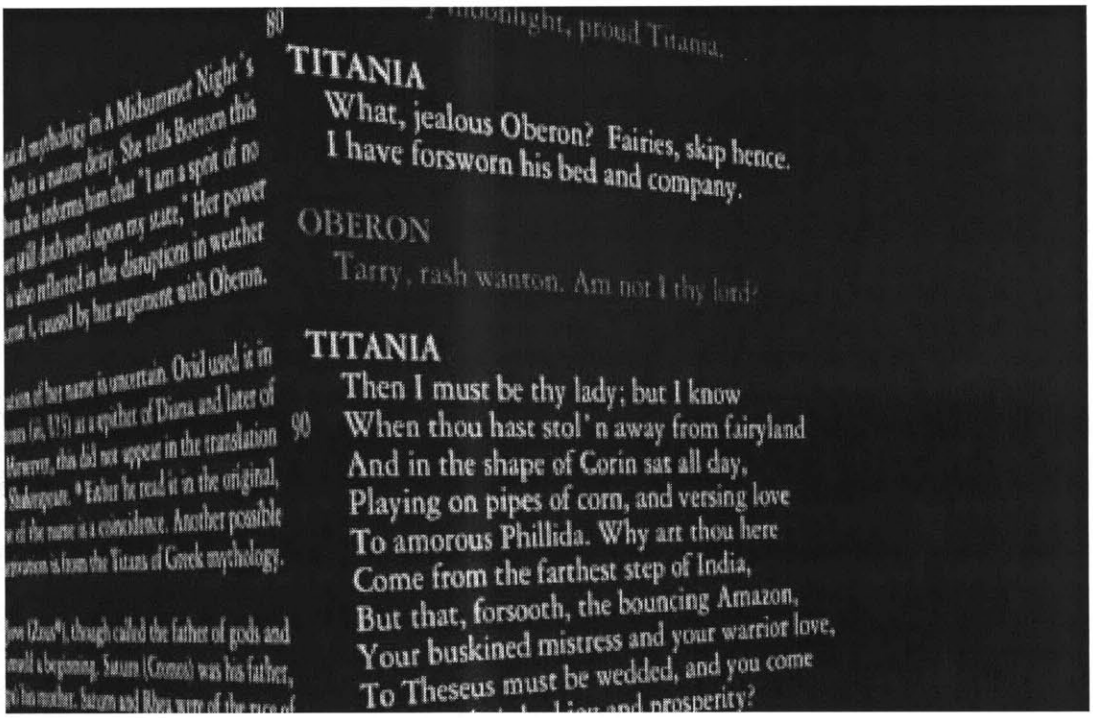

Figure 3.2: Virtual Shakespeare Project

3.2.2 Relationships

Many types of information involve relationships. One common way to visualize structures relationships is to use a graph, with nodes representing the entities and links the relationships between the entities.

Graphs work well for small information sets (tens to hundreds of nodes and links) but are easily cluttered and become visually confusing for larger sets. One promising approach for increasing the density of information on graphs involves using distortion lenses, enabling the viewer to see the detail and the general context. See Carpendale et al. for more details on this issue as well as a new contribution that extends distortion techniques from two to three dimensions.

One method for guiding the user's attention is by blurring the less relevant parts of the display while sharply displaying the relevant information. Using that same idea to blur objects based not on their distance from the camera but on their cur-

rent relevance in the application makes it possible to direct the viewer's attention. Kosar et al. call this method semantic depth of field (SDOF). SDOF allows users literally focus on the currently relevant information. Thus, it's possible to display the results of queries in their context and make them easier and faster to comprehend. Blur uses a visual feature that's inherent in the human eye and therefore is perceptually effective [Kosar et al.]. As an example, they developed LesDOF, an application that supports text display and keyword search. See Figures 3.3 and 3.4 taking from [Kosar et al.] to get a better idea of how this system looks:

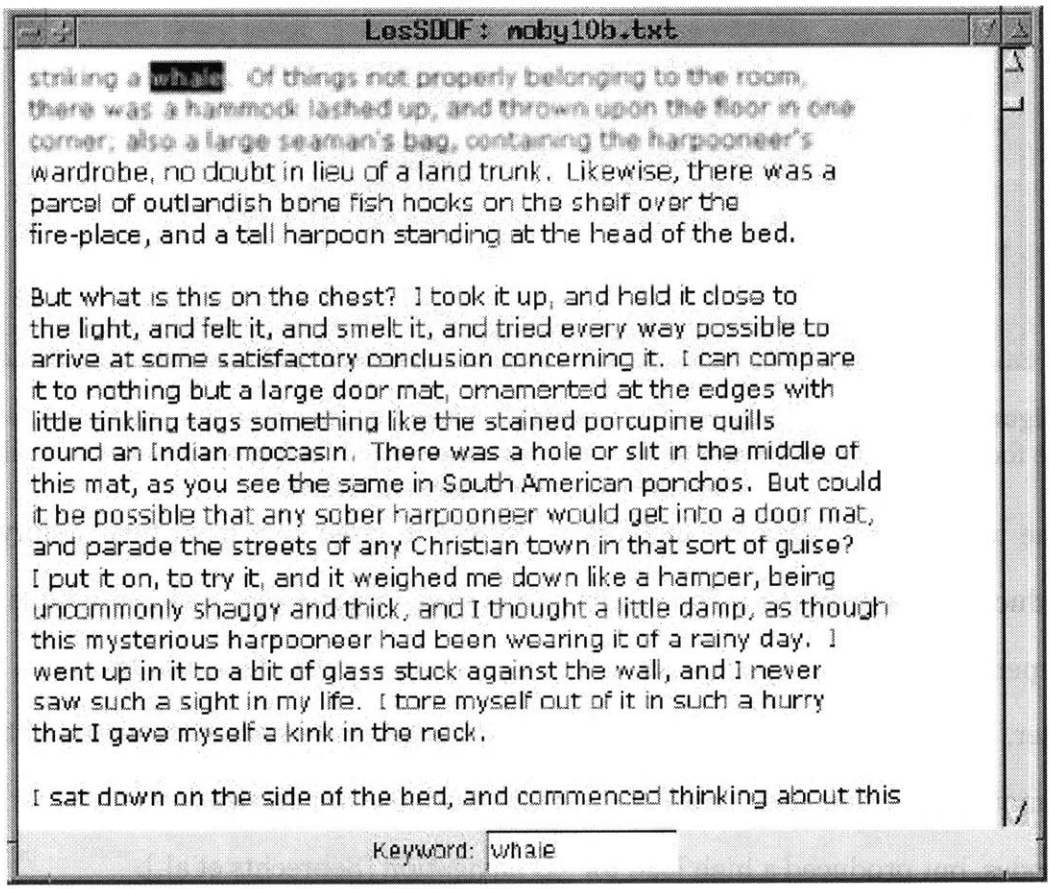

Figure 3.3: Scrolling in LessDOF. Three lines on the top are context from the last page, and therefore blurred, but still readable.

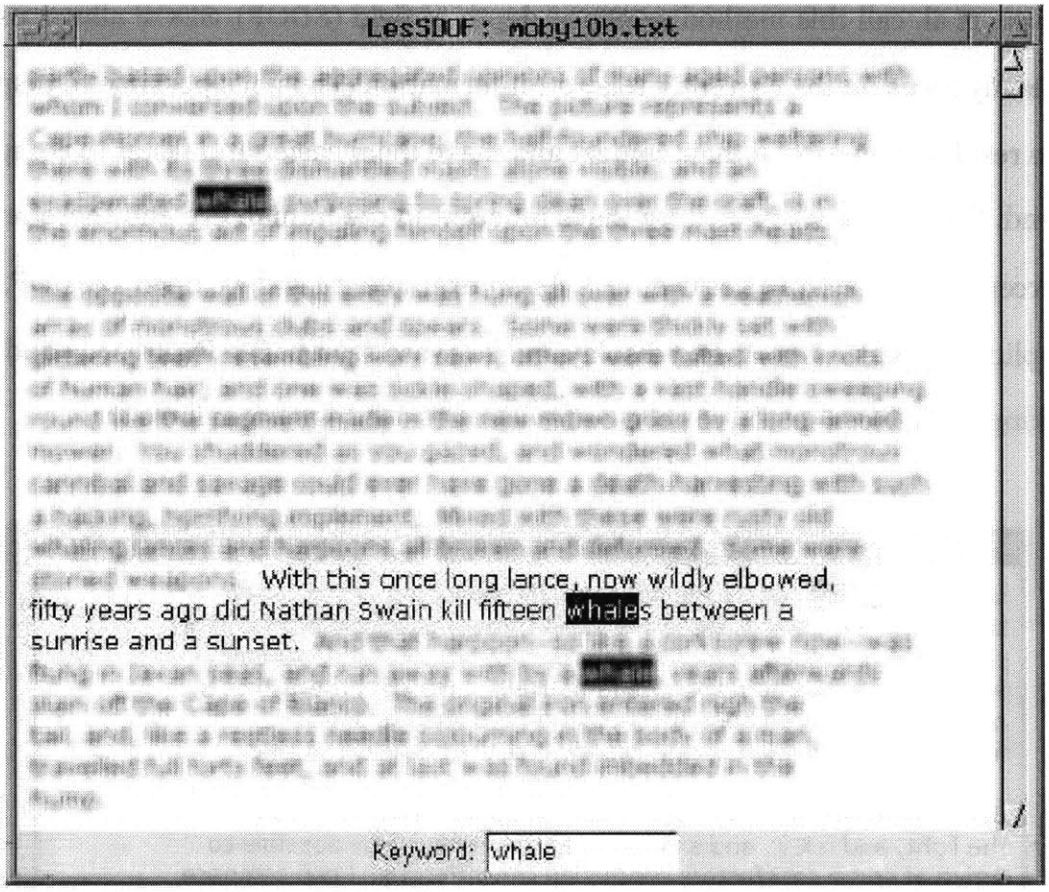

Figure 3.4: Finding a keyword in LessDOF. Three hits appear on this screenshot, with the focus currently on the middle one. The sentence around the keyword is clearly visible, while the rest of the context is blurred.

The utility of visualization techniques derives in large part from their ability to reduce mental workload. The results of this study suggest that such reductions are dependent upon an appropriate mapping among the interface, the task, and the user. In the case of the interface, for example, pilot studies had demonstrated that a mouse provided reasonably low mental workload for navigation in text or 2D modes, but produced a high load for 3D navigation [Sebrechts et al.].

3.2.3 Hierarchies

The traditional way of depicting hierarchical information is to structure it in a tree-like node and link diagram. For large trees, however, these diagrams rapidly become cluttered and unusable. One of the earliest instances of information visualization —the cone tree, developed by Card et al.— is a 3D representation of hierarchical information. The idea is to display information in a 3D representation. Information is broken down in a hierarchical structure. The highest node represents the generalized concept. The system can then be explored vertically, each level contains more details of elements belonging to the premier node. This process is then duplicated at each secondary node delivering many different levels until all the information has been represented. The use of shadow and transparency offers a view of the whole structure to the user. By incorporating rotational methods, the user can bring to the fore information nodes which exist behind or in the background of the display. The implementation of a fisheye view then allows the user to extract and explore certain elements of data within the context of a larger structure. See Figure 3.5.

3.2.4 Ease of use

In contrast to scientific visualizations, which focuses on highly trained scientists, interfaces created for manipulating information may be broadly deployed among a diverse and potentially nontechnical community. The demand for good and effective visualization of information embraces all walks of life and interests. This user community is diverse, based on varying levels of education, backgrounds, capabilities, and needs. We need to enable this diverse group to use visual representations that will be tailored to their specific needs and the specific problem at hand.

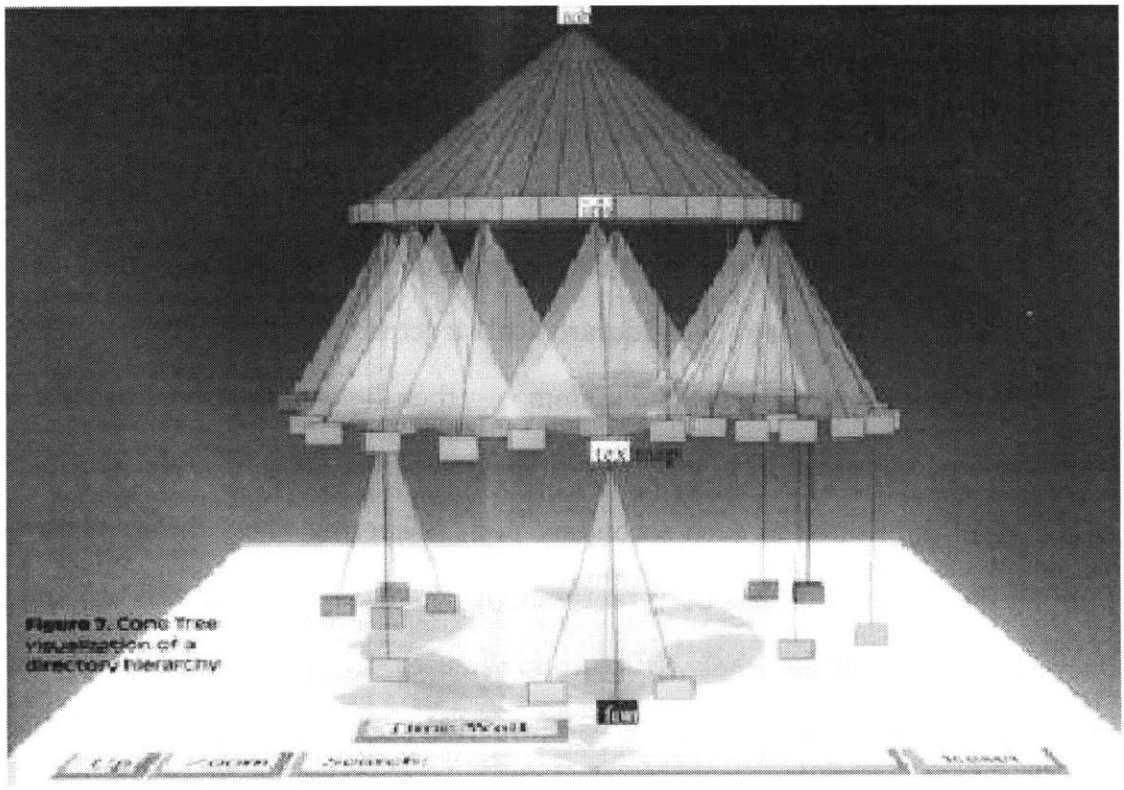

Figure 3.5: Cone tree visualization

3.2.5 Text vs. images

Recent developments in visual display computer hardware on the one hand and computer graphics and visualization methods and software on the other have generated new interest in images and visual representations. "A picture is worth a thousand words," goes the popular saying. However, images may have some disadvantages, and words are sometimes more effective (or powerful) than pictures. The disadvantages of using images include difficulty in representing information clearly; dependency of visual and information perception on memories, experiences, beliefs, and culture; and difficulty in making effective use of color. To use images effectively in science, education, art, and life in general, we need to understand the power and frailty of images. We need to understand when they are

equivalent to words, when they are more appropriate to represent information than words, and when they are not. This issue has become extremely important with the spread of the Web, whose many document authors use graphics inappropriately [Gershon et al.].

Another example of work related to the problem of browsing a collection of interrelated documents on the World Wide Web is *Dexter* by Murtaugh [Murtaugh 1996]. *Dexter* represents an alternative model to HTML for creating a browseable collection of keyword-annotated documents. *Dexter* presents viewers with a dynamic graphical interface that supports browsing based on association. Dexter was used in the award-winning Web-based HyperPortrait *Jerome B. Wiesner: A Random Walk through the Twentieth Century*. Dexter has two main parts: the Concept Map and the Material Listing. The Concept Map gives a graphical overview of the full set of keywords used to describe the set of documents. The Materials Listing provides an overview of the complete set of documents available to the viewer. See Figure 3.6.

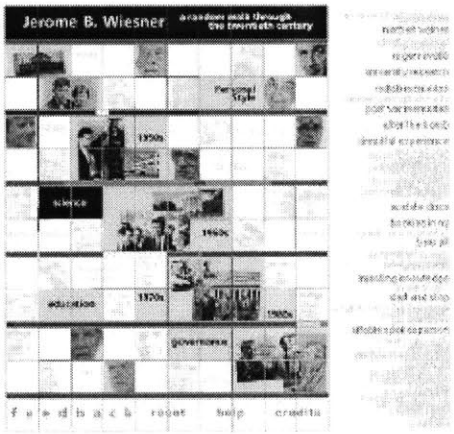

Figure 3.6: Jerome B. Wiesner: A Random Walk through the Twentieth Century.

Chapter 4

Design and Implementation

Here, we present the design details of our software tool. There are two main parts: one is what we called the *text-mining* and, second the *visualization* part.

4.1 Text-mining

We start by introducing our approach for the processing of the information. There are three pieces in the text processing: first, the profile extractor, second, the cluster generator and third, the community profiler. Figure 4.1 shows the text-mining task in a nutshell.

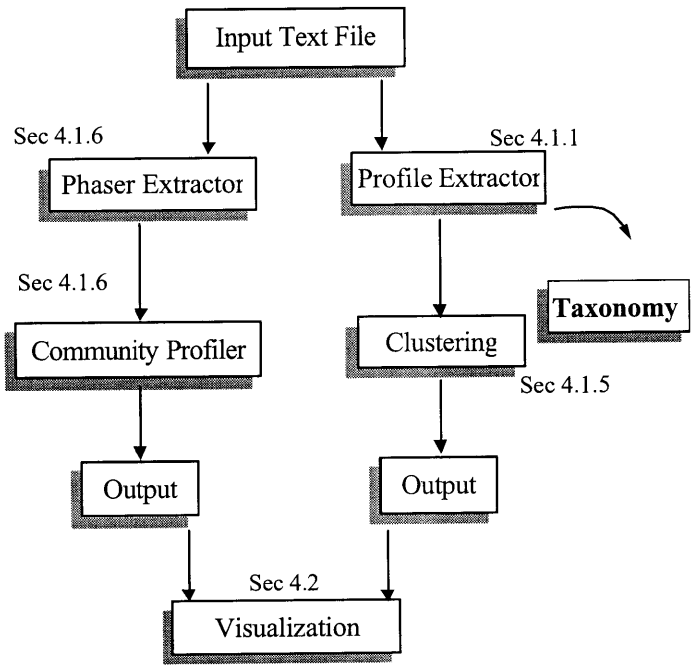

Figure 4.1: General Architecture

4.1.1 The Profile Extractor

Information Extraction (IE) systems try to identify and extract specific types of information from natural language text. Most IE systems focus on information that is relevant to a particular domain or topic. For example, IE systems have been built to extract the names of perpetrators and victims of terrorist incidents, and the names of people and companies involved in corporate acquisitions. IE systems have also been developed to extract information about joint venture activities [MUC1-5 Proceedings, 1993], microelectronics [MUC-5 Proceedings, 1993], job postings[Califf, 1998], rental ads [Soderland, 1999], and seminar announcements [Freitag, 1998]. The nature of the task of IE can be summarized as follows:

1. Delimited criteria of relevance/topics are specified in advance
2. Fixed and limited representational format
3. Clear criteria of success are at least possible
4. Corollary features:
   - •Typically only parts of the text are relevant.
   - •Often only part of a relevant sentence is really relevant.
   - •Can be targeted at large corpora.

Historically, the field of IE has employed discrete manipulations in order to process sentences into the critical noun and verb groups. An incoming sentence is tagged for part-of-speech and then handed off to a scaled-down parser or DFA(deterministic finite automaton) which uses local syntax to decide if the elements of a fact are present and to divide the sentence up into logical elements. However, some tasks of IE have been recently addressed successfully using Hidden Markov Models. A Hidden Markov Model is a particular kind of probabilistic model based on a sequence of events e.g., sequential consideration of the words in a text. It is presumed that the individual events are parts of some larger constituents, like names of particular type. In a Hidden Markov Model, it is hypothesized

that there is an underlying finite state machine (not directly observable, hence hidden) that changes state with each input element.

The probability of a recognized constituent is conditioned not only on the words seen, but the state that the machine is in at that moment [Appelt and Israel 1999].

Recent research has demonstrated the strong performance of Hidden Markov Models applied to Information Extraction. Bikel [Bikel et al. 1998] showed in their work that their system, *Nymble* outperforms previous attempts in the task of name finding and other numerical entities. In their work, they performed, using a fairly simple probabilistic model, an F of 90 (See section 4.1.2 for an explanation of the F measure) and above, which means "near-human performance". Andrew McCallum [McCallum et al. 1999] investigated learning model structure from data and the role of labeled and unlabeled data in the training of HMM 's for the task of extracting information from the headers of computer science research papers. Tymothy Rober Leek [Leek 1997] demonstrated the power of Hidden Markov Model to extract factual information from a corpus of machine-readable English prose. His model was used to classify and parse natural-language assertions about genes being located at particular positions on chromosomes.

Our goal is to build a tool that can automatically built a "who knows what"-like system of a scientific community. In our attempt to extract people's name and areas of expertise we decided to apply Hidden Markov Models due to their success in similar tasks as mentioned above. One of our goals is to extract from natural language text names of people, areas of expertise, and then match the similarities between those people. We first describe our HMM model in the next section and then the similarity task.

The corpus used in the design of our model was extracted from personal web

pages of some faculty members from five different universities: Stanford, Harvard, Yale, Princeton and Pennsylvania. Our model consists of nine states as shown in Figure 4.2.

Our model allow us to extract general areas of expertise like "electrical engineering" and specific areas of research. The idea of modeling specific areas of research as in "Stochastic Analysis for Fluid Queueing Systems" is because we would like to provide a better understanding of what is the researcher doing in his/her general field of expertise. In our model each state is linked with a class that we want to extract, such as *person-name*, *field* or *title*. Each class emits words from a class-specific unigram distribution. The model topology (initials, transitions and observations probabilities) were learned from training data. In order to label a new documents with classes, we treat the words from such documents as observations and recover the most-likely state sequence with the Viterbi algorithm. The Viterbi algorithm is a dynamic-programming solution which executes with time linear in the length of the sequence faster than other information extraction approaches that evaluate a super-linear number of sub-sequences. See McCallum and Freitag for more details on this issue.

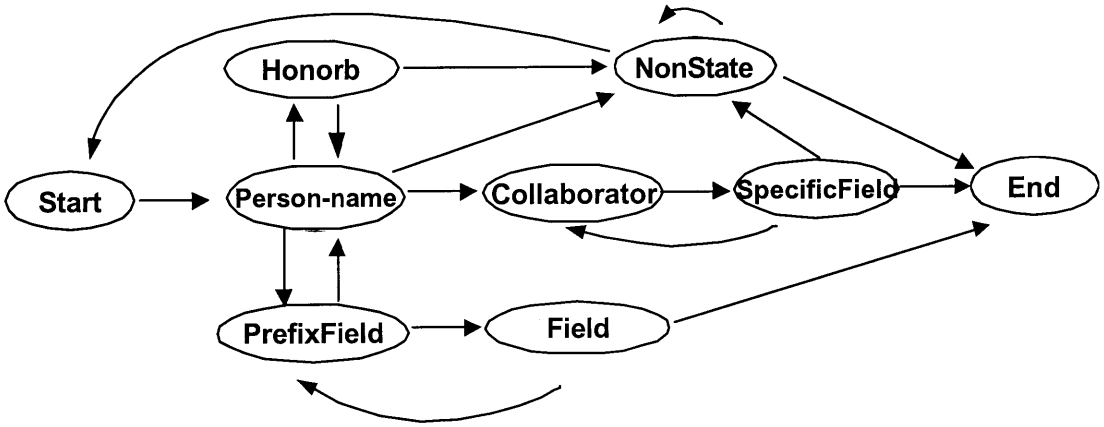

Figure 4.2: Hidden Markov Model

4.1.2 Implementation and test

Our training set consisted of two hundred and fifty six personal faculty web pages from five different universities which were manually labeled. For training we randomly hold-out half of the data and then we performed our test with the rest of the data. In our evaluation we measured both precision and recall.

$$F = (\beta^2 + 1) * RP / (\beta^2 * R + P)$$

where:

P = number of correct answers / answers produced

R = number of correct answers / total possible correct

These two measures are combined to form one, the F-measure (term borrowed from Information Retrieval): where β is a parameter representing relative importance of P and R.

We show results on our main targets: *person-name*, *field*, and *specificfield*

Table 4.1: F-scores for *person-name* task

Random Half-hold (1st run)	83
Random Half-hold 2nd run	81

Table 4.2: F-scores for the *field* task

Random Half-hold (1st run)	81
Random Half-hold 2nd run	79

Table 4.3: F-scores for the *specificfield* task

Random Half-hold (1st run)	85
Random Half-hold 2nd run	84

4.1.3 Further tests of the model

Our model was tested with the research project list of the Media Laboratory (Technotes) consisting of 300 text files. To evaluate the results provided by the tool, we asked people familiar and non-familiar with the Media Laboratory to use the visualization tool to complete some scenarios. For more details on this, see next chapter.

4.1.4 An example

We will show in this section an example of what is our tool taking as input, as well the final output. For this purpose, we first show the original file used for training as well as how it looks after hand tagging. Then we will illustrate how a new (unseen) sample file is processed automatically with our tool by showing the final structure. The following text corresponds to a researcher's profile found publicly available on the web:

Frederick Abernathy Faculty Research Profile

Professor of Engineering Gordon McKay Professor of Mechanical Engineering

B.S., 1951, Mechanical Engineering Newark College of Engineering

S.M., 1954, Ph.D., 1959, Mechanical Engineering Harvard University

Industrial Productivity

Professor Abernathy and colleagues in DEAS, the Economics Department, Harvard Business School, and the Boston University School of Management are concerned with the productivity of the entire manufacturing channel from raw materials to finished products sold at retail. These interests led to the formation at Harvard of a Center for Textile and Apparel Research, which is a part of the Alfred P. Sloan Foundation's Industry Centers Program.

Professor Abernathy his students, and collaborators are involved in developing understanding of how to achieve optimal sourcing of the steps of the production process while minimizing inventory risks. The current focus of the research concerns product proliferation, the logistics of items in production, overall production cycle time, and real options to reduce inventory risk

Abernathy, F. T. Dunlop, J. H. Hammond and D. Weil Retailing and Supply Chains in the Information Age Technology in Society 22 (1) (2000): 5-31.

Abernathy, F. H., J. T. Dunlop, J. H. Hammond and D. Weil

And here is the structure of the manual label:

<NAME>Frederick Abernathy</NAME>

<NONSTATE> Faculty Research Profile </NONSTATE>

<HONORB> Professor</HONORB> <NONSTATE> of </NONSTATE>

<FIELD> Engineering</FIELD>

<HONORB>Gordon McKay Professor</HONORB>

<NONSTATE> of </NONSTATE>

<FIELD>Mechanical Engineering</FIELD>

<NONSTATE> B.S., 1951, </NONSTATE>

<FIELD>Mechanical Engineering</FIELD>

<NONSTATE> Newark College of Engineering </NONSTATE>

<NONSTATE> S.M., 1954, Ph.D., 1959, </NONSTATE>

<FIELD>Mechanical Engineering</FIELD>

<NONSTATE> Harvard University </NONSTATE>

<NONSTATE> Industrial Productivity </NONSTATE>

<HONORB>Professor</HONORB>

<NAME>Abernathy</NAME>

<NONSTATE> and colleagues in DEAS, the Economics Department, Harvard Business School, and the Boston University School of Management </NONSTATE> <PREFIXFIELD>are concerned</PREFIXFIELD> <SPECIFICFIELD>with the productivity of the entire manufacturing channel from raw materials to finished products sold at retail</SPECIFICFIELD>.

<NONSTATE> These interests led to the formation at Harvard of a Center for Textile and Apparel Research, which is a part of the Alfred P. Sloan Foundation's Industry Centers Program. </NONSTATE>

<HONORB> Professor <HONORB> <NAME> Abernathy <NAME> his students, and collaborators are <PREFIXFIELD> involved in </PREFIXFIELD> <SPECIFICFIELD> developing understanding of how to achieve optimal sourcing of the steps of the production process while minimizing inventory risks. </SPECIFICFIELD> <PREFIXFIELD>The current focus of the research concerns</PREFIXFIELD> <SPECIFICFIELD> product proliferation, the logistics of items in production, overall production cycle time, and real options to reduce inventory risk </SPECIFICFIELD>

<NAME> Abernathy, F </NAME>. <COLLABORATOR>Dunlop, J.</COLLABORATOR> <COLLABORATOR>H. Hammond </COLLABORATOR> <NONSTATE> and </NONSTATE> <COLLABORATOR>D. Weil </COLLABORATOR>. <SPECIFICFIELD> Retailing and Supply Chains in the Information Age </SPECIFICFIELD> <NONSTATE> Technology in Society 22 (1) (2000): 5-31. </NONSTATE>

<NAME>Abernathy, F.</NAME> <COLLABORATOR>H., J.</COLLABORATOR> <COLLABORATOR>T. Dunlop </COLLABORATOR> <COLLABORATOR>J. H. Hammond </COLLABORATOR> and <COLLABORATOR>D. Weil</COLLABORATOR>

Now, here is how it looks another researcher's profile. This one was taken from the Media Laboratory.

Justine Cassell is an associate professor at MIT's Media Laboratory, where she directs the Gesture and Narrative Language Research Group. She holds a master's degree in Literature from the Université de Besançon (France), a master's degree in Linguistics from the University of Edinburgh (Scotland), and a double Ph.D. from the University of Chicago, in Psychology and in Linguistics. Cassell and her students study natural forms of communication and linguistic expression, and build the technological tools that enable and enhance these activities, in particular face-to-face conversation and storytelling. After having spent ten years studying verbal and non-verbal aspects of human communication

through microanalysis of human data, Cassell began to bring her knowledge of human conversation to the design of computational systems, designing the first autonomous animated agent with speech, gesture, intonation and facial expression in 1994 during a sabbatical spent at the University of Pennsylvania Center for Human Simulation. Aong with her students, she is currently implementing the newest generation of Embodied Conversational Agent -- "Rea" -- a life-size animated humanoid figure on a screen that can understand the conversational behaviors of the human standing in front of it (using computer vision techniques), and respond with appropriate speech, animated hand gestures, body movements, and facial expressions of its own. The architecture for this new "conversationally intelligent" agents based on an analysis of conversational functions, allowing the system to exploit users' natural speech, gesture and head movement in the input to organize conversation, and to respond with automatically generated verbal and nonverbal behaviors of its own. As well as being a pioneer in this new research area of Embodied Conversational Agents, Justine Cassell has also played a key role in investigating the role that technologies such as these play in children's lives. Interactive technologies such as Sam, the virtual storytelling preen have the potential to encourage children in creative, empowered and independent learning. They can also demonstrate ways for new technology to live away from the desktop, supporting children's full-bodied, collaborative, social play-based learning. Cassell and her students have built a suite of Story Listening Systems that encourage children to tell stories and in doing so to practice decontextualized language of the kind that is essential for literacy. Justine Cassell's other current research and projects include: technological toys for both boys and girls that encourage them to express aspects of self-identity that transcend stereotyped gender categories; technologies that are accessible both to children with high technological fluency, and children with no technological fluency.

Justine Cassell was the director of Junior Summit '98, a program that brought together online more than 3000 children from 139 different countries to discuss how technology could be used to help children, and then gathered 100 of those children in Boston 5 day summit, where they presented their ideas to world leaders and international press.

And here is what our tool extracts (omitting the *nonstate* and *prefixfield* fields)

<name> justine </name>

<FIELD> linguistics </FIELD>

<FIELD> technology </FIELD>

<SPECIFICFIELD> technological toys </SPECIFICFIELD>

<SPECIFICFIELD> students study </SPECIFICFIELD>

<SPECIFICFIELD> stereotyped gender categories </SPECIFICFIELD>

<SPECIFICFIELD> self-identity </SPECIFICFIELD>

<SPECIFICFIELD> particular face-to-face conversation </SPECIFICFIELD>

<SPECIFICFIELD> other current research </SPECIFICFIELD>

<SPECIFICFIELD> nonverbal behaviors </SPECIFICFIELD>

<SPECIFICFIELD> non-verbal aspects </SPECIFICFIELD>

<SPECIFICFIELD> no technological fluency </SPECIFICFIELD>

<SPECIFICFIELD> new technology </SPECIFICFIELD>

<SPECIFICFIELD> natural speech </SPECIFICFIELD>

<SPECIFICFIELD> natural forms </SPECIFICFIELD>

<SPECIFICFIELD> microanalysis </SPECIFICFIELD>

<FIELD> literacy </FIELD>

<SPECIFICFIELD> linguistic expression </SPECIFICFIELD>

<FIELD> learning </FIELD>

<FIELD> language </FIELD>

<FIELD> knowledge </FIELD>

<SPECIFICFIELD> international press </SPECIFICFIELD>

<SPECIFICFIELD> independent learning </SPECIFICFIELD>

<SPECIFICFIELD> human conversation </SPECIFICFIELD>

<SPECIFICFIELD> human communication </SPECIFICFIELD>

<SPECIFICFIELD> high technological fluency </SPECIFICFIELD>

<SPECIFICFIELD> head movement </SPECIFICFIELD>

<SPECIFICFIELD> facial expressions </SPECIFICFIELD>

<SPECIFICFIELD> facial expression </SPECIFICFIELD>

<SPECIFICFIELD> conversational functions </SPECIFICFIELD>

<SPECIFICFIELD> computer vision techniques </SPECIFICFIELD>

<SPECIFICFIELD> computational systems </SPECIFICFIELD>

<SPECIFICFIELD> communication </SPECIFICFIELD>

<SPECIFICFIELD> body movements </SPECIFICFIELD>

<SPECIFICFIELD> autonomous animated agent </SPECIFICFIELD>

<SPECIFICFIELD> appropriate speech </SPECIFICFIELD>

<SPECIFICFIELD> animated hand gestures </SPECIFICFIELD>

<FIELD> systems </FIELD>

<FIELD> psychology </FIELD>

<SPECIFICFIELD> pennsylvania center </SPECIFICFIELD>

<SPECIFICFIELD> narrative language research group </SPECIFICFIELD>

<SPECIFICFIELD> media laboratory </SPECIFICFIELD>

<SPECIFICFIELD> junior summit '98 </SPECIFICFIELD>

<SPECIFICFIELD> interactive technologies </SPECIFICFIELD>

<SPECIFICFIELD> human simulation </SPECIFICFIELD>

<SPECIFICFIELD> conversational agents </SPECIFICFIELD>

<SPECIFICFIELD> conversational agent </SPECIFICFIELD>

<SPECIFICFIELD> world leaders </SPECIFICFIELD>

Here is another example from the other source of information considered in our tool: research papers. Most of the time one will find research papers in pdf, or ps format. In either case, we used gnu tools (ps2ascii, pdf2ascii) to convert those format to pure text. So here is part of the output from a research paper:

<FIELD> knowledge </FIELD>

<FIELD> information </FIELD>

<FIELD> interaction </FIELD>

<SPECIFICFIELD> matching apartments </SPECIFICFIELD>

<SPECIFICFIELD> apartment features </SPECIFICFIELD>

<FIELD> technology </FIELD>

<FIELD> representation </FIELD>

<SPECIFICFIELD> electronic profiles </SPECIFICFIELD>

<SPECIFICFIELD> personal data </SPECIFICFIELD>

<FIELD> systems </FIELD>

<SPECIFICFIELD> artificial intelligence </SPECIFICFIELD>

<SPECIFICFIELD> simple consumer goods </SPECIFICFIELD>

<SPECIFICFIELD> strict query e-commerce </SPECIFICFIELD>

4.1.5 Similarity Task

Once the *profile extractor* was done, we proceeded with the clustering task among people. The *profile extractor* provided us with three pieces of information used: *area(s) of expertise, specific areas* and *collaborators*. The idea behind clustering is to first, match those people who share general areas of expertise; after that, with the goal of being more specific in the process of matching people, we used the *specific field* information (usually titles of papers, or research interests) to provide a better understanding of the similarity; and finally, with the *collaborators* information we provide names of people who have already collaborated with the given researcher in a published paper or those who are mentioned as references. In other words, our one-to-one matching algorithm process as follows:

Given two researchers x and y:

1.-FindSimilarCollaborators(x,y)

2.-ExtractSimilarAreasOfExpertise(x,y)

3.-ExtractSimilarSpecificAreasOfExpertise(x,y)

4.-Show similarity found (degree of similarity)

Let us clarify the "degree of similarity", which is used as input to the visualization tool which is described in section 4.2. Our goal is to be able to connect people, so the first thing that our algorithm looks for, is to figure out if the two researchers being compared shared a common collaborator (See Chapter 3 for an explanation on this). This information is obtained from the output of the *profile extractor* as mentioned before. If that is the case, then we look for potential areas of similar interest. We use a basic taxonomy Appendix A to show similarity in terms of closeness in those cases when an exact match between areas was not found(e.g. if we find someone who is doing research in cinema/video and another doing journal-

ism and photography we can say at least that they are doing media arts, and that chances are that it might be a good idea if they collaborate in a project involving some of the areas mentioned or that someone interest in cinema and video perhaps should take a look at journalism and photography). We borrowed the idea of Resnik [Resnik 1995] but we applied to the taxonomy of areas of study instead of the taxonomy in *WordNet* used by Resnik. His work was mainly about establishing the similarity between concepts in terms of closeness in a given taxonomy. Due to speed performance, we used Isearch, which is an open source C++ package for indexing and searching text documents. Our main reason comes from the fact that our preprocessing files (those generate by the *profile generator*) are with SGML-style mark up and Isearch, as mentioned in its documentation, is capable of indexing such format easily. The Isearch architecture consists of the shell, the search-engine library, and the doctypes. Using this tool we processed the files generated by the *profile extractor* and we, finally, generated the information needed by the visualization tool.

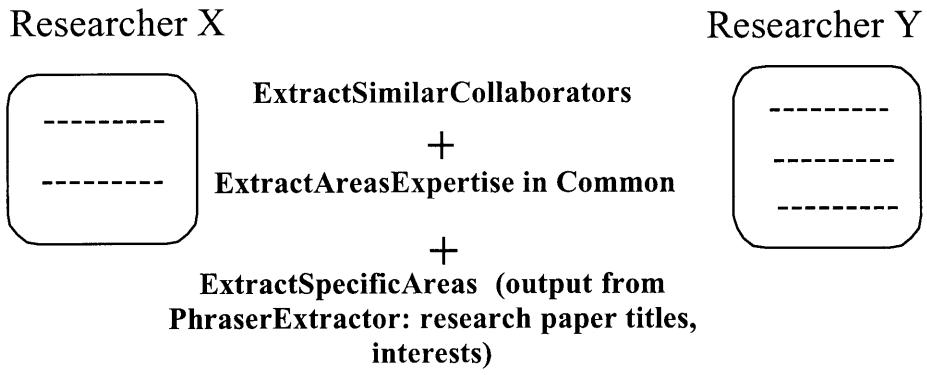

Figure 4.3: This schema illustrates the matching process

4.1.6 Community profiler

This part of the software tool has two main goals: one is to generate those areas of expertise that defines the research done in a given scientific community and two, present the relation that these areas have in terms of which is the most active area, which is the area where the scientific institution is more concern with and who are the people doing more research in those topics. The output from the *profile extractor* is used to extract these areas of research. We decided to augment the information provided by the *profile extractor* by using a POS (part-of speech-tagger) to be able to show some areas of interests that even though are not considered as areas of expertise, help to get a better idea of what is the research institution about (e.g. community development, community building were topics discovered during the processing of the Media Lab research list (technotes)). We called this module the *phraser extractor*. In other words, this part of the tool is capable of showing areas of interests which are useful to get a better understanding of what a given scientific community is about. We use the tagger and build from that a unigram phraser model to extract those topics not included from the list generated by the *profile extractor*.

4.2 Visualization Tool

The main conceptual guideline during the development of the visualization tool was to be able to present relationships between people and institutions, as well as the hierarchy of research areas. During this design process, we followed the references outlined in Chapter 3. In particular, our work includes:

•Dynamic visual presentation of information to present relevant information according to the interaction performed by the user

•Animation to illustrate relationships

•Zooming to show details in those cases where the information is not fully readable

•Panning to give the experience of navigation through the space to focus in some specific information

We used an open-source library for the design of the interface called Jazz. As defined in its documentation, Jazz provides a Java API for building Zoomable User Interfaces (ZUI). It provides support for a general purpose scenegraph with multiple cameras (views). Jazz provides support for many basic operations, visualizations, and interactions. Jazz makes no specific policy about visual or interaction design, but instead provides overridable default behaviors.

Jazz is completely open source. Initially developed, and currently managed at the University of Maryland's Human-Computer Interaction Lab.

4.2.1 Visualization of a Scientific Community

In order to visualize a scientific community, we used the files generated by the *Community Profiler* and the *Phraser Extractor,* in that order. From the *Profile Extractor* we extracted the *field* field. We used different font sizes to illustrate the areas where the scientific community is more interested. So size here helps the

user identify the most important areas of research. The left-hand side of the interface shows the information extracted from the *Community Profiler* and on the right-hand side we show those topics extracted from the *Pharser Extractor*. The user can perform zooming in those areas where he/she is interested or can perform navigation by panning in the scenegraph to focus in the information that she desires. Figure 4.4 illustrates what we have just described.

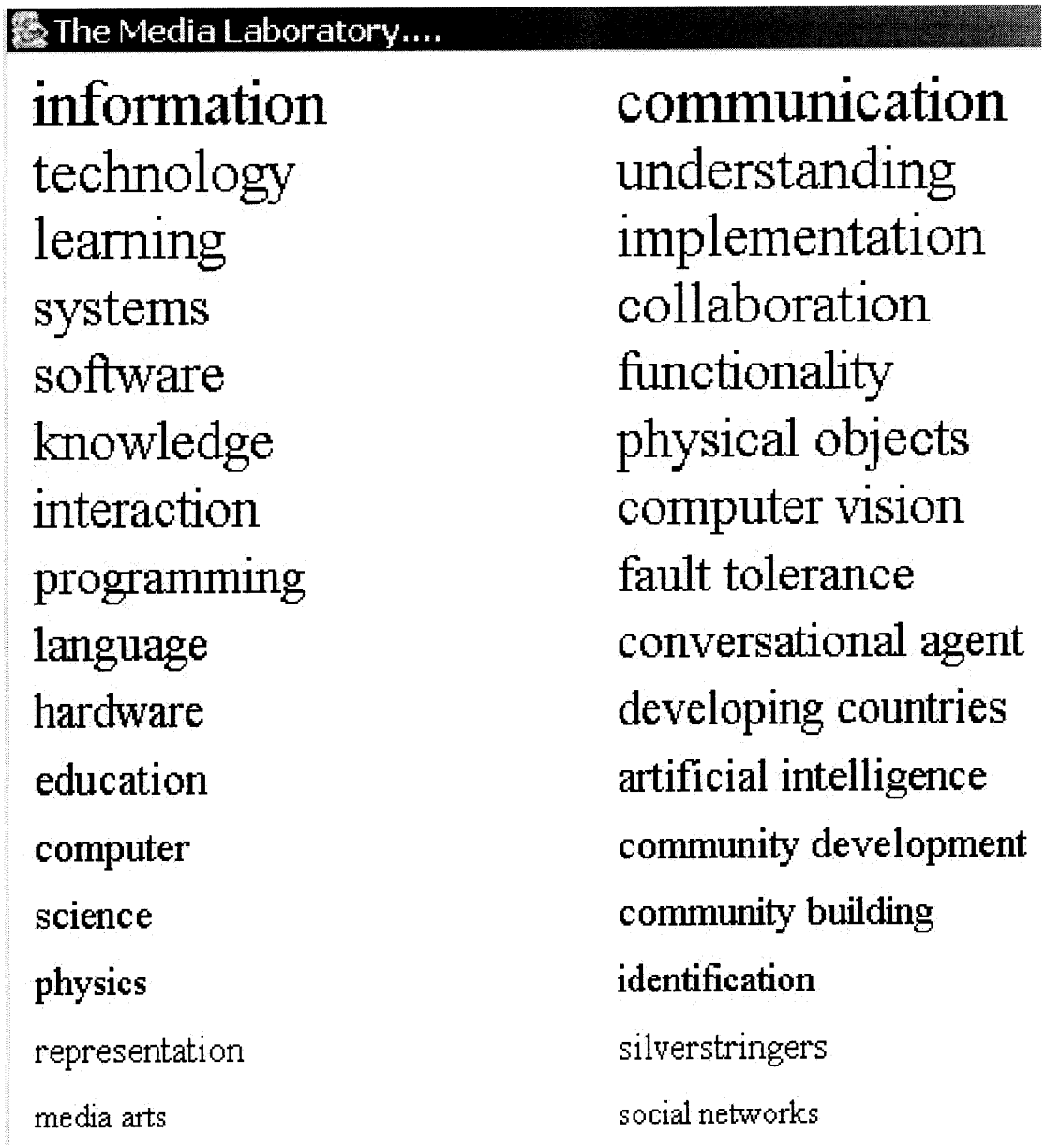

Figure 4.4: Outputs: the Community Profiler and the PhraseExtractor.

Another representation for a given scientific organization is showing its researchers and allowing the user to click on their names to get the researcher relation with his/her colleagues.

Figure 4.5: An example of researcher's similarities

Figure 4.5 shows an example. On the left-hand side one can see a list of researchers. By clicking on one of them, the user can get two things: first, the areas of research the selected member is working on and second, the names of the other members who share interest in each of those areas (right-hand side). Another piece of information the user can have access is to some details of each of the members who are similar to the one selected at the beginning of the interaction by rolling the mouse over the desired researcher.

4.2.2 Visualization of relationships

In the process of designing the visualization of relationships, we used the information generated by the *Profile Extractor*. There are two cases where this visualization would be use: first, when analyzing a given scientific institution and second,

when contrasting two of them.

4.2.3 Contrasting two scientific institutions

This representation consists of showing the list of the researchers of one institution on the left-hand side and the other on the right-hand side. In the middle, we show common research areas. The user can just click in a given area that he /she is interest and he/she is presented with the researchers working in the area selected. Moreover, by just rolling the mouse over the name of a given researcher, the user has direct access to more specific information. To achieve the effect of blurring (see Chapter 3) while showing the information a user is interested, we change the color (for speed performance) of the information not relevant in a given context. Figure 4.6 illustrates how the system presents the user comparing two scientific institutions. Font size determines those areas where there are more similarity among the given institutions.

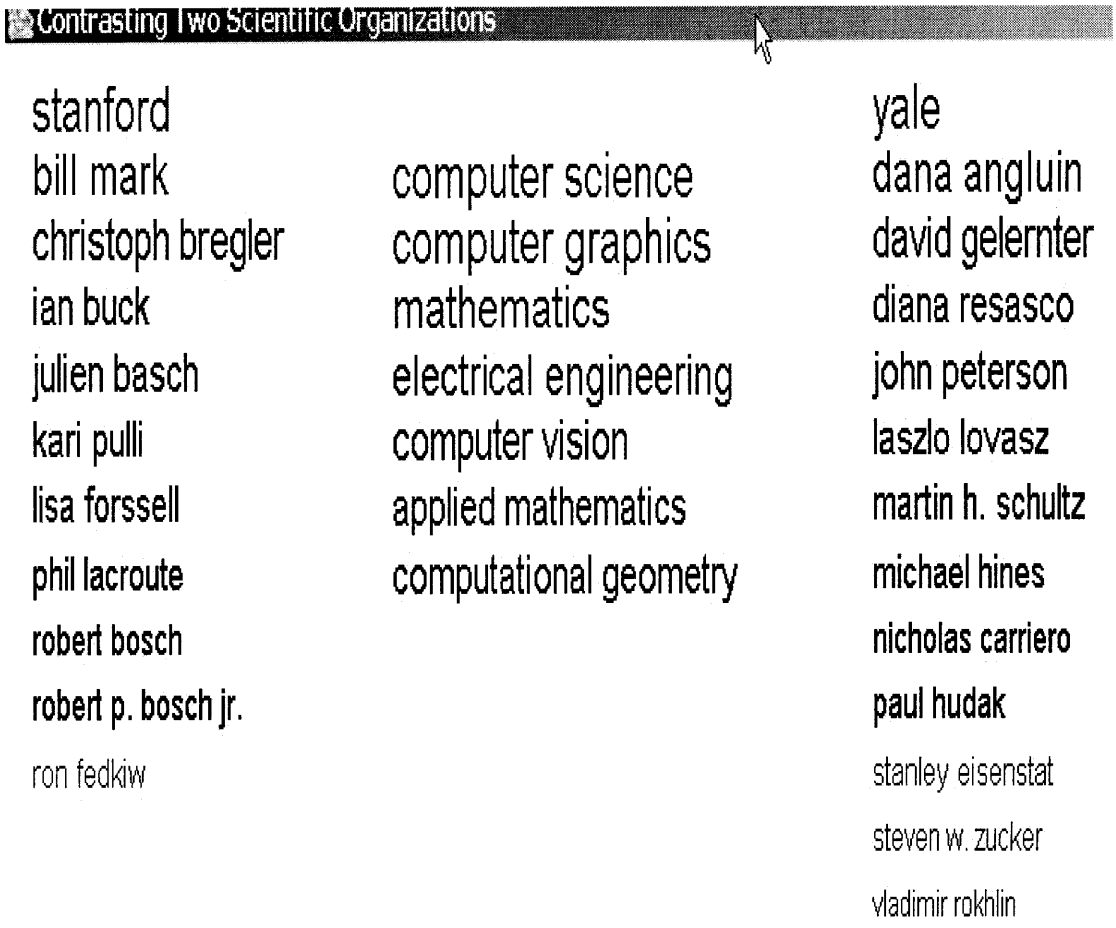

Figure 4.6: Contrasting two scientific organizations

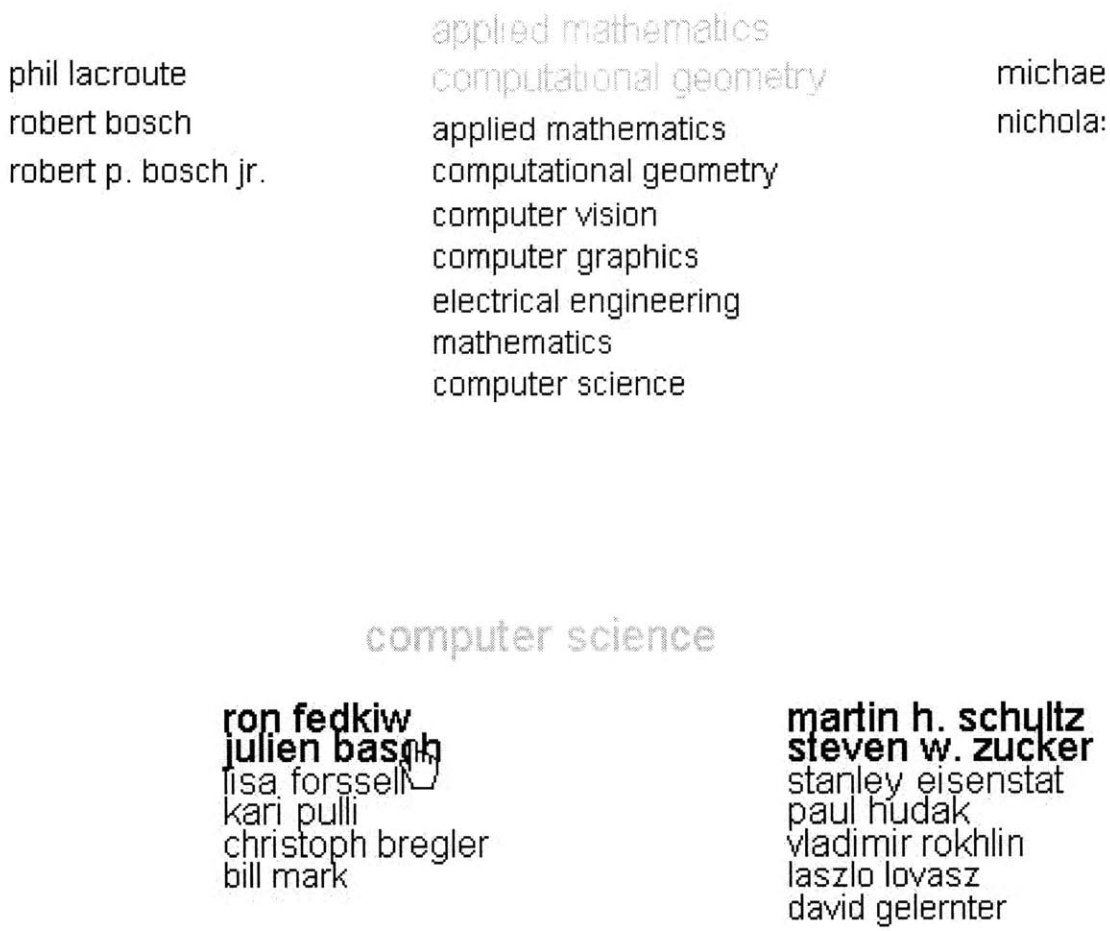

Figure 4.7: Illustrates the common researchers working on the selected area.

Figure 4.7 illustrates an example when a user clicks on *computer science* and the system presents those members of both organizations who are working on it. The example also shows what happens when a user rolls the mouse over a name of a researcher: shows specific areas the researcher is working on. Font size is used to emphasize those members who are working more in the selected area.

Chapter 5

Evaluation

One of the goals of the work proposed here was to provide a tool to facilitate and motivate collaboration and team formation in a scientific community. In this chapter we describe the test performed to evaluate the tool.

5.1 The task

The test consisted in having users performing tasks in two scenarios. In the first scenario, the user was asked about the research areas of a given scientific institution (The Media Laboratory). In the second scenario, the user was asked to form a team capable of building a robot that communicates with humans by voice and maintains a conversation.Two groups of ten people each were randomly selected. One of the groups used our system, referred as Kinix in the following discussion and the other group, referred as Control, used the Media Lab website. Users in the Control group needed at least 40 minutes to complete the tasks. Users in this group, were told to use the resource provided (a computer connected to the Internet) to complete the tasks. Users in the Kinix group were told to use our tool to complete the tasks and they were told about the basic functions of our tool (mouse over, click, double click, zooming in and zooming out). They usually needed at least 25 minutes to complete the tasks. We used a questionnaire (Appendix B) to evaluate the users. We identified three different categories from the results of our questionnaire: ease of use, user satisfaction and system efficiency.

5.2 Method

Because the type of the data (ordinal)) collected by the questionnaire, we used the Mann-Whithney Test. The Mann-Whitney Test is a nonparametric test for the significance of the difference between the distributions of two independent samples, A and B, of sizes n_a and n_b, respectively [Frank 1994]. For details in the procedure of the Mann-Whitney Test see [Lowry].

5.3 Hypothesis tests

In the three categories we used a directional hypothesis, claiming that the Kinix group will be more effective in the completion of the tasks than the Control group that used the Media Lab web site.

5.3.1 Ease of use

We need to prove that $U_a < U_b$ ($T_a > T_b$). In this case, U_a=11 and U_b=89. Using a table with the critical values of U, we can say that the result is significant with the.01 level for a directional test. Table 5.1 shows the critical values of U for n_a = 10 and n_b = 10 and Table 5.2 shows mean ranks for both samples.

Table 5.1: Critical values of U for n_a = 10 and n_b = 10

Level of significance for a directional test			
	0.05	0.025	0.01
lower limit	27	23	19
upper limit	73	77	81

Table 5.2: Mean ranks ease of use category

Mean ranks for:	
Kinix	Control
14.4	6.6

Users in the Control group found that even though it is easy to navigate a web site, if they are presented with tasks such as in our experiment, it is hard to find the information they need to complete such a task. Users in the Kinix group found that, even though at the beginning of the task they were confused with the layout of the data, it was easy to learn how to use the tool provided. Once they understood it, in general, they concluded that it was easy to complete the task.

5.3.2 User satisfaction

With $U_a = 10$ and $U_b = 90$ we can say that the result, as the one above, is significant with the .001 level for a directional test. Table 5.3 shows the mean ranks for both samples.

Table 5.3: Mean ranks for user satisfaction category

Mean ranks for:	
Kinix	Control
14.5	6.5

Users in the Control group found in the context of completing the task of team formation, that the layout provided by the Media Lab web site is not clear enough. In particular, they found that using the web site is definitely not pleasant for this type of task. The Kinix group reported that the layout of the information in Kinix was not clear. Some of them did not understand that font size was actually telling something. However, they found the layout of the information convenient for the team-building task.

5.3.3 System efficiency/functionality

In order to evaluate system efficiency, we present two results here: one the outcomes of this category from the questionnaire; and, second, the results from the task of team formation. Results from the questionnaire showed a $U_a = 10$ and $U_b = 90$ just as in the user-satisfaction task. So the conclusion is the same. To evaluate

results from the team-formation task itself, we used a t-test, because our data is not ordinal in this case. We selected the five "ideal" people for the task (using our system) and compare it against user's selection. Users in the Control group used the Media Lab search engine in the web site to perform the selection of team members. In general, the Control group looked for keywords such as robotics, hardware, system, software, artificial intelligence and communication. Table 5.7 summarizes the results where each number corresponds to the number of correct answers of each subject.

Overall, there were two points that users in the Control group complained about: one was the time it took for them to complete the task —they bookmarked web pages of researchers they thought could be potential members of the team and then going back to compare them among new findings; and two was in regard to the information provided by the web site —they said that in the way that was provided, it was not easy to understand relationships among members of the community. Those users who tried to search on a per-research group basis were disappointed by the lack of connections among groups.

Users in the Kinix group found two things important: 1) the way the information was structured helped them to decide their team members; and 2)they required little time to complete the task.The Kinix group did suggest improvements to the functions and capabilities of the system. In particular, users asked for more levels of detail —they wanted to be able to see for example, research papers at different levels of details i.e., abstracts at one level, introductions at a second level, conclusions at a third level, and the complete paper as a last level of detail.

With $t = 3.8$ and $df = 18$ we proved our directional research hypothesis that people using our tool, the Kinix group, will perform better than the Control group, and as

our observed result, $M_{xa}-M_{xb}=1.40$, proved consistent with that hypothesis, the relevant critical value is between 2.88 and 3.92 for a .005 level of significance. Table 5.5 summarizes the data and Table 5.6 shows a fragment of the critical values of t used.

Table 5.4: Number of correct answers of each subject.

Kinix	Control
5	3
3	3
4	2
3	3
4	2
2	3
5	2
4	2
3	1
4	2

Table 5.5: Summary Data

Kinix	Control
$N_a = 10$	$N_b = 10$
$M_a = 3.7$	$M_b = 2.3$
$S_{sa} = 8.1$	$S_{sb} = 4.1$
$M_a - M_b = 1.4$	

Chapter 6

Conclusions

This research has introduced a software tool capable of facilitating team formation in an organization setting as well as contrasting two given scientific organizations in terms of areas of expertise. Furthermore, the visualization module designed was proved to be effective in illustrating those capabilities. In this chapter we describe our contributions and possible improvements and future work that could enhance this research.

6.1 Contributions

The initial hypothesis that our tool could facilitate team formation was proved by the evaluation described in Chapter 5. Our approach in the text-mining module, using a novel application of HMM as a more powerful processing tool, was also proved to be effective. It was because of the performance of the processing phase that we were able to implement the matching process quite efficiently as the results of the evaluation revealed. The analysis of the tools developed by this research, to the best of our knowledge, is the most complete in the context of characterization of scientific communities in terms of areas of expertise. We not only evaluated the performance of the text-mining itself but also performed subjects tests to figure out whether the text-mining combined with the visualization designed were capable of providing the information necessary to prove our hypothesis.

6.2 Improvements and future work

With the feedback and comments of the users, we expect to add more functionality to the visualization tool, such as the capability of displaying the full content of papers and web pages within the application in those cases where the user desires to have this ultimate level of information detail. We expect also to be able to perform more training in the text-mining part to make the tool more robust. Another interesting future work is to add commonsense capabilities to the tool developed. Commonsense, as Marvin Minsky in his book "The Society of Mind" defines is the mental skills that most people share. Commonsense thinking is actually more complex than many of the intellectual accomplishments that attract more attention and respect, because the mental skills we call "expertise" often engage large amounts of knowledge but usually employ only a few types of representations. In contrast, common sense involves many kinds of representations and thus requires a larger range of different skills. The idea is to be aware of the limitation we are facing here: there is not going to be (ever) enough data to train our tool in order to make it perfect. One of the things to do about it, is to add some "commonsense" methodology to respond in cases where the best "guess" of the system is below certain threshold or when a response of the system is not good enough because it does not "make sense". As an example, let us consider what happens when our system finds a researcher whose areas of expertise are computer science and electrical engineering and there is no more details about him. The system extracted such information and now it needs to figure out whether there is someone else similar in the scientific community or not. Without more information, our tool will perform a match with those people who have some work and/or background in the areas mentioned. However, is this enough? Can we do it better? Probably,

one can imagine having a basic commonsense-like module which somehow "knows" that the given researcher is a member of a group where research in nanotechnology is being done and "understands" that it will make sense, based on that, to request more details of each of the areas addressed by the nanotechnology group and from there seek for other members of the scientific community working in similar topics.

Appendix A Taxonomy

- Arts
  - Arts-History, Theory, and Criticism
    - Art Therapy
    - Arts Administration
    - Arts History
    - Art History & Criticism
  - Design Arts
    - Computer Arts
    - Fashion/Textiles Design
    - Graphic Design
    - Illustration
    - Industrial Design
    - Interior Design
    - Jewelry/Metalsmithing
    - Medical Illustration
    - Regional/Urban Design
    - Set/Theatre Decoration/Design
  - Media Arts
    - Cinema/Video
    - Journalism
    - Photography
    - Radio

See http://mati.eas.asu.edu:8421/hed/stats.html for the complete taxonomy

Appendix B System Evaluation Questionnaire

This is the questionnaire used in the evaluation (Chapter 5).

1. Overall, I am satisfied with the ease of completing the scenarios
2. It was simple to use this system
3. I can effectively complete my task using this system
4. I am able to complete my task quickly using this system
5. I am able to easily complete my task using this system
6. I feel comfortable using this system
7. It was easy to learn to use this system
8. I believe I became productive quickly using this system
9. It was easy to find the information I needed
10. The information provided for the system was easy to understand
11. The information is effective in helping me to complete the tasks and scenarios
12. The layout of the information is clear
13. The interface of this system is pleasant
14. I like using the interface of this system
15. This system has all the functions and capabilities I expect it to have
16. Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario
17. Overall, I am satisfied with this system

Each user answered each question using a likert scale from 1 to 7 (strongly disagree to strongly agree).

Questions 2,4,5,7 were used as the source for the ease of use category

Questions 1,6,12,13,14 were used as the source for the user satisfaction category

Questions 3,8,9,10,11,15,16,17 were used for the system efficiency/functionality category

This questionnaire was adapted from the one designed by James R. Lewis at IBM. See [Lewis,1995] for more details.

References

Appelt Douglas E. and Israel David J. *Introduction to Information Extraction Technology.* A Tutorial Prepared for IJCAI-99. 1999.

Bikel M. Daniel, Miller Scott, Shwartz Richard and Weischedel Ralph. *Nymble a High-Performance Learning Name-Finder.* 1998

Card, S.K., Robertson, G.G., and York, W. *"The WebBook and the Web Forager" An information workspace for the World Wide Web".* Proceedings of CHI 1996, ACM Conference on Human Factors in Software.

Dugad Rakesh and Desai U. B. *A tutorial on hidden Markov Models.* Signal Processing and Artificial Neural Networks Laboratory Department of Electrical Engineering. Indian Institute of Technology.

Erickson Thomas and Kellogg A. Wendy. *Knowledge Communities: Online Environments for Supporting Knowledge Management and its Social Context* IBM T.J. Watson Research Center

Foner, Leonard N. *Yenta: A Multi-Agent, Referral-Based Matchmaking System.* Proceedings of the First International Conference on Autonomous Agents 1997, Marina del Rey, CA

Frank, Harry and Althoen, Steven C. *Statistics: Concepts and Applications.* Cambridge University Press. 1994.

Gershon Nahum, Erick Stephen G. and Card Stuart *Information Visualization* Interactions, march+april 1998

Grishman Ralph *Information extraction: Techniques and Challenges* Computer Science Department New York University 1996

Kosara Rober, Hauser Helwig and Miksch Silvia. *Focus+Context Taken Literally.* IEEE Computer Graphics and Applications. January/February 2002.

Leek R., Timothy. *Information Extraction using hidden Markov models.* Master's thesis, UC SanDiego, 1997.

Lewis R.,James IBM Computer Usability Satisfaction Questionnaires. International Journal of Human-Computer Interaction. 1995 v. 7 n. 1 p 57-78.

Lowry, Richard Vassar College Poughkeepsie, NY (http://faculty.vassar.edu/lowry/webtext.html)

McCallum, Andrew, Seymore Kristie and Rosenfeld Ronald. *Learning hidden Markov model structure for Information Extraction.* AAAI 1999 Workshop on Machine Learning for Information Extraction.

McCallum Andrew and Freitag Dayne *Information Extraction with HMM Structures Learned by Stochastic Optimization.* Just Research

Minsky, Marvin. *The Society of Mind.* New York: Simon and Schuster, 1986.

Murtaugh M *The Automatist Storytelling System: Putting the Editor's Knowledge in Software* MIT Masters Thesis, 1996

Rabiner L.R. *A tutorial on hidden Markov models and selected applications in speech recognition.* Proceedings of the IEEE, 77(2), February 1989

Resnik, P. *Using information content to evaluate semantic similarity in a taxonomy.* In Proceedings of IJCAI-95, pages 448–453, Montreal, Canada. 1995

Sebrechts Marc M., Vasilakis Joanna, Miller Michael S. Cugini John V. and Laskowski Sharon J. *Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces.* SIGIR 1999 Berkley, CA USA

Small, David. *Navigating large bodies of text.* IBM System Journal, Vol. 35, No. 3&4, 1996

Vivacqua, A. S. *"Agents for Expertise Location".* In The Proceedings of the AAAI Spring Symposium on Intelligent Agents in Cyberspace, Stanford, CA, March 1999

Ware, Colin. *Information Visualization: Perception for design.* Morgan Kaufmann 1999