Internet Information Aggregation using the Context Interchange Framework

by
Benny Suen

Submitted to the
Department of Electrical Engineering and Computer Science

May 19, 1998

In Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Electrical Engineering and Computer Science
and Master of Engineering in Electrical Engineering and Computer Science

Author____
Department of Electrical Engineering and Computer Science
May 18, 1998

Certified by____
Stuart E. Madnick
Thesis Supervisor

Accepted by____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Internet Information Aggregation using the Context Interchange Framework

by
Benny Suen

Submitted to the
Department of Electrical Engineering and Computer Science

May 19, 1998

In Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Electrical Engineering and Computer Science
and Master of Engineering in Electrical Engineering and Computer Science

Abstract

Context differences in the Internet can lead to large-scale semantic heterogeneity and obstruct meaningful information exchange. The Context Interchange (COIN) project aims to tackle the challenges in this area and targets to provide context mediation services. This project, Internet Information Aggregation Agent (IIAA), using the COIN framework, aims to facilitate the gathering of information from the Internet with a focus on company research. We will look into the company research process and discuss the issues involved in automating the process. Two main topics for discussions are *Question Answering Process* and *Knowledge Capturing*. Three problem decomposition techniques we can use are *computation decomposition*, *unknown substitution* and *iterative expansion*. Knowledge on data sources, ontology (company information, industry, geography, Internet domain name), question keywords, statistics, financial reports and accounting are important to the system.

Thesis Supervisor: Stuart E. Madnick

Title: John Norris Maguire Professor of Information Technology,
Sloan School of Management

Acknowledgments

I would like to thank my thesis advisor, Professor Stuart Madnick, and Michael Siegel, for giving me the opportunity to work in the Context Interchange project. I would like to thank Professor Madnick for his advice, support and consideration. I would like to thank the entire Context Interchange team, especially Steve Tu for his guidance and support and Tom Lee for his insightful suggestions.

I would also like to thank all of my friends at MIT. It definitely has been one of the greatest moments, with all of you in my life.

Finally, I would like to thank my parents for their love and support.

Contents

1 INTRODUCTION 6

1.1 MOTIVATION 6

1.2 GOAL OF THE INTERNET INFORMATION AGGREGATION AGENT (IIAA) 7

1.3 IMPACT OF THE IIAA PROJECT 9

1.3.1 Impact on the COIN framework: 9

1.3.2 Impact on Automated Research on the Internet 9

1.4 OBJECTIVE OF THESIS 10

1.4.1 Research Questions 10

1.4.2 Focus of Thesis 11

1.5 STRUCTURE OF THESIS 11

2 BACKGROUND 13

2.1 CONTEXT INTERCHANGE SYSTEM 13

2.1.1 Motivation 13

2.1.2 Relationship between IIAA and the COIN framework 13

2.1.3 Integration of IIAA and the COIN framework 14

3 RESEARCH EXERCISE 15

3.1 MOTIVATION 15

3.2 GOAL 15

3.3 SCENARIO 15

3.4 QUESTIONS 16

3.5 GATHERING DATA SOURCES 17

3.5.1 Source Description 17

3.5.2 Example 20

3.6 LOGIC FLOW OF RESEARCH EXERCISE 20

3.6.1 Conducting a research on Affymax 20

3.6.2 Conducting a research on Affymetrix 26

3.7 OBSERVATIONS 30

3.7.1 Problem Solving Methodology 30

3.7.2 Knowledge 31

4 QUESTION ANSWERING PROCESS 35

4.1 CATEGORIZATION OF QUESTIONS 35

4.1.1 Keyword Question vs. Essay Questions 35

4.1.2 Asking for something common / uncommon 36

4.1.3 Example of Questions 37

4.1.4 Resolving Obscurity 39

4.1.5 Dealing with the different types of questions 40

4.2 HOW DOES HUMAN ANSWER A QUESTION? – A THOUGHT EXPERIMENT 41

4.2.1 On Problem Solving Methodology: 43

4.2.2 On Knowledge – What do we need to know? 43

4.2.3 Section Summary 44

5 PROBLEM SOLVING METHODOLOGY 45

5.1 OVERVIEW 45
5.2 BACKWARD CHAINING 45
5.2.1 Backward Deductive Chaining *46*
5.2.2 Issues that needs to be resolved *47*
5.3 PROBLEM DECOMPOSITION 47
5.3.1 How do we do it? *48*
5.3.2 How may the system do it? *48*
5.3.3 Decomposition Techniques *50*
5.4 COORDINATION BETWEEN BACKWARD CHAINING AND DECOMPOSITION 52

6 KNOWLEDGE 53

6.1 KNOWLEDGE ON DATA SOURCES – SOURCE DESCRIPTION 53
6.1.1 Introduction *53*
6.1.2 Challenge *53*
6.1.3 Issues and suggestion on the current source description scheme *54*
6.2 ONTOLOGY 60
6.2.1 Ontology of Company Information *60*
6.2.2 Industry Ontology *62*
6.2.3 Geography Ontology *67*
6.2.4 Internet Domain Name Ontology *67*
6.3 KEYWORD KNOWLEDGE 67
6.4 FINANCIAL REPORT / ACCOUNTING KNOWLEDGE 67
6.5 STATISTICS KNOWLEDGE 68

7 OPEN ISSUES 69

8 CONCLUSION 70

9 APPENDIX 71

APPENDIX A DATA SOURCES GATHERED IN EXERCISE 71
Appendix A-1 General Data Source *71*
Appendix A-2 Computer Industry *82*
Appendix A-3 Telecommunication Industry *84*
Appendix A-4 Pharmaceutical Industry *89*
APPENDIX B LOG OF LOGIC FLOW 94
Appendix B-1 Computer Industry *94*
Appendix B-2 Telecommunication Industry *96*
APPENDIX C YAHOO! CLASSIFICATION 98

1 Introduction

1.1 Motivation

The phenomenal growth of the Internet has provided a vast source of information. However, due to the inherent semantic heterogeneity in human languages, correct interpretation of information has become a major challenge. An information source may operate in a different context than the potential receivers'. For example, a US reader may misinterpret an amount in pounds (posted by a UK web site) as a US dollar amount. When the actual currency is not mentioned, people in the US assume the amount is in US dollar while people in the UK assume it is in pounds. This is an example of different contexts. Context differences lead to large-scale semantic heterogeneity and obstruct meaningful information exchange.

The Context Interchange (COIN) project aims to tackle the challenges in this area and targets to provide context mediation services. The key to the COIN framework is the notion of context, which determines the underlying meaning and interpretation of the information. A context is the collection of implicit assumptions about the *context definition* (i.e., meaning) and *context characteristics* (i.e., quality) of the information.

One natural application of the COIN framework is the aggregation of information on the Internet with context mediation. Before the emergence of the Internet, people may have to rely on magazine articles, journals and human experts to gather information on a topic. As the Internet expands, people can reach an abundance of information with a click on the mouse. There are web sites and data sources that provide information on almost every topic. People can find out virtually anything from the web: from microorganisms to galaxies light-years away, and from prehistoric myths to the latest technologies. The Internet is taking on a more and more important role as the center of research.

In the business world, particularly in the management consulting and finance areas, one needs to research a company or an industry frequently. Management consulting firms need to know the clients, the competitors of the clients and the industry in order to give advice to their clients. Equity research analysts and portfolio managers need in-depth knowledge about a company's financial situation.

Researching a company or an industry can be extremely laborious. A researcher may need to visit industry-specific sites to get acquainted with the industry. She may have to visit company web sites to get an overview of the company, and to some investor web sites to get financial data. She may also need to use search engines, and pick out the few useful web pages from many irrelevant links.

The research experience can be especially inefficient and laborious when there are numerous companies to be compared, and when it involves various data sources. The different sites may have discrepancies in formats, definition of terms, currencies, denominations, etc. These context differences make direct comparison impossible.

As a joint effort from the industry and the academia, this project – Internet Information Aggregation Agent (IIAA) – aims to facilitate the gathering of information from the Internet with a focus on company research.

1.2 Goal of the Internet Information Aggregation Agent (IIAA)

IIAA facilitates the process of conducting a company research. IIAA uses the Context Interchange framework to assist the company / industry research on the Internet. We suggest to develop IIAA in several stages:

Stage 1

The system should be able to handle a number of preset questions, which would include some data look up. The exact questions should be determined by the availability of information, which will be discussed later.

Stage 2

As an extension to the preset questions, the system should be able to answer these type of questions:

- Single company data lookup - searches for a company's information on selective fields.
- Multiple company data comparison - compares multiple companies' operation information.

Stage 3

The system should also answer more difficult questions. The system should handle queries that look for companies with certain characteristics using Boolean operators. For example, companies with a P/E ratio smaller than 15, or an ROA larger than 50%, etc.

The system should allow the user to control the cost of the research. This will allow the user to take advantage of both free and fee-based sites. The system should also enable the user to control the quality of the research. As a central issue in the COIN framework, the user should also be able to track the quality of information.

Although there are three stages of the project, we will not limit our discussion to Stage 1 of the project in this thesis. We will discuss issues that are important to any stages.

Primarily, the focus will be placed on companies in the computer, telecommunication and pharmaceutical industries. However, an open, scalable architecture for the system will be adopted for future expansion.

1.3 Impact of the IIAA project

The goal of the IIAA project is manifold. We would like to find out the feasibility of the current COIN framework and also the feasibility of an Internet Information Aggregation Agent.

1.3.1 Impact on the COIN framework:

A main focus of the project is building an application using the Context Interchange framework. We hope to identify potential weak points of the framework and augment the framework for more flexibility. IIAA represents a experimental infrastructure that is ideal for exploring a number of ideas. We would answer questions such as:

- Is the current COIN framework appropriate for a task such as conducting a research on the Internet?
- What augmentation is needed to deal with the semi-structuredness of web sites.
- What changes in the descriptor file language or domain model is required to capture the context definition and context characteristics of web sites?

1.3.2 Impact on Automated Research on the Internet

IIAA signifies one of the first attempts in automating the research process with semi-structured web sites. Source selection, data extraction, and problem solving mechanisms are all required to enable the automation of the company research process.

Research in these areas will facilitate the building an intelligent agent in information aggregation.

1.4 Objective of Thesis

1.4.1 Research Questions

We will look into the company research process and discuss the issues involved in automating the process. The following research questions are central to IIAA. These questions are arrived with a top-down approach. We asked ourselves, "What do we need to know in order to build IIAA?" The questions can be grouped into five main areas:

1. Company Research

- What are some questions that are commonly asked?
- What are some interesting questions for which the answering process is non-trivial?
- What are the representative web sites for company information? What information do they provide and, what are the differences in the contexts of these sites?

2. Question Answering Process

- How do we answer a question? Is there a general methodology that we use?
- How do we know if we can answer a question with the available data sources?
- How do we know if we need to break up the question into smaller pieces before we can answer it?
- How do we break up a question?

3. What Knowledge do we need to Capture?

- What does the system need to know in order to answer users' questions?

- What does the system need to know to determine if an answer is answerable?
- What does the system need to know about the sources?

4. Source Selection

- How do we know which data sources are needed to answer a question?
- How do we choose among the data source if multiple sources lead to the answer?
- How does the agent take fees and quality of information into consideration?

5. Data Attribution

- How do we know if the answer is accurate, up-to-date answer?
- How do we compare the quality of the answers if we can get the answer in multiple ways?

1.4.2 Focus of Thesis

For this project, we have conducted a research exercise to illustrate the parameters and range of accounts pertaining to the problem. In this thesis, we will summarize the research exercise, and the topics that have surfaced as important issues. Due to the complexity of the problem, addressing all the issues would be impossible. We will focus on issues in *Company Research, Question Answering Process* and *Knowledge Capturing*. Some *Source Selection* issues will also be touched upon.

1.5 Structure of Thesis

In chapter 2, we give background. We will discuss the COIN framework that IIAA bases on. We will also give a quick overview of how IIAA will integrate into the COIN framework.

In chapter 3, we will summarize the research exercise we have conducted. We will discuss the scenario of our exercise, our first trial in source description, and the logic flow of answering the questions. Then, we report our observations and allude to the interesting issues.

In chapter 4, we discuss the question answer process. We categorize different types of questions in order to determine what questions the system can handle, and to find a common question answering methodology. Through a generalized example, we realize that we need a problem solving methodology and a set of knowledge in order to answer a question.

In chapter 5, we will discuss the problem solving methodology. Backward deductive chaining and problem decomposition are two techniques that can be used in solving a problem.

In chapter 6, we will discuss the different bodies of knowledge required for IIAA: data source knowledge (source description), ontology, keywords, financial report / accounting knowledge, and statistics knowledge.

2 Background

2.1 Context Interchange System

2.1.1 Motivation

The immense growth of the Internet has provided vast source of information; however, each source of information and potential receiver of that information may operate with a different context. This leads to large-scale semantic heterogeneity, and correct interpretation of information has become a major challenge. A context is the collection of implicit assumptions about the context definition (i.e., meaning) and context characteristics (i.e., quality) of the information. The Context Interchange (COIN) project aims to tackle the challenges in this area and targets to provide context mediation services.

Context may vary in three major ways. First context varies due to geographical differences. Second, there are functional differences, even within the same organization and location, different functional areas interpret and use information differently. Third, there are organizational differences. The information used in the same function, in the same industry, in the same country, can have different meanings between two companies. For example, a receiver of a piece of information on price may misinterpret a price in pounds to be in US dollars.

2.1.2 Relationship between IIAA and the COIN framework

The core function of IIAA would include meditating contextual differences in the different sources. For example, when a user wants to find the companies in the computer industry with the largest market share, we want to make sure that the market share numbers are of the same market. (This is not an easy question to answer by the way!)

2.1.3 Integration of IIAA and the COIN framework

Here is a high level description to give the reader a general idea on where the COIN engine would fit in IIAA.

1. IIAA provides a graphical user interface for users to select the type of research to be conducted.
2. IIAA translates the graphical request into a query.
3. IIAA devises the plan to answer the questions - selects data sources, breaks the query into smaller queries if necessary.
4. IIAA passes the requests to the context mediator.
5. The context mediator rewrites the query posed in a receiver's context into a mediated query and resolves all potential conflicts explicitly
6. The context mediator retrieves company data from different web sites and databases through their respective wrappers.
7. The context mediator translates the information according to the user's context and returns them to IIAA.
8. IIAA displays the information. IIAA can possibly be extended to generate charts and reports automatically

3 Research Exercise

3.1 Motivation

In order to better understand the nature and scope of the problem, we have conducted a research exercise to illustrate the parameters and range of accounts pertaining to the problem.

3.2 Goal

From this exercise, we would like to

1) identify a number of non-trivial web sites that contain relevant information about companies,
2) identify interesting questions that one may pose,
3) find out what issues would become important in automating the research process, and
4) find a research methodology that is applicable across different industries, while customizable to the specific requirement of different industries.

We have performed the exercise across three different industries - the computer industry, the telecommunications industry and the pharmaceutical industry - to find out the varying types of sources and a common methodology in conducting a research. In this thesis, we will only discuss the exercise on the pharmaceutical industry. However, the discussion can be generalized to other industries. The logic flows of the exercise on computer and telecommunications industry is included in Appendix B.

3.3 Scenario

We based our exercise on a "conflict of interest" scenario. Suppose that a company has a policy of not taking a new client that is a competitor of its existing clients, because working for both sides would cause a conflict of interest. Now the company is considering taking a new client, Affymax. To ensure the observance of the policy, the company needs to make sure that none of Affymax's subsidiaries or parent is a competitor of its existing clients. The company needs to find all the subsidiaries and parents of Affymax, and then check if there are any industry and product overlaps with the existing clients. To double check for maximum safety, the company also decides to collect some directory information, such as the subsidiary companies' addresses, to inquire about subtle competitive relation.

3.4 Questions

We have come up with a list of ten questions as questions of interest:

- What are the subsidiaries / parent of Affymax?
- What industry is Affymax in?
- What is the incorporate/headquarters address of Affymax?
- What products does Affymax produce?
- What parts of the world do these products get sold to?
- Who are the competitors of Affymax in terms of similar SIC code and products?
- Who are the competitors of Affymax in terms of other competitive relations?
- Is there any officer changes of Affymax?
- Who are the officers and members of the board of directors (useful for conflict of interest detection)?
- Benchmarking figures for performance (not necessarily financial measures) analysis?

3.5 Gathering Data Sources

We have gathered a number of general and industry-specific sources. We will try to answer the questions with these sources.

3.5.1 Source Description

As a first run, we tried to describe the sources with the following fields.

Data Source Name

- the name of the data source

Original Source

- the original data sources that this source is compiled from

URL

- the URL address to either the data source, or to the site with links to the actual data source

Query Interface

- Types of interface for accessing the data source

1. *Restricted Domain:* Users are presented with columns of selections to formulate conditions, where the possible values of each column are pre-coded
2. *Unrestricted Domain:* Users are presented with columns of selections to formulate conditions, where the possible values of each column are not pre-coded but the type of each column is pre-defined
3. *Full text search:* Users are allowed to type in any keyword without either value or type restriction
4. *Browsing:* Users are presented with a list of anchors to articles which the user can read/browse

Query input

- The information that needs to be typed in the case of restricted domain

Query result

- The structure of which results are returned

- HTML formatted table
- HTML tagged text documents
- HTML anchors to documents

Attribute domain (category)

- The category of information to be returned

1) Stocking trading
   a. High
   b. Low
   c. Volume
   d. Average
   e. Sec Exchange
   f. ...
2) Financial
   a. Sales
   b. Asset
   c. Debt
   d. Profit
   e. ...
3) Directory
   a. Address
   b. CEO names
   c. Board of director
   d. Phone#
   e. ...
4) Background
   a. Employee#
   b. Subsidery
   c. Founders
   d. ...
5) Marketing
   a. Product
   b. Brand
   c. SIC code
   d. Market share
   e. ...

Attribute domain (items)

- The actual returned attribute items under the categories

Entity domain

- The various description to characterize companies

1) Geography: Continent/Country/State/City
2) Industry subtype: SIC classification
3) Status: Public/Private, Major/Large/Small
4) Top n: Top 100/500/1000

Quantity

- The number of companies covered, with respect to the entity domain.

Usage Fee

- Charged fee for accessing the data source

1) *Free:* No registration or fee is required
2) *Free Registration:* Registration is required, but no fee for access
3) *Charged registration:* Registration is required, and certain fee is charged for access

3.5.2 Example

Table 1 is an example with SEC Edgar as an entry. The entire list of data sources we have gathered are included in Appendix A.

Table 1: An example description of one of our data sources - SEC Edgar

Data Source Name	SEC EDGAR Archive
Original source	SEC filings
URL	http://www.sec.gov/cgi-bin/srch-edgar
Query Interface	unrestricted domain- one column search
Query input	company name or any keywords
Query Result	full-text
Attribute Domain (Category)	ownership info, directory, background, financial
Attributes Domain (Items)	ownership info: subsidiaries directory: address, phone number, executives; background: company statement; financial: financial statements
Entity Domain	publicly held companies
Quantity:	almost complete for US public
Usage Fee	free

3.6 Logic Flow of Research Exercise

3.6.1 Conducting a research on Affymax

In picking a company to conduct the research exercise on, we avoided companies that we had prior knowledge of. This would ensure a low level of assumption on our part. From a list of pharmaceutical and biotech companies, we tried the first few not-so-well-known companies. Affymax turned out to be quite an intriguing example.

Much of the information about Affymax required inference due to the lack of structured information available on the company. The following log records the flow of our research exercise and the lessons that we learned along the way. It will also explain why information on Affymax is limited.

1) *SEC Edgar page*

We looked at SEC Edgar, as it is an extensive source for publicly traded company in US. However, we found nothing.

What have we learned?	How do we know?
We suspect Affymax to be either not publicly traded or it is non-US	• It is not found in SEC Edgar. • A company not found in a site implies the company is probably outside the site's entity domain. In this case, the entity domain of SEC Edgar is *US publicly traded companies.* • We actually learned something even though we did not find the company.

2) *Yahoo!*

We decided to look at Yahoo!, as it is a resource of many public and non-public companies. However, there is not an entry of Affymax in Yahoo!. Yahoo! actually used Alta Vista to conduct the search and returned a list of pages which contain the word *Affymax.*

3) *The Glaxo Wellcome page*

The URL of the first article listed in the Yahoo! / Alta Vista search result, http://www.glaxowellcome.co.uk/world/affymax/affymax3.html, hints an affiliation between Affymax and Glaxo Wellcome (a UK company). On the Glaxo Wellcome page, we found these paragraphs (*please note there are two different companies – Affymax, the one we are interested in, and Affymetrix*):

> *To this end, 1993 saw the launch of Affymetrix, a company that is developing the chip technology originally created for drug discovery to produce diagnostic tests to screen for genetic diseases such as cystic fibrosis. Owning Affymax has also allowed Glaxo Wellcome to establish a physical presence in the high tech environment of the west coast of the USA...*

> *Dr Ringold: "We can serve as Glaxo Wellcome's eyes and ears to California's academic and biotech community and provide the company with a technology magnet that attracts people with interesting ideas to us."*

What have we learned?	How do we know?
Glaxo Wellcome is probably the parent company of Affymax	• From the hierarchy of the URL • From the phrase "Owning Affymax has also allowed Glaxo Wellcome..."
There is an unclear relationship between Affymax and Affymetrix.	• Inference from the content of the page
Affymetrix was launched in 1993, it develops chip technology for diagnostic tests.	• Directly stated in the page (however summarizing it requires some knowledge of the language)
Affymax – biotech related, in California	• Inferred from the content

4) *Affymax Fact File in Glaxo Wellcome's site*

At the end of the page, there is a link to *Affymax fact file.*

> *In December 1991 the first public offering of Affymax raised $92 million, the third largest in biotechnology. This allowed rapid growth in 1992 from 65 employees to 200.*
>
> *In 1993, Affymetrix Inc, a subsidiary of Affymax, was created. Glaxo Wellcome now owns about 48 per cent of the equity of Affymetrix.*
>
> In March 1995, Affymax was acquired by Glaxo, now Glaxo Wellcome. Structurally, within Glaxo Wellcome, Affymax reports to Dr Barry Ross, director of Research Strategy and Alliances. Functionally, it reports to all Glaxo Wellcome research centres.
>
> Affymax currently employs 215 people. Of these, 80 are biologists, 66 chemists and 13 engineers, with the remainder being support staff. Affymax has two sites - Palo Alto and, not far away, Santa Clara.

What have we learned?	How do we know?
Affymax went on public in Dec 1991	• Directly stated in the page
Affymax was public traded, but not anymore	• Inference from the content of the page
Affymetrix Inc. is a subsidiary of Affymax (GW owns 48% of Affymetrix)	• Directly stated in the page
Affymax was acquired by GW in Mar 1995	• Directly stated in the page
# employees = 215: 80 biologists, 66 chemists, 13 engineers	• Directly stated in the page
2 sites: Palo Alto, Santa Clara	• Directly stated in the page

5) *Revisiting SEC Edgar*

We tried to find Glaxo Wellcome on SEC Edgar. However, nothing was found in SEC Edgar, probably because Glaxo Wellcome is a UK company. However, the search returned a 13G report of Affymetrix. We noted an unspecified relation from an earlier page and thus we decided to look into the report.

To find relevant information on Affymax, we searched for the word "Affymax" in the 13G report. The word appeared under the section *Name of Reporting Person.* In that section, the addresses of Glaxo Wellcome plc., Glaxo Group Ltd., Glaxo Venture Ltd., Affymax N.V., Affymax Technologies N.V., Mr. Douglas Hurt, and Dr. Barry Ross were listed.

Under Item 7 "DISCLOSURE REGARDING SUBSIDIARIES", we found the whole chain of parent / subsidiary companies.

> *Glaxo Wellcome plc ("Glaxo Wellcome") is the ultimate parent holding company with respect to all of the other Glaxo Reporting Persons, other than Mr. Douglas Hurt and Dr. Barry Ross. Glaxo Wellcome owns, directly and indirectly, 100% of Glaxo Group Ltd. ("Glaxo Group"). Glaxo Group in turn owns 100% of Glaxo Venture Ltd. ("Glaxo Venture") which owns approximately 99% of Affymax N.V. ("Affymax") of which Affymax Technologies N.V. ("Affymax Technologies") is a wholly-owned subsidiary. Neither Glaxo Wellcome, Glaxo Group nor Glaxo Venture directly holds any shares of Affymetrix Common Stock.*

What have we learned?	How do we know?
Glaxo Wellcome plc -> owns 100% Glaxo Group Ltd -> owns 100% Glaxo Venture Ltd. -> owns 99% Affymax N.V. -> owns 100% of Affymax Technologies	• Directly stated in the page (under "*Disclosure regarding subsidiaries*")
Affymax Technologies owns 6,746,592 shares of Affymetrix directly. Affymax owns 958,475 shares directly.	• Directly stated in the page (under "*Disclosure regarding subsidiaries*")
Addresses of the above companies	• Directly stated in the page (under "*Disclosure regarding address of principal business office*")
Director of Affymax is Dr. Barry Ross	• Directly stated in the page (under "*Disclosure regarding subsidiaries*")

6) *Affymax Research Institute homepage*

We found nothing else on Edgar. We went back to Yahoo! and found a homepage of Affymax. The URL to the homepage was

http://www.affymax.com/ari.home.rev091297.html

However, this was only a particular version (the 09/12/97 revision) of the homepage!

We tried http://www.affymax.com/ instead and found the *real* home page.

We found a mission statement of the company.

> *To create the leading center for invention, attraction and implementation of technologies for drug discovery with responsibility to transfer the best tools available throughout the Glaxo Wellcome research organization.*

There were some other similar data in "The Company" page to those we had found.

What have we learned?	How do we know?
About the research process: • Should have tried www.*company*.com first! • Try the base location of a URL returned by a search engine may be useful	
About the company: • Mission, management committee members, patents and publications, research technologies, etc.	• Directly stated in the site.

7) *Network Science's BioTech Yellow Pages*

This is another link returned by Yahoo!. We found a one-page entry of Affymax.. We identified Affymax's Address, Phone #, Fax #, URL, Status, Officers, and Description. It is stated that Affymax is a subsidiary of Glaxo Wellcome (but not the complete chain of ownership relationship) in the Status section.

What have we learned?	How do we know?
About the research process: • These directory sites are useful to get general information of a company. • We should not go down the Yahoo! result page one by one; we should have looked at the results and evaluate which one is more relevant first.	

8) *Pharmaceutical data sources that we gathered*

There is little information on Affymax in the sites we gathered in the Pharmaceutical Data Source page. The result is a little disappointing. But Affymax has more a research / academic nature to its work; most of the sites we gathered are more commercial oriented. This is an area that we can develop on.

9) We found some useful information from the Network Science's page. We found out more information about Affymax in turns of its products / competitors.

> *Molecular Simulations Inc. is the leading provider of molecular modeling and simulation software for both life and materials science research.*
>
> *Pangea was first to market with a software architecture to integrate the drug discovery process... In April 1996, Pangea launched GeneWorld(TM), a flexible, open, and automated workflow architecture for gene discovery and annotation... Founded in 1991..., Pangea Systems began as a consultancy organization providing database integration and analysis solutions for* ***Incyte Pharmaceuticals****,* ***Affymax****, and others.*

What have we learned?	How do we know?
Incyte Pharmaceuticals can be a competitor of Affymax, it also uses a drug discovery software architecture called Pangea.	• Inferred from the paragraph

10) However, we have problems in finding the answers to:

Have there been any officer changes of a company?

Benchmarking figures for performance (not necessarily financial measures) analysis?

3.6.2 Conducting a research on Affymetrix

Since Affymax is a non-public company, information is very limited. In order to study a more common case of public companies, we decided to do a similar exercise on Affymetrix, the public traded subsidiary of Affymax. There should be more information on Affymetrix, and the information should be more straight forward. Thus, we will focus less on the inference of the company information, but instead we will focus on what we would learn on the research process.

1) SEC Edgar

From our list of General Company Sources, we see SEC Edgar is an extensive source for publicly traded company in US. We found some documents filed by Affymetrix.

In the 10-Q form's *Overview*, we found:

> *OVERVIEW Affymetrix has developed and intends to establish its* ***GeneChip****-Registered Trademark- system as the platform of choice* ***for acquiring, analyzing and managing complex genetic information in order to improve the diagnosis, monitoring and treatment of disease****. The Company's GeneChip system consists of disposable DNA probe arrays containing gene sequences on a chip, reagents for use with the probe arrays, a scanner and other instruments to process the probe arrays, and software to analyze and manage genetic information.*
>
> *The business and operations of the Company were commenced in 1991 by* ***Affymax*** *N.V. ("Affymax") and were initially conducted within Affymax. In March 1992, the Company was incorporated as a California corporation and wholly owned subsidiary of Affymax.*

> *Beginning in September 1993, the Company issued equity securities which diluted Affymax' ownership in Affymetrix. In March 1995, Glaxo plc, now Glaxo Wellcome plc ("Glaxo"), acquired Affymax, including its ownership interest in Affymetrix. As of September 30, 1997* ***Glaxo owned approximately 33% of Affymetrix****.*

Lessons:

10Q's Overview in Edgar SEC is good for getting a General Description and maybe the Ownership status of the company

2) Yahoo! Finance

We decided to look at Yahoo! Finance since Affymetrix is a public traded company for some third party summary of the company. The ticker symbol is AFFX. We found general financial info about AFFX. In *Profile*, we found the

- Address
- Phone
- Fax
- Industry
- Sector
- Number of Employees
- Officers
- Business Summary
- and other Financial Data

> *Affymetrix, Inc. is focused on developing* ***"GeneChip"*** *based products and related technology for the* ***acquisition, analysis and management of complex genetic data.*** *For the nine months ended 9/97, revenues rose 97% to $12.6 million. Net loss increased 50% to $15.4 million. Results reflect increased funding from the Advanced Technology Program and NIH. Higher losses reflect the hiring of additional research and development personnel and associated purchases of research supplies.*
>
> In the *Industry* category, there are other companies (around 400) in the *Biotechnology & Drugs* industry.

Lessons:

Yahoo! Finance is a good source for US publicly traded companies. After looking up a company's ticker symbol, we can obtain their classification of the company's industry and sector. There are also the number of employees and business summary,

3) Affymetrix's company website.

We found complete product list in these categories:

1. GeneChip Instrumentation Systems;

2. GeneChip Information Systems;

3. GeneChip Service;

4. GeneChip Expression Analysis Probe Arrays;

5. GeneChip Assays;

6. GeneChip Controls.

Try to seek management changes and strategic partners / competitors from the *Company News* section:

- Affymetrix and Oncormed to co-develop and commercialize gene expression database
- Affymetrix and BioMérieux Expand Collaboration to Include HIV Genotyping and Microbial Contamination Testing
- Affymetrix Files Patent Infringement Suit Against Synteni and Incyte
- Affymetrix and Metabolex Sign Agreement for Supply of Gene Expression Monitoring Arrays
- Affymetrix and Novartis Sign Agreement for Supply of Gene Expression Monitoring Arrays
- Affymetrix and Amersham Pharmacia Biotech Enter into Distribution Agreement for GeneChip Products
- Affymetrix and Molecular Dynamics Enter Agreements to Expand Access to DNA Array-Based Genetic Analysis Tools
- Dr. Alejandro Zaffaroni to Retire from Affymetrix and Alza Boards

Lessons:

- The company website is a good, accurate source for administrative data (address, phone number, officers)
- The company website is probably a good source for its product line.
- The news section can help in finding out the company's strategy, its alliances / competitors and maybe officer changes.

4) Tradeport California

We found the Product SIC Code to be 8731.

5) SIC Manual Online

We found that 8731 stands for Commercial Physical and Biological Research.

6) Sales Lead USA

We tried to search for other companies in the 8731 category. We found sub-categories within 8731. The relevant sub-categories are:

SIC code	Industry	# companies
8731-08	PHARMACEUTICAL RESEARCH LABORATORIES	198
8731-24	LABORATORIES-BIOCHEMICAL	22
8731-26	BIOTECHNOLOGY PRODUCTS & SERVICES	223

We obtained a preview list of the companies in these three categories. These can be considered to be affymetrix's competitors. We could have further filter the list with size, location, etc.

7) For benchmarking figures, we have decided to compare Affymetrix's P/E with the industry average. Having found affymetrix's competitors, we can find their P/E's to calculate the industry average. (See Section 5.3 Problem Decomposition)

8) Questions left to answer:

What parts of the world do these products get sold to?

Who are the competitors of this company in terms of other competitive relations?

Have there been any officer changes of a company?

These questions are difficult to answer, since few sites have these answers compiled. The best hope is to gather some information from articles.

9) We attempted to find some more information from the list of sites on our pharmaceutical data sources.

PhRMA - Genomics Today - returned two yahoo articles on Affymetrix

Network Science - returned some industry news, which we had seen before. It also returned a form on Second International Lake Tahoe Symposium on molecular Diversity. Under the *Drug Screening and Functional Genomics*, we found Affymetrix's name, and also some possible **competitors**.

NIH Web Search – we found some new articles on Affymetrix.

3.7 Observations

We found that we can categorize what we learned in the research exercise into two main parts: 1) Problem Solving Methodology, and 2) Knowledge. Let us briefly summarize what we learned in this research exercise. We will discuss the issues in depth in the chapters 4-6. From the research exercise, we see there are different knowledge modules that system need to use in order to solve a problem: 1) Data source, 2) Domain / Ontology, 3) Accounting / Financial Report, 4) Statistics, and 5) Keyword.

3.7.1 Problem Solving Methodology

The question on Problem Solving Methodology is how to use the different knowledge modules and utilize users' prior knowledge to solve a problem. Here are some observations we made regarding Problem Solving Methodology.

- Some questions are difficult to answer because they ask for qualitative information that is hard to clearly define. (See Section 4.1.1 Keyword Question vs. Essay Questions) For example,

  1. *Who are the competitors of this company in terms of competitive relations?*

- Some questions are difficult to answer because they ask for information that is rarely compiled by data sources. (See Section 4.1.2 Asking for something common / uncommon.) For example:

1. *What parts of the world do these products get sold to?*
2. *Have there been any officer changes of a company?*

- There are different types of questions and we may need to capture knowledge about *questions*. We can develop different strategies to deal with different kind of questions. (See Section 4.1 Categorization of Questions.)
- When questions are difficult to answer, or no data are readily compiled, we may rely on articles or journals to infer the answer. However, this requires understanding of the language. IIAA could refer the page to the user and ask the user to supply the system with more knowledge afterwards. To identify relevant articles and even paragraphs, we can use keyword triggering. For examples, when seeking for subsidiary relationships, we can search on words like: "own", "subsidiary", "parent", etc. Knowledge on what words are important for which questions would be useful. (See Section 6.3 Keyword Knowledge.)
- When the company is listed with other companies in an article, those other companies are probably its competitors.

3.7.2 Knowledge

Knowledge on Data Sources

General Observations of Sites

- Finance data sources generally provide data of higher quality, but mostly cover publicly traded company only.
- Industry-specific (at least in the pharmaceutical / biotech sector) sites are less established.
- Company web site is generally the best site for product information.

- Some sites focused on companies that provide services to pharmaceutical industry - e.g. equipment, chemicals suppliers, and not companies that provide pharmaceutical products or services to the public.
- Some sites act as an advertising place. Companies need to pay a fee to be listed. Quality of the information may vary.

Site Characteristics

- Query Input Syntax – different sites have different query input syntax. For example, in SEC Edgar, an "OR" is assumed between separate words. A search for "Glaxo Wellcome plc" would return any entry that contains any of the three words. This search returns many irrelevant entries as many entries contain "plc". (See Section 6.1.3 Issues and suggestion on the current source description schemePage – Query Syntax.)
- Difficult to categorize into "different" industries; sites often have overlapping domain. For example, how do you separate sites among pharmaceutical, biotech and health care? Besides, how general can be considered general? Should a site that covers several industries considered general? Dividing sites into industry group is inappropriate. The range of companies covered should be reflected in entity domain.
- Entity domain should not be limited the domain of company covered only, but the scope of information in general. (See Section 6.1.3 Issues and suggestion on the current source description scheme: Site – Entity Domain.)
- Page or site? Which page should we reference to? The site's homepage or the search page? Should we have multiple entries for different pages? Typically, the homepage has some browsable pages and then a search the site page. Sometimes there is also a

site map. For the Fields we have right now, should we describe the page or the site? (See Section 6.1.3 Issues and suggestion on the current source description scheme: Site vs. Page.)

Domain Knowledge

Company Information Ontology

Address can be thought of a piece of directory information. A site containing directory information probably contains address information. This is why an ontology of company information may be useful. (See Section 6.2.1 Ontology of Company Information.)

Industry Ontology

How do we categorize the industries? Should we use the existing industry ontology scheme such as the SIC code, Yellow Pages classification? (See Section 6.2.2 Industry Ontology.)

Internet Domain Name Ontology

As in the example, we inferred that Glaxo Wellcome is a UK company from its URL. (See Section 6.2.4 Internet Domain Name Ontology.)

Keyword Knowledge

As mentioned previously, keyword can be used to identify relevant paragraphs. For example, for subsidiary information, useful keywords would be: "own", "parent", "subsidiary". And for merger information, useful keywords would be: "acquire", "merge", "spin off", "go public", etc. (See Section 6.3 Keyword Knowledge.)

Financial Report / Accounting Knowledge

We learned that several sections are particular sections of some financial reports are most useful in finding some information. For example, in 10Q, the "Name of reporting person" and "Disclosure regarding subsidiaries" often contain ownership information. In 13G, the “Disclosure regarding subsidiaries” section also contain subsidiary information. 10Q's Overview in Edgar SEC is good for getting a General Description and maybe the Ownership status of the company. (See Section 6.4 Financial Report / Accounting Knowledge.)

Statistics Knowledge

As in the example, when we calculate the industry’s average P/E, we need to know what an average is. Other useful statistics measures may include: minimum, maximum, total number of, standard deviation, variance, etc. (See Section 6.5 Statistics Knowledge.)

4 Question Answering Process

Let us now look at the Question Answering Process in more details. There are different types of questions. In Section 4.1, we will categorize questions and decide which type of questions we should focus our effort in answering.

In Section 4.2, we discuss how to answer a definite question with a thought experiment outside the company research domain. Using a general example, instead of a specific one like the research exercise we conducted, would allow us to examine the question answering process in general. We will discuss the question answering process in terms of two aspects: 1) Problem Solving Methodology and 2) Knowledge.

4.1 Categorization of Questions

We want to identify interesting types of questions that are plausible for the system to handle. Interesting types of questions are questions that the system can feasibly answer, but would require some non-trivial ways to answer the questions.

In this section, we try to categorize questions and identify questions that are *answerable* by the system. Using this approach, we can develop category-specific strategies, as well as general methodology in answering questions.

4.1.1 Keyword Question vs. Essay Questions

Some information can often be summarized into several words or numbers – let us describe this kind of information as *keyword.* For example, a company's P/E ratio is a piece of *keyword* information. *Keyword* data are usually listed in table format. Questions asking for *keyword* information are easier to handle – the data can be easily recognized.

Some questions are more like essay questions. An example would be "What does a company do?". The answers to these questions are lengthier and one can answer these questions in many ways. To answer questions of this type, one needs to recognize relevant phrases and reorganize the information to actually answer a question.

4.1.2 Asking for something common / uncommon

Common information refers to data that are readily found in sites. These data are usually free of charge and are compiled into table format by data sources. Some other information is less commonly organized and compiled by data sources. These data may be buried inside articles or journals.

For example, a user may want to find out the parent and subsidiaries of a company. This kind of information is definite; however, they are not compiled systematically by sites. In this situation, we need to infer ownership information from, for example, the company web site and its financial reports. To extract relevant information would require the ability to understand a piece of writing and infer conclusions.

Table 2 summarizes the characteristics of definite / vague and common / uncommon questions in a 2 x 2 matrix.

Table 2: A 2 x 2 Categorization of Questions

	Keyword	***Essay***
Common	***A*** • Example: P/E ratio • can be answered in a few key words, or a list of items • readily found in sites • usually free of charges • usually compiled into table format by data sources • may come from same original • definition can be unclear	***C*** • Example: What does the company do? • essay-type question; can be answered in many ways • readily found in sites • usually free of charges • subject to human interpretation
Uncommon	***B*** • Example: Merger Dates and Partners • can be answered in a few key words, or a list of items • harder to find organized data • fees may apply • Possible source: news articles, journal, proprietary data sources	***D*** • Example: "What is the future trend of an industry?" • essay-type question; can be answered in many • harder to find organized data • fees may apply • subject to human interpretation • possible source: news articles, journals

4.1.3 Example of Questions

To examine a wider range of questions, we have added more questions. Here, we tried to categorize the questions into the four types we mentioned. This gives a sense of each type and allows us to see how we can deal with the different types of questions. The original questions in our research exercise are bolded in the following list.

Type A (keyword, common)

- **What is the incorporate/headquarters address of a company? (or any other common directory information)**
- **Who are the officers and members of the board of directors?**
- **What is the industry / sector of a company?**

- What is the latest stock quote of a company? (or any other common financial data)
- Does the company have a lower P/E ratio than the industry average?

Type B (keyword, uncommon)

- **What products does this company produce?**
- **What are the subsidiaries / parent of a given company?**
- **What parts of the world do these products get sold to?**
- **Is there any officer changes of a company?**
- **Who are the competitors of this company in terms of similar SIC code and products?**
- What college did the CEO attend?
- What companies have the CEO worked for before?
- What is the merging / acquiring / spinning off dates and partners of a company?
- What price is charged for a product?
- What securities has the company issued?
- What is the profit margin of a product?
- What is the market share of a particular product?
- What is the cash flows projection of a company?
- Is the industry a cyclic industry?
- What is the geographic concentration of the industry?
- Which cellular service provider has the largest number of customers?
- What is the revenue of the domestic long distance telecom market?
- What are the regulations on satellite communications?
- What percentage of the US telephone network employs optical fiber technology?
- Who sells aspirin?
- What are the drugs that are currently available for curing AIDS?
- Which are the top ten pharmaceutical companies in the world by revenue?

Type C (essay, common)

- What does a company do?
- What service does a company provide?
- What is a company's mission / culture? (commonly found in company web sites)

Type D (essay, uncommon)

- **Who are the competitors of this company in terms of other competitive relations?**
- What is a company's image to the consumer?
- What is a company's market strategy on a new product?
- What is a company's competitive advantage to its competitor?
- Why are there a sudden improvement / decline of the company's revenue?
- What is the limitations/ advantage of satellite communication technology?

4.1.4 Resolving Obscurity

Some questions are inherently less definite while some questions are badly phrased. A question on market size is inherently ambiguous – what is the market of a product? Should Dell Computer's market be considered as the PC market or specifically mail-order PC market? Is it merely the US market or the international market, etc.

Some questions are badly phrased. For example, "Which cellular service provider is the most popular?" is a bad question. *Popular* is a vague word; by *popular*, one can refer popularity to "having the most user accounts", "being the most well-known", "having the highest revenue", etc? It can be asked in the following ways instead: Which cellular service provider has the largest number of users / has the highest revenue / covers the largest area?

Some quantities, though vague, do have a common conception of what it is. For example, take "What does a company do?", although it is vague in that we can answer the question in thousands of ways, we often see one or two fairly standard answer. There is an indication that these vague quantities do have a generally accepted meaning.

Even some seemingly definite quantities can have different definitions. For example, there are many categorization of a company industry. There are the Yahoo! categorization, the SIC / NATCS code, the yellow pages categorization, etc. There may be more than one value for a company's industry. In this case, we need to decide whether to capture the different values, and note which categorization scheme we are using for each value we capture.

If it is a simple question of what is the company's industry, we could return all the values that the system can find. We can return the company's Yahoo categorization, SIC/ NATCS code and yellow pages categorization. (This brings us to another question that how we limit and control the output amount to the user. What is too much? When should we return all findings to the user and when should we select some to the user?) However, in a question in which the company's industry is used, but not directly asked for, for example, "Is the company's P/E ratio lower or higher than the industry average?", which industry categorization should we use? Resolving obscurity is one issue that we need to look at.

We would first reduce vague questions into definite questions. That is, we have made the vague question into one question by choosing one particular definition, or several questions by using some or all of the definitions. (How exactly this is done is outside the scope of the thesis; this is definitely a topic for research.)

4.1.5 Dealing with the different types of questions

There are a range of difficulties in dealing with the different types of questions, with Type A being the easiest and Type D being the most difficult to handle.

Type A questions require us to find the relevant data source and then probably identify a number or phrase from a structured page.

For Type B questions, we probably need to find relevant news or journal sites (or some special unstructured sites that we identified, ex. for rankings of a company). These questions require command over the language to interpret the information and infer the answer. Instead of attempting to find the answer, the system will refer the relevant pages or sections to the user by recognizing some keywords for certain questions (see Section 6.3 Keyword Knowledge). And then the user can read the section to find the answer. If necessary, she can refine her question to get other information.

For Type C and D questions, if it happens that there is an exact entry in a more-or-less structured web page, we can return the entry to the user. Otherwise, we would return relevant pages for users to read as for Type B questions.

In the following section, we will, as a first step, focus on answering Type A questions. We will focus on answering questions that require only keyword (more definite) data; and let us assume no parsing of text is required to get those information.

4.2 How does human answer a question? – A thought experiment

How do human answer a question? There are several different paradigms. When asked to write an expository essay on a question, one may brainstorm for some ideas and then organize them into logical order to present to the reader. When asked to do an old problem, one probably remember how they did it last time and repeat the same process if nothing seems wrong with the method. These methods maybe useful, but a methodology in solving a new problem is probably most useful to our research.

Let us pretend we have never seen this question before:

> "An egg is thrown off a building, what is the impact when
> it hits the ground?"

We will probably gather the relevant pieces of information. But how do we know what information is relevant? We probably find out some equations relating *impact*:

1. $Impact = \int F\, dt$ *or*
2. $Impact = m\, \Delta v$

In order to answer what the impact is, we can decompose the original question into either

- "What is the *force* when the egg hits the ground, at every moment during the impact?" or
- "What is the *mass* of the egg and the *change in velocity*?"

The second approach is obviously easier. But why do we know? How do we choose which path to take in answering the question? How should one value which path is the best? It probably depends on which piece of data is available and which piece is easiest to get if unavailable.

The *mass of egg* is not too difficult to get. For "What is the *change in velocity?*", we can decompose the question into

- "What is the final velocity (after the transfer of momentum)?" and
- "What is the initial velocity (when it hits the ground)?"

We know that the final velocity is 0 – the egg collapses. Thus, the question that remains is "What is the initial velocity when it hits the ground?" The question is equivalent to "What is the impact velocity (velocity just before the ground)?" We know that we can answer this question using the four kinematics equations:

1. $v_t = v_0 + a\, t$
2. $d = v_0 t + 1/2\, a\, t^2$
3. $v_t^2 = v_0^2 + 2\, a\, d$
4. $d = 1/2\, (v_0 + v_t)\, t$

Here, we have reduced the question to a simpler question – "What is the impact velocity?" – and we can continue solving this problem. But let us not going to belabor the point. (Probably we will use equation 1. We will find the elapsed time and measure the initial vertical velocity when it leaves the thrower's hand.)

We have already found some patterns in this example. Let us discuss these patterns in the following section.

4.2.1 On Problem Solving Methodology:

1. We use a ***backward deduction*** methodology. We start from the goal – the answer we want to get.
2. We try to connect the goal to information that are available or information that are easy to obtain. In other words, we want to minimize the *effort* in arriving at the answer. How we define *effort* is another topic that we should discuss.
3. We may need to decompose a problem into simpler problems. (*Maybe in some situations, we will need to convert a problem to a more complicated problem to reach an answer? Probably, we would not call the problem more complicated, it is probably a simpler problem with more steps.)*

4.2.2 On Knowledge – What do we need to know?

There is a large amount of implicit knowledge required in this example. We probably cannot list it all. Here is some examples of what knowledge we need to know:

1. We need to know some physics phenomena, e.g. egg is an object and object falls to the ground on Earth.
2. Some knowledge of physical entities? E.g., impact, mass, velocity, time, etc.
3. Some physics equations that relate to the question. What does the symbol stands for? e.g. misunderstanding of the various initial velocities and final velocities in the equations would yield an incorrect answer.
4. Algebraic skills.
5. What information is easy to get? What measurements are hard to make?

4.2.3 Section Summary

In this section, we have looked at an example of answering a new question that requires only *keyword* data, that no distillation of the data from the data source is required. We have recognized the important issues in the Question Answering Process. In the following chapters, we will in turn discuss:

1. Problem Solving Methodology and
2. What knowledge do we need to capture for IIAA?

5 Problem Solving Methodology

5.1 Overview

In answering a question, problem solving methodology and knowledge modules are closely coupled. How we solve a problem depends on what knowledge we have. In a trivial case, if we have the answer in our knowledge modules already, no problem solving is required. We only need to retrieve the answer from the knowledge modules. In normal cases, how we solve a problem also depends heavily on what knowledge we have. Problem solving methodology involves how to use knowledge to answer a question.

In this chapter, we will first discuss problem solving methodology, and in the next chapter we will discuss the knowledge modules. Some issues are related to both chapters, but can be better explained in the next chapter. These issues may be mentioned in this chapter, but will be discussed in detail in the next chapter.

Problem solving methodology consists of two major techniques: 1) Backward Chaining, 2) Problem Decomposition.

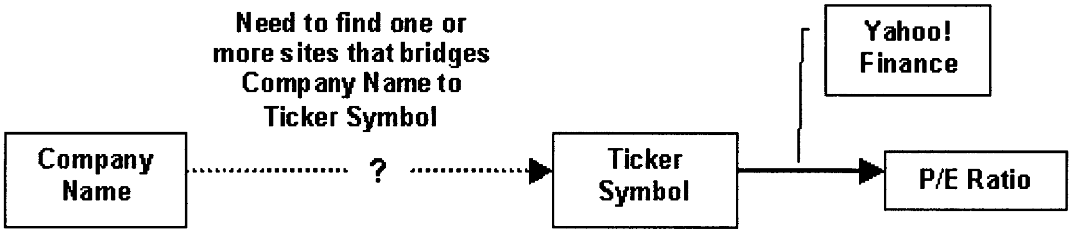
5.2 Backward Chaining

Multiple steps are often required to solve a problem. For example, for obtaining Affymetrix's P/E, we may need to get the Affymetrix's ticker symbol first, and then use the ticker symbol to get the P/E ratio.

How do we find a path to answer a question? In this simple case, it is quite obvious that we will go find the ticker symbol first and then find the P/E with the ticker symbol. But what if we have a less obvious case when the path may involves tens of steps?

5.2.1 Backward Deductive Chaining

As we have exemplified in our egg-impact problem and this P/E ratio problem, we have used a backward-chaining deductive method to answer the question. We start from the *Goal* (information we are seeking) and see how we can connect to the *Origin* (the information that is available or that are easy to obtain).



Need to find one or more sites that bridges Company Name to Ticker Symbol
Company Name
?
Ticker Symbol
Yahoo! Finance
P/E Ratio

Figure 1: Using Backward Chaining to find P/E Ratio of a company

Figure 1 depicts the process to answer the P/E question. First, we identify sites that return P/E ratio. Yahoo! Finance is one of them and it takes ticker symbol for input. Now we need to find one or more site that bridges ticker symbol to company name. In this simple case, Yahoo! Finance also serves as a link as it can take a company name and return the ticker symbol.

Of course, we did not know whether ticker symbol would be a useful immediate data. An immediate data would only be useful if the immediate data leads to a completed path to the origin.

In this model, we will view a data source as an input-output box. When given inputs, it returns some outputs. Input domain is the input that a data source takes and attribute domain is the output it returns.

The system would basically look at the data source database and identify data sources with P/E ratio in its attribute domain. Then, check if the *origin* is in these data sources' input domains. If so, we have found a complete path; otherwise, take the values in the new input domains as new *goals*. Repeat the process until the *origin* is reached.

An additional requirement is that the entity domain of the data source must cover the entity in interest. In other words, we need to know that Yahoo! Finance covers Affymetrix.

Whether we can answer a question depends we can find a path that connect the *Goal* and the *Origin*.

5.2.2 Issues that needs to be resolved

Multiple paths consideration

When multiple paths exist, how do we choose? Do we stop when we find one path? Do we find all possible paths and choose the best one? Or do we explore all the paths to get the most accurate answer? There are also performance issues.

Selecting the best path

How should one value the best path? We need to incorporate quality and cost issues in evaluating the paths. The set of possible paths depends on which data is available. The objective of choosing an optimal path should be to connect the *goal* (information that we are seeking) to information that are available or information that are easy to obtain. In other words, we want to minimize the *effort* in arriving at the answer. *Effort* should be a function of fee, time, computation power and quality.

5.3 Problem Decomposition

One assumption that we made in previous section is that we have at least one data source providing the data we are seeking. In other words, we know the data is out there, we just need to find a path to connect the data we have to the data we want. What if none of the data source contains the piece of information we are looking for, such as the industry average P/E ratio. Problem Decomposition may lead us to the answer. We have two questions regarding problem solving methodology:

- How do we know if we need to break a problem up?
- How do we break a problem into sub- problems?

5.3.1 How do we do it?

We may need to decompose some problems into simpler problems before we can solve them. For example, as in our research exercise, we wanted to know whether a company's P/E ratio is high or low relative to the industry average. We will need to break the problem up into simpler problems that we know how to solve it:

1) What is Affymetrix's P/E?
2) What is the Affymetrix's industry?
3) What are other companies in the industry?
4) What are their P/E's?
5) Finally, we need to calculate the average of the P/E's of the companies and see if company A's is higher or lower than the average.

5.3.2 How may the system do it?

Figure 2 is a graphical flow chart of how the system may solve the problem and we will discuss what the figure means in words. The question, once again, is whether Affymetrix's P/E is higher or lower than the industry average. The system needs to know that the question involves a comparison of two numbers – Affymetrix's P/E and the industry average. Therefore, the system sets out to find these two numbers. Finding

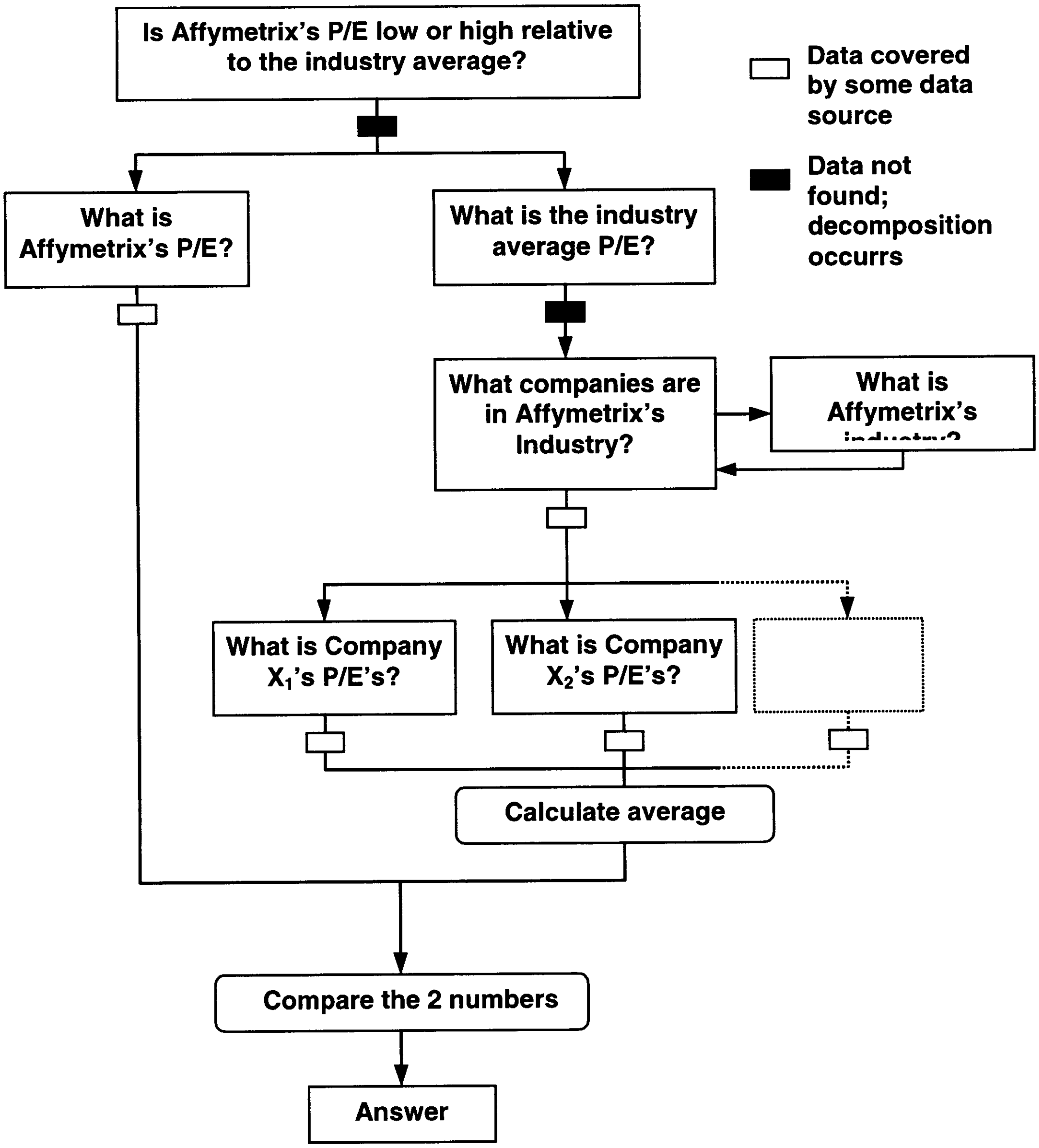

Is Affymetrix's P/E low or high relative to the industry average?
Data covered by some data source
Data not found; decomposition occurrs
What is Affymetrix's P/E?
What is the industry average P/E?
What companies are in Affymetrix's Industry?
What is Affymetrix's industry?
What is Company X1's P/E's?
What is Company X2's P/E's?
Calculate average
Compare the 2 numbers
Answer

Figure 2 Decomposing a question

Affymetrix's P/E ratio is achievable; the system can query a financial data source to get the data. (The black boxes in the figure represent that no problem decomposition is required as there exists sources that output the data. The white boxes signify that problem decomposition is required.)

The system needs to understand concepts in statistics such as average, minimum, maximum, etc. in order to decompose the question. For example, a question involving *average* can be decomposed into – summing all the members and dividing the sum by the total number of members. (see also Section 6.5 Statistics Knowledge).

In order to find the industry average, the system needs to know "What is the P/E's for each of the company in the Affymetrix's industry?" However there is an unknown in this question: *Affymetrix's industry*. The system should find *Affymetrix's industry* first and substitute its value back into the question. Gathering *Affymetrix's industry* requires no decomposition; Affymetrix is in the biotechnology industry.

Now the system needs to answer "What is the P/E's for each of the company in the biotechnology industry?" This question contains an iteration process – "For all the companies in the biotechnology industry, find their P/E's." This is how the question should be broken down. We need to find all the companies in the biotechnology industry and substitute each instance into the variable and find the P/E of the company.

After gathering all the P/E's, the system needs to perform a computational step of calculating the average and comparing it to Company Affymetrix's P/E.

5.3.3 Decomposition Techniques

Learning from the above example, we have summarized our findings in three decomposition techniques: 1) comparison or computation decomposition; 2) unknown substitution; and 3) iterative expansion.

Comparison or Computation Decomposition

The system needs to realize if a question requires a comparison or computation step of a set of data. If so, break up the main question into gathering each element in the required set of data.

In our example, the system would break “Is Affymetrix’s P/E high or low relative to the industry average?” into

1. “What is Affymetrix’s P/E?”,
2. “What is the industry average P/E?” and
3. attach a comparison step after gathering the data.

The system would also decompose “What is the industry average P/E?” into

1. “What is the P/E for each of the company in the biotechnology industry?” (assuming we have substitute *biotechnology industry* for Affymetrix’s *industry* already),
2. attach a computation step, namely calculating the average, after gathering the data.

Unknown Substitution

When there is an unknown, *X*, in a question. Find out the value of *X* first. In other words, insert the question “What is *X*?” before the original question. Then substitute the value of the unknown into the original question.

In our example, we have “What is the P/E for each of the company in Affymetrix’s industry?” The unknown is *Affymetrix’s industry*, so the system should

1. insert “What is Affymetrix’s industry?” before that question, and
2. substitute the answer with the unknown.

Iterative Expansion

Iterative Expansion is often used jointly with Comparison and Computation Decomposition. It is often used when a statistical inference is required. For example, when a user wants to find out about the average, maximum, minimum, etc. of some information.

The system would expand a question relating all the elements in a set into an iterative process. The system would ask the same question to every element in the set. Let X be a set of {X_1, X_2, ..., X_n}and the question is "What is each member's C in X", while C is a characteristic of the X_i's. Iterative Expansion would substitute the original question with a set of questions:

1. "What are the members of X?"
2. "What is X_1's C?"; "What is X_2's C?"; ... "What is X_n's C?"

In our example, we would expand "What is the P/E for each of the company in the biotechnology industry?" into an iterative process:

3. "What are the member companies in the biotechnology industry?"
4. "What is *biotech company 1*'s P/E?"; "What is *biotech company 2's* P/E?"; ... "What is *biotech company n*'s P/E?"

5.4 Coordination between Backward Chaining and Decomposition

As in our previous example, we have already used an algorithm implicitly. The basic algorithm is: 1) try solving the problem with backward chaining first, 2) if no completed path is found, try decomposing the problem using the three decomposing techniques.

6 Knowledge

6.1 Knowledge on Data Sources – Source Description

6.1.1 Introduction

Knowledge on the data sources is central to the system. As we have already illustrated in the question answering process section, how we go about answering a question depends on what we know about the available data sources. We have discussed Question answering process in Chapter 4 Question Answering Process. And we will now discuss what data source knowledge we need to capture to enable source selection in answering a question. There are many things we need to know about the data sources. The most important piece of information is obviously what data it contains. As we have attempted in the research exercise, we need to describe and categorize the different data sources.

The range of knowledge to capture depends on what we want to do with source selection. We want to capture the entity domain of the data source so that we know which data source to explore when we need some information. We want to capture the fee in accessing a data source so that we can select data sources with the fee taken into account. Every piece of knowledge we capture should serve a purpose in source selection.

6.1.2 Challenge

Accurately describing a source is challenging because most web sites are not structured. We attempt to describe a data source by first constructing a list of attributes and defining the domain of each attribute. Then for each attribute, we find the best value to categorize a data source. The wide range of format and structure in web sites make it very difficult to categorize a site with a fixed set of attributes and attribute values.

In our research exercise, we tried a first run for describing sources (See Section 3.5.1 Source Description). We found that it is difficult to categorize sites in several attributes, and that some additional attribute values can better describe web sites. There are also other issues concerning the different attributes. Some of the issues are not of great importance, but it is important that we resolve them for the system to work.

Some other issues are important but complicated. Many of them are not resolved in this thesis. The exactly solution would require further investigation into source selection and weighing of the pros and cons of different implementation. We need to concretize the system requirements before these decisions to be made. The following section lays out the issues that need to be considered with different implementations. Reading the section would not give you answers to many of the problems, but the issues that need to be considered.

6.1.3 Issues and suggestion on the current source description scheme

Site vs. Page

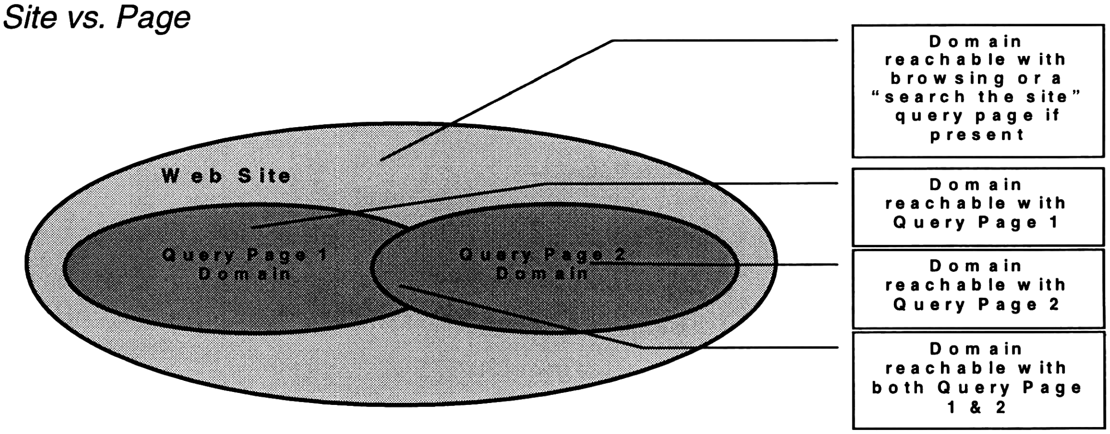

Web Site
Query Page 1 Domain
Query Page 2 Domain
Domain reachable with browsing or a "search the site" query page if present
Domain reachable with Query Page 1
Domain reachable with Query Page 2
Domain reachable with both Query Page 1 & 2

Figure 3: A graphical representation of a web site

Typically, the root page of a site has some browse-able links and also links to one or more search pages. Generally, one can reach all the content of a site by browsing. There are maybe one or more pages to search on different information, or it can take the form of restricted domain query. For example, a company web page may allow you to select a domain, out of products, company news and technology. Sometimes, typically in more established sites, there is a "search the site" page which should in theory covers the entire site. This view of a web site is captured in *Figure 3*.

Originally, we have the following attributes in the research exercise for describing a source:

- Data Source Name
- Original source
- URL
- Query Interface
- Query input
- Query Result
- Attribute Domain (Category)
- Attributes Domain (Items)
- Entity Domain
- Quantity
- Usage Fee

In the research exercise, we did not distinguish between site properties and page properties. For example, does *URL* refer to the site's homepage, or does it refer to the query page? A natural organization to would be, for each site, to include one set of *site characteristics* and one or more sets of *page characteristics* for each search / query page.

I suggest modifying the existing site record to the following format. Bolded items are new attributes. The addition and modification of attributes will be discussed.

Site Characteristics

- Data Source Name
- Original Source
- Root URL

- Registration and Cost
- Quantity
- Entity Domain

Query Page Characteristics

- Page Name
- Page URL
- Original Source
- Query Interface
- Query Input
- **Query Syntax**
- Query Result
- Attribute Domain (Category)
- Attributes Domain (Items)
- **Query Domain**

Representation

Instead of one database table of all the sources, we now should have one table for sites and another table for pages. In the page table, we should include the *parent site* attribute to indicate the Site → Page relationship.

Site – Data Source Name

The name of the data source is ambiguous. Sites have several names. There are a page title, an HTML title (the title encoded in HTML code that are shown on the browser's title bar). There is also the name of the company that provides the data. The name is probably mainly for human use and will not affect the system. However, we do need to clearly define what data source name is. The question is "Which of the names (can be more than one) do we want to capture?"

Site – Original Source:

We identified several issues associated with this field after conducting the research exercise.

The original data source may not be obtainable in many cases. Some sites do not state where they obtain their information. Sometimes, the information is just not probably credited. Sometimes, maybe the data is original.

It is difficult to classify who original own the data in many web sites. Maybe the data is aggregated from a number of different places. For example, Four11 compiles its data from various sources and add some data that they gather themselves (e.g., via user submission). It is difficult to classify the originality of the information.

There is an issue of defining what is original. If there are some written material that the online version originates from, that written material qualifies to be called an original source of the web content. However, nowadays, many information are presented both online and in paper version. The paper version cannot be described as the original source; it is merely a different presentation format of the same content.

I suggest that this field to be made optional. The original source should only be stated if there is a clear indication that the web obtains most of its content from the source. Otherwise, we can include the sources as partial sources.

Site – Root URL

This is now specifically the URL of the site's root page.

Site - Usage Fee → Registration and Cost

In our research exercise, we had:

1) Free: No registration or fee is required
2) Free Registration: Registration is required, but no fee for access
3) Charged registration: Registration is required, and certain fee is charged for access

I suggest separating Usage Fee to Registration and Cost, as the two attributes are independent of each other. Also, on the observation that many free sites do provide premium service for a fee, I suggest incorporating this kind of charging structure into our description.

1) Registration: either required or not required.
2) Cost:
   a) Free / Charged
   b) Premium Service available for a fee
3) Charging Structure
   a) per usage / per article / flat rate (monthly, yearly charge).

Site – Quantity

In our research exercise, the number is often difficult to find or even estimate. In many cases, we cannot find a documented number nor it was possible that we can see a complete list of companies / people covered.

Site – Entity Domain

Site Entity Domain is now specified as the domain that is covered by the entire site, not merely the searchable domain. Site Entity Domain is represented by the large oval in *Figure 3*.

Page – Query Page Name

There may not be a Query Page Title, because the query input box may be embedded in, for example, the root page. Should we even include such a title? And as we have discussed in the *Site – Data Source Name* section, should we capture the title on the page, or the HTML title?

Page – Query Page URL

Query Page URL is now specifically the URL of the query page.

Page – Query Interface

Now we have the following Query Interface:

1) Restricted Domain
2) Unrestricted Domain
3) Full text search
4) Browsing

I suggest adding Script Session to the list.

5) Script Session: Users are presented with choices (in the form of list boxes, radio buttons, check boxes, etc.) After submitting their selection, users proceed to a next list of choices depending on their selection. Generally sites use this interface to filter out companies that fit the criteria progressively.

   Example: SalesLead USA uses this interface to filter out the company of interest. If you would like to find companies in a particular industry, you check Type of Business, instead of All Business. Then you are presented with choices of enter type of business by 1) Yellow Pages Headings, 2) Major Industry Groups and 3) SIC codes. The selection process afterwards is essentially the same.

Page – Query Syntax

I suggest adding this attribute. Syntax varies in different query pages and understanding of the query syntax allow us to obtain relevant and useful results. For example, in SEC Edgar, an "OR" is assumed between separate words. A search for "Glaxo Wellcome plc" would return any entry that contains any of the three words. This search returns many irrelevant entries as many entries contain "plc".

Page – Query Domain

Page Query Domain is now specifically the domain covered by the query. Page Query Domain is represented as the small circles in *Figure 3.*

6.2 Ontology

Ontology contains a set of entity, and their relations. Ontology is useful because it provides hierarchy information of the world, which can be used for subsumption checking. For example, if we are researching a computer-software company, we should consult data source covering not only computer-software industry, but also the computer industry. In other words, we should also consult data sources that cover a super-concept of the computer-software industry.

6.2.1 Ontology of Company Information

We have reviewed a wide range of questions. From the list, we found that the questions are usually after a common set of information. These company data can be grouped into categories. Note that these company data are central to our system – all questions are after them, all ontology we discuss later stem from them, and all entity domain are described with them.

Some of the items listed below are well covered by the sites we have, for example directory information, and financials of public companies. Some domains are less well-covered; inference from articles may be required.

- Company
  - Industry
    - SIC / NAISC code
    - Yellow pages classification
    - Trend and development
    - Characteristics – e.g. cyclical, macroeconomic-dependent
  - Directory Info

    - Name
    - Address
    - Phone number
    - Fax number
    - URL of company web page, etc.
- Ownership
    - Public / Private
    - Owners
    - Parent
    - Subsidiaries
- Officers
    - Directory Info
        - Name
        - Address (Home)
        - Phone #, etc.
    - Employment
        - Position
        - Salary / compensation
    - Biography
        - Work history
        - Education background
- Financials
    - P/E
    - revenues
    - Cash flows
    - Earning forecasts
    - Stocking trading
        - High
        - Low
        - Volume
        - Average
        - Sec Exchange
    - etc.
- Mission
    - What does the company do?
    - Goal / mission statement
- Product
    - Market
        - Target customers
        - Target Location
        - Strategy
        - Demand
        - Competition (link to companies with close *industry* and *product*)
    - Raw Materials

        - Suppliers
        - Cost
    - Technology
    - Legal - Regulations
- History
    - Officer changes
    - Mergers / spinoffs
- Brand Equity
    - Image
    - Brand recognition by consumers
    - Industry influence

Please note the object-orientedness of this organization. For example, Officer and Owners are of type *People*, it should inherit the *People* characteristics:

- Directory Info
    - Name
    - Address (Home)
    - Phone #, etc.
- Employment
    - Position
    - Salary / compensation
- Biography
    - Work history
    - Education background

And the Directory Info, whether for *People* or for the *Company*, should contain the same items:

- Name
- Address
- Phone number
- Fax number
- email
- URL of homepage
- etc.

6.2.2 Industry Ontology

One field from the above company data ontology is *industry*. There are many existing industry ontology schemes. For example, the yellow pages categorization, the

Yahoo industry hierarchy, SIC (Standard Industrial Classification) and NAICS (North America Industry Classification System)

Yahoo!

Most sites in Yahoo! are suggested by users. Sites are placed in categories by Yahoo! Surfers, who visit and evaluate all suggestions and decide where they best belong. Yahoo! claims that this is done to ensure that the directory is organized in the best possible way, making the directory easy to use, intuitive, helpful, and fair to everyone.

Some Yahoo! headings are listed in multiple places within the Yahoo! hierarchy. These headings are symbolized with an "@" sign. However, these multiply-listed headings do have a primary location. Clicking on the heading will take the user to the primary location in the hierarchy for that heading.

Discussions

This means that the Yahoo! industry ontology is not strictly tree-like. There are some occasional cross-links between nodes. This can be one concern in deciding which ontology to use.

Example

To get a general idea and feel of the Yahoo! ontology, let us look at an example. At the base level of the ontology, *Business and Economy* is what concerns this project. Let us look at the *Computer* category within *Business and Economy*. Yahoo!'s classification under Business and Economy is included in Appendix C Yahoo! Classification.

- Accessories (227)
- Books@
- Classifieds (51)
- Consulting (2919)

- Employment (867)
- Furniture@
- Hardware (5074)
- Internet (43)
- Magazines@
- Media (60)
- Multimedia (470)
- Networking (3116)
- Newsletters (10)
- Product Reviews@
- Retailers (2503)
- Security (198)
- Services (1391)
- Software (16283)
- Trade Magazines (6)

And under *Software*, there are these sub-categories:

- Artificial Intelligence (115)
- Arts and Crafts@
- Books@
- Business (1778)
- CAD/CAM (621)
- CD-ROM (483)
- Character Recognition (OCR/ICR) (49)
- Children@
- Classifieds (1)
- Communications and Networking (1823)
- Consulting (665)
- Custom Programming (648)
- Data Conversion (46)
- Databases (707)
- Desktop Publishing@
- Educational@
- Employment Listings (3)
- Emulation (22)
- Entertainment (189)
- Financial@
- Games (812)
- Genealogy (26)
- Government (77)
- Graphics (711)
- Health@
- Industry Specific (222)
- Internet (1104)
- Law (178)
- Licensing (23)
- Localization (153)
- Mapping@
- Multimedia (1189)
- Navigation (20)
- Operating Systems (499)
- Organizations (24)
- Paranormal Phenomena@
- PDA Software (47)
- Personal Information Management (88)
- Programming Tools (1116)
- Recreation and Sports (311)
- Regional (20)
- Retailers (248)
- Scientific (777)
- Shareware@
- Society and Culture (159)
- Surveys and Polling (31)
- System Utilities (747)
- Text Retrieval (50)
- Training (203)
- Virtual Reality (109)
- Voice Recognition (120)
- World Wide Web@
- Writing (32)

SIC (Standard Industrial Classification)

The Standard Industrial Classification (SIC) is a 4-digit code originally developed by the Office of Management and Budget to facilitate statistical economic analysis and reporting of the state of the U.S. economy based on enterprises engaged in production, trade, and service. Last published in 1987, the SIC is going through a major revision for the 1997 publication. (See NAICS in the following section.)

NAICS (North America Industry Classification System)

The 1997 North American Industry Classification System (NAICS), will replace the 1987 SIC system. NAICS '97 will become effective in the United States on January 1, 1997. The definitive US NAICS Manual will be available June 1, 1998.

NAICS is being developed in collaboration with Statistics Canada; Mexico's Instituto Nacional de Estadistica, Geografia e Information; and the U.S. Office of Management and Budget, Economic Classification Policy Committee. The NAICS system provides common industry definitions that cover the economies of the three North American countries and will aid in statistical economic analyses for all three countries.

Features

Important Features of the North American Industrial Classification System:

- Groups establishments with similar production processes
- Follows the production oriented economic concept
- Is relatively compatible at the 2-digit level with the International Standard Industrial Classification of All Economic Activities (ISIC rev. 3)

NAICS is organized in a hierarchical structure much like the existing SIC in that it groups establishments with similar production processes.

- The first two digits designates a major *Economic Sector* [formerly Division] such as Agriculture or Manufacturing.
- The third digit designates an *Economic Subsector* [formerly Major Group] such as Crop Production or Apparel Manufacturing.
- The fourth digit designates an *Industry Group*, such as Grain and Oil Seed Farming or Fiber, Yarn and Thread Mills.

- The fifth digit designates the *NAICS Industry* such as Wheat Farming or Broadwoven Fabric Mills.
- Optionally, each country may add additional detailed industries below the 5-digit level so long as the additional detail aggregates to a 5-digit level of NAICS.

Example

We searched for possible industries relevant of the computer software industry.

334611 – Software Reproducing (334 Computer and Electronic Product Manufacturing)

42143 – Computer and Computer Peripheral Equipment and Software Wholesalers (421 Wholesale Trade, Durable Goods)

51121 – Software Publishers (511 Publishing Industries)

Discussions

From the example, it looks as though 51121 (software publishers) is what we would group, for example, Microsoft under. However, NATCS categorizes the software publisher under the Publisher *Economic Subsector*. NATCS categorization bases on similar *production process*. This may or may not be what the system needs.

Reference

For more information, please refer to the NTIS homepage at http://www.ntis.gov/business/sic.htm.

Suggestion

We can take advantage of different industry ontology schemes. The difference should be noted and different schemes can be used for different situations.

6.2.3 Geography Ontology

Knowledge on geography can help in that the system would know a Massachusetts company will not be covered by Tradeport Californian Company.

6.2.4 Internet Domain Name Ontology

In our research exercise, we have inferred that Glaxo Wellcome is a UK company from its URL. The system can make more inference if it can make use of the Internet domain name ontology.

6.3 Keyword Knowledge

As mentioned previously, keyword can be used to identify relevant paragraphs. For example, for subsidiary information, useful keywords would be: “own”, “parent”, “subsidiary”. And for merger information, useful keywords would be: “acquire”, “merge”, “spin off”, “go public”, etc.

6.4 Financial Report / Accounting Knowledge

We learned that several sections are particular sections of some financial reports are most useful in finding some information. For example, in 10Q, the “Name of reporting person” and “Disclosure regarding subsidiaries” often contain ownership information. In 13G, the “Disclosure regarding subsidiaries” section also contain subsidiary information. 10Q's Overview in Edgar SEC is good for getting a General Description and maybe the Ownership status of the company.

6.5 Statistics Knowledge

As in the example, when we calculate the industry's average P/E, we need to know what an average is. Other useful statistics measures may include: minimum, maximum, total number of, standard deviation, variance, etc.

7 Open Issues

This thesis has explored a large number of issues that are important in automating the research process. We have provided some suggestions to some of the issues; however, there are still many issues that are left unresolved. In this section, we will summarize and list these issues again.

Question Answering Process

- How do we resolve obscure questions to more definite questions.
- Can we allow user to control the completeness, and the verbosity of the output in terms of length, extension of different definitions of a term?
- How much inference intelligence are we aiming for? How much inference can the system do?
- How do we define effort when evaluating different paths? We can define *effort* as a function of fee, user's time, computation power and network usage. There are also the issue of tradeoff between accuracy and effort.

What knowledge do we need to capture?

- What should we do when *quantity* and *original source* is not available.
- How do we define *original source*?
- How do we deal with the different industry categorizations?

Data Attribution

- How do we evaluate the qualify of a piece of information?
- How do we estimate of the accuracy of the answer?

8 Conclusion

Through the study of two research exercise and a thought experiment, we have discussed the issues involved in automating the company research process. We have focused on issues in *Company Research, Question Answering Process* and *Knowledge Capturing*. This thesis has brought about a number of interesting research topics in facilitating the aggregation of information from the Internet. With the development of the system, research on the Internet will no longer be a long a tedious process. Users will be able to access the exact information they need in a short time with a few keystrokes.

9 Appendix

Appendix A Data sources gathered in exercise

Data sources were collected by Tom Lee, Benny Suen and Steve Tu.

Appendix A-1 General Data Source

Data Source Name	**SEC EDGAR Archive**
Original source	SEC filings
URL	http://www.sec.gov/cgi-bin/srch-edgar
Query Interface	unrestricted domain- one column search
Query input	company name or any keywords
Query Result	full-text
Attribute Domain (Category)	subsidiary, directory, background, financial
Attributes Domain (Items)	subsidiaries (int. & domestic), address, executives, financial statements
Entity Domain	publicly held
Quantity	almost complete for US public
Usage Fee	free

Data Source Name	**Yahoo Company Search**
Original source	Yahoo
URL	http://www.yahoo.com/Business/Companies/
Query Interface	Unrestricted domain
Query input	any keyword
Query Result	full-text
Attribute Domain (Category)	web page indexing and URL addresses
Attributes Domain (Items)	N/A
Entity Domain	US and international
Quantity	
Usage Fee	free

Data Source Name	**Yahoo! Finance (Yahoo! Finance Europe & Japan are also available)**
Original source	Yahoo
URL	http://quote.yahoo.com

Query Interface	Unrestricted domain
Query input	Ticker Symbol
Query Result	full-text
Attribute Domain (Category)	web page indexing and URL addresses
Attributes Domain (Items)	NA
Entity Domain	US
Quantity	
Usage Fee	free

Data Source Name	**Tradeport Foreign Company**
Original source	U.S. Dep. of Comm, user-added
URL	http://www.tradeport.org/ts/foreign/
Query Interface	restricted domain- three column search
Query input	name, product/service, country: and/or combination
Query Result	structured text
Attribute Domain (Category)	directory, background
Attributes Domain (Items)	product/service, inc. address, phone
Entity Domain	25 sparse major trading countries
Quantity	ok
Usage Fee	free

Data Source Name	**Tradeport Californian Company**
Original source	Database Publishing Company
URL	http://www.tradeport.org/cgi-bin/banner.pl/ts/companies/index.html
Query Interface	restricted domain- four column search
Query input	name, product/service, SIC code, CA county: and/or combination
Query Result	structured text
Attribute Domain (Category)	directory, background
Attributes Domain (Items)	product/service, inc. SIC code, address, phone
Entity Domain	California
Quantity	broad
Usage Fee	Free

Data Source Name	**SIC Manual Online**
Original source	US Comm. Dep. NTIS

URL	http://www.osha.gov/oshstats/sicser.html
Query Interface	unrestricted domain, browsing
Query input	SIC code, industry
Query Result	full-text
Attribute Domain (Category)	background
Attributes Domain (Items)	industry type, specific industry entry
Entity Domain	complete industrial classification
Quantity	xxxx range
Usage Fee	free

Data Source Name	**Thomas Register Search**
Original source	Thomas Register Online
URL	http://www2.thomasregister.com/ss/.519709034/firstsearch.cgi
Query Interface	restricted domain - three column search (exclusive)
Query input	name, product/service, brand name
Query Result	structured text
Attribute Domain (Category)	directory
Attributes Domain (Items)	inc. address, phone, clearer product description
Entity Domain	US manufacturers only
Quantity	155,000, 60,000, 124,000
Usage Fee	free registration

Data Source Name	**Companies Online**
Original source	D & B
URL	http://www.companiesonline.com/
Query Interface	restricted domain & browsing
Query input	name, city, state, industry, ticker, URL
Query Result	structured text
Attribute Domain (Category)	directory, background, financial
Attributes Domain (Items)	head quarter addr., sales, employee#, phone
Entity Domain	US public and private
Quantity	100,000
Usage Fee	free

Data Source Name	**Market Guide Investment Center**

Original source	(affiliated with Yahoo! Finance)
URL	http://www.marketguide.com/
Query Interface	Restricted Domain
Query input	Ticker Symbol, Company Name
Query Result	table
Attribute Domain (Category)	Company info
Attributes Domain (Items)	address, sector, industry, business summary, financial statistics
Entity Domain	publicly traded companies
Quantity	
Usage Fee	free for regular service, fee for detailed version

Data Source Name	**Yahoo! White Pages** Can find out about e.g. the CEO's, personal info
Original source	published white page directories and other publicly available sources
URL	http://yahoo.four11.com
Query Interface	unrestriced domain
Query input	first name, last name, city, state, coutry, current organization, domain, old e-mail address
Query Result	html tagged, formatted text
Attribute Domain (Category)	directory
Attributes Domain (Items)	email, phone number
Entity Domain	people with publicly listed phone number or email addresses
Quantity	
Usage Fee	free

Data Source Name	**Yahoo! Yellow Pages** can find out smaller, private companies in a particular location. e.g. How many Cable TV companies are located around MA 02139? (and what are their names, address?)
Original source	
URL	http://yp.yahoo.com/
Query Interface	browsing, unrestricted domain
Query input	name, type, postal code, address
Query Result	formatted text, HTML tagged
Attribute Domain (Category)	Directory

Attributes Domain (Items)	Name, address, map, phone number
Entity Domain	companies listed in yellow pages
Quantity	
Usage Fee	free

Data Source Name	**555-1212.com**
Original source	Metromail, ProCD
URL	http://555-1212.com/telephone.htm/
Query Interface	Restricted domain
Query input	Name, state, industry
Query Result	html formatted text
Attribute Domain (Category)	directory
Attributes Domain (Items)	Address, email, phone#, fax#, etc
Entity Domain	listed people and company
Quantity	
Usage Fee	free

Data Source Name	**SalesLead USA**
Original source	American Business Information
URL	http://www.salesleadsusa.com/
Query Interface	Java script session (choose sth, proceed to next list of selection...) to narrow down your search progressively
Query input	industry, geography, # employees, sales revenue
Query Result	HTML structured table
Attribute Domain (Category)	directory
Attributes Domain (Items)	Name, state
Entity Domain	public / private, US / candadian companies
Quantity	11 Million business

Usage Fee	free to preview the list, .50 for each detail report including:aa Directory info, Key Contact includes: Job Function/Title & Gender ABI Number, Location Type, Employee Size Range,Sales Volume Range, Credit Rating Description, Download Date, Primary SIC Code, Franchise/Brand Description, Secondary SIC Code #1, Secondary SIC Code #2

Data Source Name	**Search Hoover's Company Capsules**
Original source	
URL	http://www.hoovers.com/search.html#capsules
Query Interface	restricted domain
Query input	company name, ticker symbol, keyword
Query Result	HTML formatted page
Attribute Domain (Category)	Directory, Financials, news
Attributes Domain (Items)	**Top Competitors**, address, phone #, fax #, URL, business summary, key executives, sales, 1 year sales growth, employee, employee growth, net income, earnings estimates, SEC filings, news
Entity Domain	companies
Quantity	
Usage Fee	free, charges for in-depth company profiles

Data Source Name	**Search Industry Snapshots**
Original source	
URL	http://www.hoovers.com/search.html#industry
Query Interface	restricted domain
Query input	preset group of industry
Query Result	HTML text
Attribute Domain (Category)	industry info
Attributes Domain (Items)	vaires: major players, size of the industry...industry jargon, industry specific sites
Entity Domain	
Quantity	
Usage Fee	

Data Source Name	**Hoover's Online: The List of Lists** can find out "the top 100 largest cable company" "most highly paid CEO" "PC Magazine's 100 Most Influential Computer Companies (1997) "
Original source	Forbes, Fortune,
URL	http://www.hoovers.com/features/whosontop/bgrichlist.html
Query Interface	browsing
Query input	N/A
Query Result	N/A
Attribute Domain (Category)	
Attributes Domain (Items)	links
Entity Domain	links to articles on Top x largest company, richest person, **THE BIGGEST COMPANIES & RICHEST PEOPLE** •The Biggest Companies & Richest People •Biggest Companies by Country **Advertising & Marketing** •Top Agencies & Commercials •Top Brands •Top Salaries •Top Advertisers **Media** •Top Companies •Top Magazines •Top Companies on the Internet •Top Salaries of Media Company Executives (Advertising Age) **High-Tech Industry (Biggest & Best-Performing Companies)** **Entertainment & Sports** **Emerging Companies** **Social Responsibility, Corporate Governance & Workplace**
Quantity	
Usage Fee	free

Data Source Name	**BigChart** Stocks with the largest % gain
Original source	
URL	http://www.bigcharts.com/
Query Interface	browsing, restrcited domain
Query input	ticker

Query Result	html structured page
Attribute Domain (Category)	financials
Attributes Domain (Items)	stock quote, charts...
Entity Domain	public traded companies
Quantity	
Usage Fee	free

Data Source Name	**BigChart Historical Quotes** historial stockquotes
Original source	
URL	http://chart4.bigcharts.com/report?r=sp-hlookup
Query Interface	browsing, restrcited domain
Query input	ticker
Query Result	html structured page
Attribute Domain (Category)	financials
Attributes Domain (Items)	stock quote, adjustment factor (after splitting)
Entity Domain	publicly traded companies
Quantity	
Usage Fee	free

Data Source Name	**Database America - PeopleFinder**
Original source	
URL	http://www.databaseamerica.com/html/gpfind.htm
Query Interface	restricted domain
Query input	last name, first name, city, state, phone # (for reverse lookup_
Query Result	HTML formatted table
Attribute Domain (Category)	directory
Attributes Domain (Items)	Address, phone #
Entity Domain	people listed in publicly available directories
Quantity	
Usage Fee	free

Data Source Name	**Web100**
Original source	
URL	http://www.w100.com/
Query Interface	browsing
Query input	N/A
Query Result	N/A
Attribute Domain (Category)	Directories, links
Attributes Domain (Items)	Name, links to a list of subsidiaries
Entity Domain	Web100 US comapanies; can also go by industry; Global Web100
Quantity	
Usage Fee	free

Data Source Name	**Forbes Toolbox**
Original source	
URL	http://www.forbes.com/tool/html/toolbox.htm
Query Interface	browsing
Query input	N/A
Query Result	N/A
Attribute Domain (Category)	financial / market
Attributes Domain (Items)	Ranking
Entity Domain	The Forbes 500 Annual Directory of America's Leading Companies The 500 Largest Private Companies in the US 200 Best Small Companies 400 Richest People in America Forbes ASAP: Technology's Richest 100 The International 500 Corporate America's Most Powerful People 1997 World's Richest People
Quantity	
Usage Fee	free

Data Source Name	**Business Wire**

Original source	
URL	http://www.businesswire.com/search.shtml
Query Interface	unrestricted domain
Query input	concept or keyword
Query Result	HTML formatted page
Attribute Domain (Category)	URL to articles
Attributes Domain (Items)	URL's
Entity Domain	news aritcles
Quantity	
Usage Fee	free

Data Source Name	**PR Newswire**
Original source	
URL	http://www.prnewswire.com/cnoc/cnoc.html
Query Interface	restricted domain
Query input	company
Query Result	list fo URL
Attribute Domain (Category)	URL
Attributes Domain (Items)	URLs to articles
Entity Domain	news articles
Quantity	
Usage Fee	free

Data Source Name	**Walker Information**
Description	Walker Information is a global resource for ***measuring and managing stakeholder relationships***, including customers, employees and other influentials affecting business success. We use information-based products and measurement sciences to access, discover and apply new knowledge and insights critical to successful business performance.
URL	http://www.walkernet.com/
Query Interface	restricted domain
Query input	industry
Query Result	not sure

Attribute Domain (Category)	Market
Attributes Domain (Items)	Customer Satisfaction Measurement and Management; Corporate Reputation and Stakeholder Assessment; Marketing Research; Organizational Culture Assessment; Data Management; Database Marketing; and Clinical Research
Entity Domain	stakeholder relationship information
Quantity	
Usage Fee	For $79, an individual use subscription allows you to select a single industry and customer type. Based on your selection, you will receive all data available from the CSD. For $199, the annual subscription gives you unlimited access to the CSD for a 12-month period.

Appendix A-2 Computer Industry

Data Source Name	**Tennessee's High-tech database**
Original source	Tennessee Department of Economic and Community Development: self-added
URL	http://www.state.tn.us/ecd/scitech/index.htm#search
Query Interface	restricted domain- 10 column search
Query input	zip code, SIC code, county
Query Result	structured text
Attribute Domain (Category)	directory
Attributes Domain (Items)	name, address, e-mail, phone
Entity Domain	Tennessee small tech-intensive and knowledge-intensive business
Quantity	comprehensive
Usage Fee	free

Data Source Name	**The CorpFiNet**
Original source	Top 100 revenue computer companies
URL	http://www.corpfinet.com/techintro.htm
Query Interface	restricted browsing
Query input	name or revenue
Query Result	HTML table
Attribute Domain (Category)	product, sales, background
Attributes Domain (Items)	product description, sales level, home page URL
Entity Domain	top 100/500
Quantity	hundreds
Usage Fee	free

Data Source Name	**Silicon Investor Profiles**
Original source	stock exchanges
URL	http://www.techstocks.com/profiles.html
Query Interface	alphabetical browsing
Query input	comparison group selection
Query Result	stock prices, news, and histories
Attribute Domain (Category)	competition by category groups
Attributes Domain (Items)	low, high, volume
Entity Domain	public traded
Quantity	very good
Usage Fee	free

Data Source Name	**Corptech**
Original source	CorpTech high-tech database
URL	http://www.corptech.com/
Query Interface	restricted domain - three column search
Query input	name, ticker, URL
Query Result	structured text and HTML tables
Attribute Domain (Category)	directory, background, financial, competitor
Attributes Domain (Items)	address, sales, CEO, product, link anchors
Entity Domain	US high-tech manufacturers
Quantity	45,000
Usage Fee	free, free membership, membership

Appendix A-3 Telecommunication Industry

Data source name: British Telecom (Communications Store: business news and case studies for journalists and analysts).
URL http://www.commstore.bt.com
Query Input: full-text search of case studies and/or press releases. browse case studies by industry sector: industry, finance, government, retail, services, energy, logistics
Query Result: list of anchor-references to relevant documents.
Entity Domain case studies for telecommunications applications involving British Telecom worldwide in all business sectors. sources include case studies and press releases.
Usage Fee: "free subscription" user registration required.

Data source name: Communicating Online
URL http://commnews.com/
Query Input: Browsing based upon subject category: events, internet news, legislation, technology, people/officer-changes.
Query Result: list of anchor-references to documents by subject category.
Entity Domain Regional publication for West Coast/Northwest Communications News. includes coverage of national-level corporate Information and events surrounding and concerning membership of the Information Technology and Telecommunications Association (TCA) Short news articles Press releases Categories: events, internet news, legislation, technology, officer-changes
Usage Fee: free

Data source name: Achive of the Federal Information News Syndicate
URL http://sunsite.utk.edu/FINS
Query Input: browse only by subject area. subjects: public information, information resources management, information infrastructure, telecommunications infra.
Query Result: documents in chronological order listed by title.
Entity Domain Public policy papers, US Congressional bills, US Public Laws pertaining to global information and telecommunications systems. EPIC as well as news releases, RAND abstracts, reports, House hearings. The problem is there is no way of knowing how "complete" the source is. Neither 'necessary' nor 'sufficient' for establishing that something does or does not exist.
Usage Fee: free.

Data source name: International Technology Conultants
URL http://www.itcresearch.com
Query Input: browsing of "sample" country reports and market studies
Query Result: anchor-references to dated country reports and market studies as examples of products
Entity Domain Free text. Market studies for sectors of the telecommunications industry. Country reports by region: Latin America and the Carribean, China, India, and the Pacific Rim Russia, Central Asia, and Central Europe, USA, Western Europe, and Japan.
Usage Fee: yearly subscription for monthly reports

Data source name: Paul Budde Communication's Telecommunications and Superhighway News
URL http://www.budde.com.au
Query Input: Browsable interface to document database and telecommunications news service.
Query Result: Anchor-references to documents
Entity Domain Monthly newsletters and market research reports. Database of country and market research files on a "pay-per-document" basis Africa, europe, Latin America, US, Europe, South Pacific Data has a decided Australia/South Pacific bent. Telecommunications News Service, updated weekly, focus on Asian and the Pacific rim Free press releases on the Australian, "Asian", and New Zealand markets.
Usage Fee: Pay-per-document for market research Monthly newsletters for a yearly fee Free news stories

Data source name: Telecomm A.M. TPG Telecom Publishing Group Warren Publishing Inc.
URL http://www.telecommunications.com
Query Input: Browsing interface by year and month to calendar Browsing interface to paper-based products
Query Result: Lists of products which may be purchased.
Entity Domain Calendar - industry related conferences and events.
Usage Fee: On-line ordering of telecommunications market intelligence data. Calendar is free Yearly subscription on-line and/or CD products.

Data source name: Telecom Strategies Business research databse of Investext Group.
URL http://telecom.securities.com
Query Input: enable thesarus. structured query-unrestricted domain (free-text) OR. structured query-restricted domain (but the domains are VERY LARGE on the order of 100's of entries) Schema: company, business subject, product mentioned, geographic region, title, source, date of publication. Search within: document, page, paragraph, sentence, line.
Query Result: List of anchor-references in reverse chronological order
Covered domain: For the fee service: 7,000 plus company and industry research reports from financial org. including Merrill Lynch, Smith Barney, etc. Geographic coverge: North America and Europe (plus a few others) Includes full text of newspapers and trade journals (which?)
Usage Fee: payment on a per-document basis

Data source name: Telecom Update Angus TeleManagement's Weekly Telecom Newsbulletin News and telecom in Canada.
URL http://www.angustel.ca/update/up.html
Query Input: browse - current issue by article browse - archive by reverse chronological date ordering. full-text search of back issues
Query Result: list of articles (not anchors) list of anchor-references to issues (all of the articles that make up an issue are contained in a single file). anchor-references in relevance-ranked order as per EXCITE indexing engine.
Entity Domain Weekly update of telecom news for Canada. Calendar - telecommunications events nation wide.
Usage Fee: free subscription for free email is available.

Data source name: Washington Telecom Newswire
URL http://wtn.com/wtn/wtn.html
Query Input: not clear - all data is hidden behind the fee interface. possible browsing of past articles.
Query Result: not - clear
Entity Domain not - clear. policy data focusing on Washington, D.C.
Usage Fee: site licensed fee

Data source name: Telecommunications OnLine (Telecommunications Magazine)
URL http://www.telecoms-mag.com
Query Input: free-text search interface text-based indexing. browse by

subject area. browse previous issues by month and year. browse calendar of trade show events by month and year
Query Result: anchor-references listed by relevance ranking (what relevance measure?) to the article from a particular issue (month/year)
Covered domain: calendar of trade shows around the country
Usage Fee: Free on-line access Register for a "free" paper subscription.

Data source name: 'Telecommunications now!' Call Centre Focus
URL http://www.callcentre.co.uk
Query Input:
Query Result:
Entity Domain
Usage Fee: Free access, registration required.

Data source name: Spread Spectrum Scene (technology-based)
URL http://sss-mag.com/search.html
Query Input: full-text search
Query Result: anchor-reference and summary to documents listed in order of relevance (Excite indexing engine)
Entity Domain news, projects, and industry developments for the spread-spectrum wireless technology industry.
Usage Fee: free

Data source name: Phone Plus
URL http://www.vpico.com/pp
Query Input: full-text search
Query Result: anchor-reference and summary to documents (not clear whether they are listed in a relevance-ranked order) structured query-restricted domain for industry directory search by company product category (see below) structured query-restricted domain for long distance carriers and resellers. search by company product category (see below)
Entity Domain phone and telecommunications news list of companies in telecommunications (industry directory) addresses, phone numbers, contact info. unclear ... corporate headquarters or a sales office. company product categories: equipment, services, payphone, calling cards, repair, data, IXC, etc.
Usage Fee: subscription for paper?

Data source name: PennWell Media Online Lightwave Xtra
URL http://www.telecoms-mag.com
Query Input: full-text query for news full-text query for companies browsing for companies by company name, category (industry/product) - their own categorization. browse by product/service
Query Result: anchor-reference and article summary listed by relevance ranking (Excite engine) returns one "weekly issue". all articles from one issue are in a single file. anchor-reference to a one-page tear sheet of company profiles
Entity Domain weekly update of news and industry developments related to fiber optics technologies and telecommunications company profile: company address and brief description of products/industry sector 2000+ pages of information on vendors and products - plus 700+ categories and more than 15,000 links to vendors,
Usage Fee: free access free subscription for paper product

Data source name: E:Media online magazine with updates to global telecommunications and
URL http://www.emedia.net.uk/fnews.html
Query Input: browsable access to news by date, products,
Query Result: anchor-references and summaries to relevant articles. for the changes-in-leadership, anchor references to a list of summaries for a given month and year.
Entity Domain primarily European focus for data on: Call centres and CTI, networking and the Internet, telcos and companies, events, products, mobile and cable, telecoms services, audiotex and telemedia Changes in personnel - leadership changes in the industry (large co's worldwide)
Usage Fee: free. subscription for paper periodical

Appendix A-4 Pharmaceutical Industry

Data Source Name	**Pharmaceutical companies on the world wide web**
URL	http://members.aol.com/pharminf/pharmcom.html
Query Interface	Alphabetical browsing
Query input	N/A
Query Result	HTML tagged document
Attribute Domain (Category)	Directory, Disorder Information Sites
Attributes Domain (Items)	Company website URL
Entity Domain	Major Pharmaceuticals, incl.international companies
Quantity	Around 60
Usage Fee	Free

Data Source Name	**Pharmaceutical online**
URL	http://www.pharmaceuticalonline.com/
Query Interface	Unrestricted domain
Query input	Product, company
Query Result	HTML tagged document
Attribute Domain (Category)	Directory
Attributes Domain (Items)	Address, Phone, Fax, Website URL
Entity Domain	Pharmaceutical equipment suppliers
Quantity	500 - 1000 companies
Usage Fee	Free

Data Source Name	**PharmaIndia**
URL	http://www.pharmaindia.com/
Query Interface	Unrestricted domain
Query input	Companies; Products/Services; Raw Materials; Requirements
Query Result	HTML taggeed document
Attribute Domain (Category)	Directory, Marketing
Attributes Domain (Items)	Address, etc; Product / service
Entity Domain	Pharmaceutical equipment suppliers
Quantity	< 100
Usage Fee	Free

Data Source Name	**PharmWeb Yellow Pages**
URL	http://www.pharmweb.net/
Query Interface	full text search; browsing by sub-industries
Query input	N/A
Query Result	HTML tagged page
Attribute Domain (Category)	Articles
Attributes Domain (Items)	
Entity Domain	Pharmaceutical articles
Quantity	Many
Usage Fee	Free

Data Source Name	**Association of British Pharmaceutical Industry (ABPI)**
URL	http://www.abpi.org.uk/
Query Interface	Browsing
Query input	N/A
Query Result	HTML anchors to documents
Attribute Domain (Category)	Directory
Attributes Domain (Items)	Name / URL
Entity Domain	UK & others
Quantity	100
Usage Fee	free

(may have general pharmaceutical market info)

Data Source Name	**Pharmaceutical Research Manufacturer's Association (PhRMA)**
Original source	Articles from Yahoo, Reuters, Science..., also proprietary articles
URL	http://www.phrma.org/
Query Interface	Full text search, browsing of facts and figures
Query input	N/A
Query Result	HTML anchors to documents
Attribute Domain (Category)	Articles
Attributes Domain (Items)	URL to proprietary articles and also articles on other sites
Entity Domain	news article from Reuters, Yahoo, Science news / information on pharmaceuticals, genomics
Quantity	numerous
Usage Fee	free, depends on the original source of the article

Data Source Name	**Bio Online: Biotechnology: Companies**
URL	http://cns.bio.com/
Query Interface	unrestricted domain- 4 column search, browsing
Query input	Subject, # files to return
Query Result	HTML anchors to documents
Attribute Domain (Category)	News & Event, Companies, Career, Products & Services, Industry & Government
Attributes Domain (Items)	Address, etc.; # employeees; company statement; products/ services
Entity Domain	Biotech companies, research centers
Quantity	Numerous
Usage Fee	Free

Data Source Name	**Network Science**
URL	http://www.netsci.org/Resources/
Query Interface	full text
Query input	N/A
Query Result	HTML anchors to documents
Attribute Domain (Category)	Directory; Market; Finance
Attributes Domain (Items)	Address, etc.; Status (public or not); Company description; Therapeutical area addressed; Company size; Revenues
Entity Domain	limited to US companies now; scientific focus articles, literature reviews, industry and product updates and a biotechnology focus section
Quantity	around 200
Usage Fee	free

Data Source Name	**Virtual Library Pharmacy**
URL	http://www.cpb.uokhsc.edu/pharmacy/pharmint.html
Query Interface	Browsing
Query input	N/A
Query Result	N/A
Attribute Domain (Category)	N/A
Attributes Domain (Items)	N/A
Entity Domain	
Quantity	25

Usage Fee	free

Data Source Name	**F1RSTMARK**
URL	http://www.firstmark.com/fmkmbf/medfind.htm
Query Interface	full text
Query input	N/A
Query Result	HTML tagged document
Attribute Domain (Category)	Directory
Attributes Domain (Items)	Name, URL
Entity Domain	Pharmaceutical, Biotech companies, US
Quantity	1,350 biotechnology companies in the U.S. 957 pharmaceutical companies in the U.S. More than 10,000 biotech and pharmaceutical executives
Usage Fee	free service returns name and address; registration required; for more details 32¢ per record (minimum $595)

Data Source Name	**Publications of Interest to the Biological Sciences**
URL	http://www.nsf.gov/bio/start.htm
Query Interface	Browsing only
Query input	N/A
Query Result	N/A
Attribute Domain (Category)	Articles
Attributes Domain (Items)	Articles
Entity Domain	Biology
Quantity	Numerous
Usage Fee	Free

Data Source Name	**National Institute of Health (NIH)**
Original source	NIH
URL	http://search.info.nih.gov/
Query Interface	full-text
Query input	N/A
Query Result	HTML
Attribute Domain (Category)	news articles and other info pages
Attributes Domain (Items)	URL's, pdf, xls files and abstracts
Entity Domain	Heatlth

Quantity	numerous
Usage Fee	Free

Data Source Name	**MedMarket**
URL	http://www.medmarket.com/
Query Interface	Restricted Domain
Query input	Concept or word
Query Result	links to articles
Attribute Domain (Category)	articles
Attributes Domain (Items)	articles
Entity Domain	Medical suppliers
Quantity	around 50
Usage Fee	free

Appendix B Log of Logic Flow

The research were conducted by Tom Lee, Benny Suen and Steve Tu.

Appendix B-1 Computer Industry

1. Bill, a information specialist Bill, working for PW, has the company name *"Sun Microsystems"* in mind
    - Bill has the prior knowledge that Sun Microsystems is a US company
    - Bill also has the knowledge that Sun Microsystems is a publicly traded company
2. Bill consults ComputerDS
    - Bill finds out that *SEC Edgar* is a almost complete source for publicly traded company in US
3. Bill decides to use *SEC EDGAR*, and from there he selected 10K report
    - After scrolling down a bit, Bill noticed two subsidiaries *SunSoft* and *Nihon Sun*
    - Bill also understands from 10K report that Sunsoft is a domestic company incorporated in US California, and Nihon Sun is a foreign company incorporated in Japan
4. To collect industry information of SunSoft, Bill consults ComputerDS again
    - Assuming that Bill knows that SunSoft is not a public company
    - Bill therefore needs to explore data sources containing prrivate companies
5. Bill decides to explore *Thomas Register Online, Companies Online, and CorpFiNet*
6. None of the three sources contains information about SunSoft
    - Bill now understands that the reason is that SunSoft is a small company, while the above three mainly contain large companies across various industries
    - Bill decides to try more specific sources. He hence focuses on data source for high-tech industry
7. Bill consults ComputerDS and focuses on *CorpTech, Tradeport Californian Company*, and *Tennessee's High-Tech database*
    - *Tennessee's High-Tech database* is impossible because of the disjoint geographical scope
8. Bill is now left with only *CorpTech* and *Tradeport Californian Company*
    - Bill notes the quantity (size) difference for the two possible sources- *CorpTech* contains more high-tech companies than *Tradeport Californian Company*
9. Bill therefore decides to use *CorpTech* to find information about SunSoft
10. SunSoft is found from CorpTech.
    - Bill obtains its industry type, product, address, and SIC code
    - Now Bill is able to determine competition conflict with existing clients using the acquired information

11. Bill still wonders if product and industry type is a reasonable indication of competition for computer industry
12. Bill consults ComputerDS and decide *Silicon Investors Profile* to gain more subtle competition information
    - *Sunsoft* however is not there. But *Sun Microsystems* is included
    - Bill therefore browses around to gain more information not reported by *CorpTech*
    - Bill assumes that these additional information should also be a good indication for SunSoft
13. Bill now answers all of the questions for *SunSoft*> except the question:
    - Where are SunSoft's products get sold to?
14. Now Bill needs to repeat the same process for *Nihon Sun*. Again He consults ComputerDS
    - *Tradeport Foreign Company* and *Yahoo company Search* are the only two containing foreign companies
    - *Tradeport Foreign Company* contains more foreign companies than *Yahoo company Search*
    - Bill selects *Tradeport Foreign Company* and the same information for *Nihon Sun* is acquired from this source
15. Based on the address information, Bill sends out written inquiry to *SunSoft and Nihon Sun*

Appendix B-2 Telecommunication Industry

1. Conducting a search on wireless services PW, searches for information on *"Sprint PCS"*
   - We know a'priori that Sprint PCS is somehow associated with Sprint Communications
   - We know that Sprint is a publically traded company
2. From our list of General company sources
   - We see that *Yahoo* is one way to get introductory information
   - We also see that *SEC EDGAR* may provide an overview of the company
3. From *Yahoo*, we enter ``Sprint PCS"
   - A result table indicates that there is no separate ticker symbol for Sprint PCS suggesting that Sprint PCS is not publically traded.
   - A result table gives us a pointer to Sprint Corp
   - We retrieve a URL for Sprint PCS
4. Scanning and browsing the Yahoo profile on Sprint Corp we see that
   - Sprint PCS contributes to Sprint Corps financials
5. From the Sprint PCS Web page, a list of facts includes
   - the parent companies that make up the Sprint PCS partnership
   - the Sprint PCS headquarters
   - officers of the partnership
   - Sprint PCS uses CDMA technology
   - a list of the cities where Sprint PCS has operating licenses and where they are currently operating.
6. From *SEC EDGAR*, 10K report
7. From *SEC EDGAR*, 10K report
   - Reading the report reveals that Sprint PCS is a limited partnership of Cox, TCI, Comcast, and Sprint
   - Sprint PCS is organized under the Sprint Spectrum Holding Company
   - Sprint Spectrum is formally a subsidiary of the class Delaware Partnership under Sprint Enterprises which is itself listed as a subsidiary of the Sprint subsidiary, UCOM, Inc.
8. To search for recent officer changes in telecommunications, we might consult *Emedia*
   - No reference to Sprint PCS, Sprint, or Sprint Spectrum is reported.
   - We recognize the hazard that just because no officership change is reported, it does not mean that an officership change did not occur
9. To search for competitors and product information with respect to Sprint PCS, we might consider a number of sources including: America's Network, Phone Plus, PennWell Media On-Line, and Telecom Securities.
   - Unfortunately, PennWell Media focuses on fiber-optics technologies which does not match the wireless services offered by Sprint PCS.
   - Our limited budget makes the charge-per-document fee schedule of Telecom Securities undesirable.

10. Beginning first with America's network, references to several news stories and essays are returned from a string search on ``Sprint PCS".
    - PCS competitors include: Bell Atlantic Nynex Mobile, Nextel (Craig McCaw), BANM digital cellular, and AT&T Wireless.
    - there is no indication that this is an exhaustive list.
    - Competing technologies include: TDMA (AT&T, Nextel) and analog networks (BANM)
11. Likewise, turning next to Phone Plus, a string search also returns a number of stories and essays.
    - GSM as a competing standard to CDMA
    - Western Wireless as a GSM provider and competitor to Sprint PCS

Appendix C Yahoo! Classification

Yahoo!'s classification under Business and Economy.

Advertising@
Aerospace (530)
African American (48)
Agriculture (2056)
Animals (3806)
Apparel (4144)
Architecture (1635)
Arts and Crafts (5469)
Auctions (358)
Audio (724)
Automotive (10237)
Aviation@
Beverages (2534)
Biomedical (514)
Books (6449)
Catalogs (58)
Chemicals (1028)
Children (939)
Cleaning (477)
Communications and Media Services (2569)
Computers (33277)
Construction (11802)
Consulting (1319)
Conventions and Trade Shows (522)
Corporate Services (2014)
Design (2051)
Disabilities (383)
Education (2471)
Electronics (6358)
Emergency Services (464)
Employment (1854)
Energy (1571)
Engineering (2730)
Entertainment (5053)
Environment (1552)
Ergonomics (46)
Explosives (9)
Financial Services (22098)
Flags (79)
Flowers (496)
Food (5228)
Franchises (188)
Fund Raising (163)
Funerals (199)
Games (1065)
General Merchandise (222)
Gifts (1577)
Government (219)
Health (15829)
History (18)
Hobbies (3525)
Home and Garden (5475)
Hospitality Industry (374)
Imaging (623)
Industrial Supplies (8178)
Information (1182)
Insurance@
Internet Services (32435)
Jewelry (1310)
Law (4306)
Manufacturing (2845)
Marketing (3452)
Mining and Mineral Exploration (524)
Music (6998)
Navigation (427)
Networks (47)
News and Media (3489)
Newsletters (60)
Office Supplies and Services (1621)
Outdoors (6362)
Packaging (567)
Paranormal Phenomena (503)
Party Supplies@
Personal Care (1278)
Photography (2980)
Printing (1516)
Publishing (3218)
Quality (235)
Real Estate (18809)
Religion (915)
Research and Development (157)
Restaurants (6163)
Retailers (1338)
Scientific (1930)
Security (1055)
Seminars@
Sex (5015)
Shipping@
Shopping Centers (1125)
Software@
Speakers (269)
Sports (6335)
Storage (282)
Technology Transfer (39)
Telecommunications (3686)
Toys (646)
Trade (465)
Transportation (4498)
Travel (23212)
Utilities (694)
Vending Machines@
Weapons (436)
Weather (105)
Miscellaneous (101)