Research Plan for Leveraging Social Information Systems: Using Blogs to Inform Technology Strategy Decisions

Satwik Seshasai

Working Paper CISL# 2008-07

May 2008

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

Research Plan for Leveraging Social Information Systems: Using Blogs to Inform Technology Strategy Decisions

Satwik Seshasai
MIT Engineering Systems Division

May 19, 2008

Table of Contents

Preamble

This report represents an update on my doctoral research plans in the Engineering Systems Division at MIT. Even getting to this early stage would not have been possible without the guidance of my doctoral committee, Professors Stuart Madnick (chair), Joseph Sussman, Irving Wladawsky-Berger and Wei Lee Woon. This work has been developed within the Context group at MIT Sloan and is being done in conjunction with a joint project between MIT and the Masdar Institute.

Introduction

Blogging technology is becoming an increasingly popular means of publishing information to a mass audience – recent studies report 12 million users and growing[1]. A blog, short for "web log" is commonly defined as a web page with posts published by an author and displayed in reverse chronological order. As blogs move from personal diaries to use within research and business to share ideas, they are becoming a prime example of the growing volume of information content being produced in the domain of many engineering systems. Blogging technology has produced an information system which has social and technical aspects that can be leveraged by organizations to inform their decisions about which strategic technical alternatives to pursue. Information contained in blogs can reveal trends regarding which topics are being discussed at a growing or declining rate over time. Blogs can also inform organizations about which individuals and institutions to consult, both within and outside the organization. However, to reach this potential, we must develop an understanding of how existing methods of analyzing information can be adapted to apply to blogs, and also of how the unique aspects of blogs can help adapt system representation methods to incorporate more dynamic information content.

First, we need to define the notion of technology strategy as the overall direction for a particular technology based division. Specifically, this involves the primary goal of the division, and the various technical alternatives available to meet that goal. The technology strategy is a selection of one or more technical alternatives to meet the specified goal. In terms of decision making sources, there are a number of sources which are used to influence the selection among various technical alternatives. Market conditions, current customer base, organizational capabilities, and financial strength are factors that can influence the decision making. However, the sources we are concerned with in this study are knowledge sources which provide new insights to decision makers, and specifically within the set of knowledge sources we are concerned with blogs as the source we are seeking to understand better.

A motivating example is whether the existence of methods proposed in this project would have helped predict the growing acceptance of Ethernet over token ring as the dominant networking standard and allowed firms such as IBM to adapt earlier.

We begin by collecting information from outside the organization. Very specific existing methods exist, utilizing bibliometric analysis, to scan structured information such as patents and published literature and fit periodic data to growth curves, to forecast which technologies are growing. Applying these methods to blogs present a set of very unique challenges. Once done, this provides the potential to run frequent periodic scans of the Internet to quickly gather a representation of the information available within the system. The periodic scans provide a set of quantitative information which can be treated in the same light as a qualitative stakeholder who is used to inform a system representation. The CLIOS method exists for choosing among a set of strategic alternatives for an engineering system. The first phase of this method is system representation, generated by a manual process of interviewing stakeholders. The quantitative results provided by the bibliometric analysis will augment the human stakeholders and provide more dynamic insight into system representation in particular domains.

Organizations such as IBM have internal sources of information in the form of blogs, and in choosing among strategic technical alternatives, it is as important to leverage these internal databases. In addition to the analysis mentioned above, an additional question which may be asked inside the organization is which individuals are best suited to assist with a particular technology. Thus, the analysis software which was described above will be extended to look at internal profile systems and use this information to share context between profiles and blogs and determine an organization's capacity to handle various technical alternatives. The combination of analyzing external and internal sources will allow organizations to leverage this information to their benefit.

This work will provide both a methodological and domain specific contribution. Methodologically, the goal is to extend existing methods for system representation (as discussed in the Phase II of this report) and for bibliometric analysis of large scale systems of information. The specific domain which has been studied in preliminary work is renewable energy – driven by the complementary work being done by in conjunction with the Masdar Institute [2]. By applying these methods to this specific domain, particular insights have been gained which will drive work related to ten domains in the broader study.

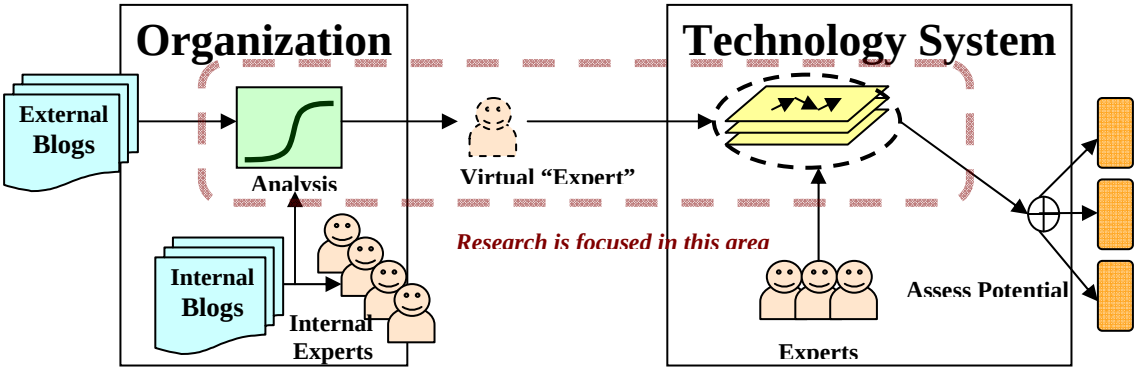

Figure 1: Overview figure.

The methodology for dynamic system representation using new means of analyzing social information (blogs, internal experts) will be done by extending bibliometric methods to include blogs, extend context interchange to analyze internal blogs/experts, and then extending system representation to be informed by analysis, introducing the "virtual expert" concept to link the information analysis to the system representation phase. This methodology will be used to assess strategic alternatives for technology strategy on a more frequent basis. In applying the above methodology to a specific domain to gather insight, we will survey organization stakeholders to understand alternatives they are choosing between and then survey organization stakeholders to validate the utility of the method.

Report Structure

This proposal will be structured as follows. In the next section, the motivation for this study will be discussed, in addition to the opportunity being presented to make an impact with this work. Based on this opportunity, a broad research question and hypothesis are presented, with specific pieces of the broad question identified. We then describe the overall intellectual contribution which results from answering this research question, along with specific methodological and domain-specific contributions, and a literature review of existing contributions in this area. The following three sections present a series of interrelated research phases with preliminary work and results. Each answers a subsequent piece of the research question and each indicates a specific contribution as well. After discussing specifics of the phases, the use of survey to validate the findings in this research is discussed. Finally, a proposed table of contents for my dissertation, and the specific project milestones and timeline for this work are presented.

Motivation and Opportunity

The motivation for this study comes from the goal of allowing organizations to use the vast array of information sources available on the Internet to inform them of the emerging trends related to their domain. The information sources available today could allow organizations to forecast which areas of technology require focus. Existing work in 'technology mining' of formal sources such as patents, published literature provides an opportunity to build a methodology for doing this with a much more dynamic and flexible system such as the blogs available on the internet. Specific technology choices in a

domain can be made with full information on what is being discussed inside the organization and outside in the public. Services such as Vantage Point[3] and the Thompson Collexis Dashboard [4] have demonstrated that the need and interest in cultivating knowledge found in blogs is increasing rapidly.

Blogs also represent a complete socio-technical system, containing both content as well as a social context of authors, readers, commentators and participating organizations. This provides an opportunity to not only determine areas of technical focus, but also to determine who the experts are, both inside and outside the company, to consult on the new areas of focus. A growing use of social software inside and outside the organization has caused knowledge to spread through blogs, wikis, shared bookmarks and other less formal means than the traditional academic journals and patents. The opportunity exists to leverage 'context interchange' technology to bridge the context between these sources – vital in an ad hoc and informal knowledge exchange – and calibrate heterogeneous sources. The systems level representation through the method discussed in phase II provides opportunity to interpret data in the context of existing systems, and proactively plan for disruptive technology.

Research Question and Hypothesis

At a very high level, the research question being considered here is what methods are useful in allowing organizations to use emerging information systems to make better decisions on technology strategy. Our goal is to test the hypothesis that improved information retrieval and representation of socio-technical information systems can inform technology strategy.

To help structure our research, all associated activities will be divided into a set of inter-related phases, each of which addresses a key component of the above question. At the micro-level, we intend to study and extend a set of computational tools which will facilitate the latter stages of our analysis. The first phase will address the need to process external information sources to gather insight on the overall strategic technology alternatives for the organization. Next, there is the need to articulate the impact of various emerging trends in a broader context. The second phase, consisting of a methodological extension of the CLIOS system representation phase, examines how the tools developed thus far fit into the greater domain-specific information eco-system for a specific domain. Based on emerging behavior identified via system representation, the third phase will evaluate the internal capacity of the organization to react to each strategic alternative with a framework for analyzing internal information sources. Finally, there is a need to aggregate and process the outputs of these tools. The final piece of the research question is whether the results gathered by these new methods are in fact useful to the organization – a survey of existing organization stakeholders will be conducted to assess the utility of the various results. This final analysis will present an overall evaluation of the various phases of using the information systems to inform technology strategy.

Contribution

The core intellectual contribution of this research is a framework for leveraging information systems with a significant volume of content and high degree of social context to the information. Such information systems have a technological basis for creating, updating and interacting with the information as well as a set of social and organizational characteristics around incentives for participation, rules for access and relationships between participants. These aspects add a level of complexity that has thus far made information within systems such as this not truly utilized to their fullest potential, beyond basic search operations against the information.

In the process of producing the contribution described above, a series of methodological and domain specific contributions will be made. Existing methodologies for analyzing information will be extended to include the concept of social information systems. For example, the existing methodology for technology forecasting through bibliometric analysis of published literature will be extended to

incorporate blogs. Well understood methods of system representation such as within the CLIOS process will be extended to the concept of dynamic system representation based on real time information retrieval and identification of the institutional and technological components of the system. Finally, the notion of context interchange which has been focused on semantic definitions of terms will be extended to the notion of context based on which individual contributed a particular piece of information within an organization – this notion of context can fill the gap of navigating the semantic web as a social information infrastructure.

On a domain-specific level, the proposed study will contribute to the method used by technology based organizations – specifically, IBM – to utilize new sources of information and incorporate adaptability within their information retrieval process. Specifically, we will focus on ten domains and key insights may be revealed within each domain based on new information found within the blogs used in the particular domain.

For the organization audience of this research, a specific set of tools and deliverables will be provided which represent the intellectual contribution described in this section. These online tools will allow individuals in an extended organization to query a heterogeneous set of databases to retrieve people and information related to specific business goals in a more relevant fashion than what is currently available. The key value proposition of these methods is that they:

- Incorporate existing databases and systems
- Accommodate systems not originally intended for this purpose
- Realize new ways to use information and social connections
- Satisfy needs which emerge after initial data is created

In the process of developing these methods, a set of requirements is expected to emerge for socio-technical data sources to be integrated by organizations into business processes. Thus, this research will propose a set of system level requirements and considerations required to adopt an information system as described above. Finally, the nature of analyzing information such as this requires advancements in context interchange technology, to incorporate the informal nature of the information as well as the profile information of the individuals contributing the information.

The true value of this work is quite simply the operationalization of blog information into the decision making process. For individuals or groups within an organization who are responsible for making technology strategy decisions (such as whether to invest in ethernet or token ring), the blogosphere is one potentially useful source in determining what the relative interest level is, who the relevant institutions and actors are. In today's world, the only recourse for utilizing this information is manual search capabilities such as Google which offer a significant volume of information at once. Thus, the application of analytical rigor to this problem will be useful not only for forecasting future growth of technology but also in describing the past and current state of technology.

Literature Review

This section focuses on the relevant literature on technology forecasting as well as the emerging research on blogs – two key focus areas for the project. The formal notion of technology forecasting has existed for over forty years, described in the first volume of the *Technology Forecasting* journal by MIT professor Edward B. Roberts as the combination of predicting technical achievements and the allocation of resources towards future technological progress[5]. A number of methods are presented, from consulting experts to observing performance of critical metrics. One recent example is Koh and Magee's approach relying on key metrics in functional categories (such as calculations per second)[6]. Ten years later, an article on the accuracy of various technology forecasting methods compared the goodness of fit and regression statistics of various approaches to show that simpler methods tend to be more useful [7].

The 20th anniversary of Roberts' article noted the application of forecasting methods to policy makers and executives in the government, military and private sector [8].

The use of bibliometrics (measurement of texts and information) in forecasting has been used with well structured information sources such as patents and scientific literature [9], employing statistical methods such as curve fitting to known growth curves. Visualization is also considered as another means for analyzing data [10], since precise quantitative results are not as critical as qualitative insight. Quantitative models of bibliometric trends have been applied to business settings as well as science and technology policy settings to help assess the maturity level of a technology and when to explore a new approach [11][12]. Cherie Courseault's Ph.D. thesis provides much of the inspiration to this project, and has proposed that firms should begin to look at bibliometric analysis of publication databases for competitive advantage [13]. Minimal reference has been made to the use of online environments in forecasting [14] and work in this area has focused largely on the aggregate user traffic rather than inspecting content such as blogs.

However, various aspects of the research community have started to explore using blogs as a data source in formal research. Initial excitement around relative ease of access to the digital assets has led to challenges such methodological adaptation, data extraction, and semantic analysis. Hookway describes the qualitative value of blogs as an online extension of diary research, and discusses a number of methodological implications such as participant/researcher interaction, soliciting a community to study, and the ethics of researching public blogs of individuals [15]. Studies on blog data either extract, index and analyze a particular corpus of data from the web, or perform queries using internet search engines across the entire space of blogs. Network metrics such as connectedness have been applied to specific communities of blogs [16]. The Neilsen Company hosts BlogPulse.com which performs very basic searches, to the level of comparing term frequency of three terms over six months within blog posts [17] and other examples exist of research on how blogs can predict general interest level in various topics [18][]. More recent works have introduced the analysis of comments to blog posts [19]. Moving from studying the nature of blog usage to examining the content of blogs for insight, studies have looked at segmenting corporate blogs and applying known methods of text analysis such as latent semantic analysis to assess the relative interest levels of various terms [20]. Anjewierden and Efimova combine text analytics with network theory to categorize blogs across the dimensions of people, documents, terms, links and time [21]. The Digital Organization Research Institute has gone even further, and introduced the notion of Semantically-Interlinked Online Communities and developed a formal ontology for expressing the main object classes within a blog [22].

The goal of our work is to build on the tools and methods described above, but make a major contribution by building a framework to act on insights gained through analysis of blogs. Elements of the described contributions also require additional progress – such as the introduction of time-varying analysis and keyword clusters instead of single keywords.

Assessing and Improving the Current State: Data Collection Plan

Data Collection Plan: Key Points
• IBM Horzonwatch includes 1700 individuals who could directly benefit
• Leadership is willing to participate
• Ten domains will be used for deep analysis

The current state is best described through the viewpoint of our selected target for this study – the population is described using the initial phase of survey methodology from Groves, et. al. [23]. Our population of interest is the overall set of large organizations driven by technology. Technology is a focus because this is a domain in which significant changes in availability of new information and tools can produce significant changes in the business, as opposed to industries which are more subject to change based on

other factors. The segment of large organizations was chosen because they are more likely to possess the ability to apply existing human resources to new problems and thus shift strategy at a broader level than most mid-size firms who are heavily invested in one area. Large organizations in this case are defined as greater than 50,000 employees. Our population of inference is firms within the computer software and hardware business – this was chosen because this domain has a high degree of activity within public blogs and may provide a large enough data set to reach conclusions which may apply to other industries in the near future. The target population for this study is the IBM Corporation, which represents a firm invested in a variety of directions but also one which has successfully made strategic decisions in the past based on market intelligence. This firm has been chosen in part because the researcher has ready access to the respondents. The sampling frame for the first phase of the study will consist of the employee directory of IBM, while the second phase of the study will focus on those who respond positively in the first phase.

It is obviously worth considering to what degree respondents from IBM represent the viewpoints of respondents from other peer technology firms of similar size. For the scope of this study we will acknowledge this as a potential gap and recommend future research to validate the findings of this study across other companies. Since a good part of this project involves testing new information analysis methods, a limited sampling frame is beneficial in getting high quality and deep feedback which can drive an improved version of the tools. This second version can be tested with a larger population.

A key piece of IBM's approach to technology strategy is a program known as "Horzonwatch", which represents a manual attempt to solve the problems discussed above [24]. Faced with a monumental volume of information on the Internet, the 1700 member community within IBM holds monthly conference calls and spends countless hours manually searching the Internet for information related to various technical topics, in an effort to identify emerging trends which the company should be aware of, and which individuals and partner institutions should be consulted. This process can be augmented, or over time replaced, by a more structured manner of analyzing the information on the Internet. The members of the Horzonwatch community will be surveyed to gather key needs and requirements, and then this same community will be presented with results and surveyed to determine the value of the methods developed in this research.

The IBM Horzonwatch group was chosen as the target audience for the pilot study, as an example of how the internet is being used today to gather market intelligence and predict emerging technologies. This is a group of 1700 IBM employees who have joined a community led by William Chamberlain in the market intelligence function of IBM, and hold period conference calls, email and blog discussions to discuss emerging trends in various fields. Following the work of the Horzonwatch group provides the essence of a “manual walkthrough” of the process which this research seeks to automate.

An initial interview was conducted with William Chamberlain, the director of the Horzonwatch group, to identify aspects of the “manual” process which he goes through in leading this group. Chamberlain described his work as very “labor intensive”, and he has even avoided some of the semi-automated options such as tracking RSS feeds because of the level of noise they bring in. He uses the concept of “Google alerts” to automatically search for the same term daily and provide him with results in his email. Due to funding pressures, he is limited in what he can do with his time, but does report that surveys of the 1700 members indicate that there is value and appreciation in what this group is doing, and that there are many areas of information which they wish they could be leveraging.

The current work being done by the Horzonwatch team is very broad in nature. There is no focus on a specific domain or industry – the goal is to scan as broadly as possible since the company desires to be ready for disruption in any domain. One of the data sources being used now is conference databases – rather than wait for academic journals to be published, Chamberlain uses conference databases to track the popularity of topics that are being spoken about in conferences. Being able to leverage information in

blogs would be an even greater step in this direction since it would allow this group to know what individuals are talking about on a daily basis.

The audience within the Horzonwatch community is a broad spectrum of employees within IBM who have cited their interest in tying sources on the public Internet to opportunities they can leverage within the organization. Members come from every functional area and many geographies within IBM. Each conference call has attendance from 80 to 100 members, and the discussion blog gets around 100 to 120 hits per day according to Chamberlain. Some of the desired improvements cited in a December 2007 report include the ability to find subject matter experts internally and externally, with better tools that enable social networking with the subject matter experts, and better search tools to utilize sources on the internet. By bridging both external and internal sources and by bridging content and people, this study is meeting the stated goals of the Horzonwatch project. Thus, this audience will serve as a valuable validation metric for the methods developed in this study.

Ten candidate domains have been chosen based on current Horzonwatch activity, and will be confirmed over the next two months before beginning the data collection:

Ten Candidate Domains from IBM:		IT for Energy	IT for Personalized Healthcare
Financial Markets	Business in Africa	Smart Cars	Customer Service Systems
Skin Sensors	Consumer Mobile	Cloud Computing	Unified Communications

Blogs as an Information System

Rather than simply a source of data, the blogs available on the internet represent a complete engineering system exhibiting social and technical characteristics which define engineering systems[25]. The information-based system consists of blogs on the public internet – the sheer volume represents a significant scale and a very broad scope in terms of users, countries, organizations and domains. It involves a technological system to drive creation and container of data, and this technological system can be modified to match the social and economic goals of its participants. For example, organizations have adapted blogging technology to include access control so that more sensitive internal discussions may be conducted in the informal manner made possible by blogs.

The technology available to search blogs stored in various sources, and to hyperlink between blogs and track who has authored as well as who has read and commented on blogs represent additional technological features which provide opportunities for the system to be more useful to the goals of this study. The technical characteristics affect how much can be leveraged from the system in both positive and negative ways. Blogs are URL addressable, searchable, and linkable, however the data is unstructured and hard to organize. The technology makes it easy to generate blog content, and it is obviously accessible from anywhere, with the data stored on the public Internet or within an organization's private intranet. Visualization technology allows for humans to process aggregate statistical information generated from usage of blogs.

At a social and organizational level, the participants are people from all backgrounds, who can be acting in an official capacity or in an informal or personal capacity, or simply be anonymous. The social issues inherent in this system are numerous, from individual identities, incentives to participate, who to trust, and how organizations can use it. Quite simply, the scale of this system – millions of blog postings in many domains – has made it challenging to be utilized for formal technology strategy or decision making by the organization.

Phase I – Bibliometric Blog Analysis

Phase I Blog Analysis: Key Points
- Method: bibliometrics and visualization
- Justification: large volume of information
- Contributes new method for bibliometrics allowing high frequency updates
- Preliminary results lay out solid plan
- Output: previously unknown trends

The energy industry was selected as the initial domain to run a preliminary study, based in part on the relationship to a project being done by the MIT-Masdar Institute collaboration. Environmental and supply concerns have forced the industry to search for new methods to create, store, and transfer energy. The resulting technology could be disruptive to current energy infrastructure, and significant R&D is required. Some may argue there is a low probability that the disruptive technologies will be invented by the organizations already involved in the domain, so it is vital for those organizations to be kept well aware of the growing areas of discussion.

Building on a set of initial bibliometric tools[26], an analysis of blogs in the public internet was conducted in the domain of renewable energy. A Python implementation was used to do HTTP requests to well specified URLs of search engines – in the initial case, Google Blog Search was used. The code takes as input a search term and retrieves the number of hits for a given query. This was extended to include blogs from Google blogs and will soon be extended to include blogs from Lexis Nexis as well. One piece of the code used is shown below as a sample:

```
      # Google blogs search
      def gen_googleblog_search (search_term, search_year=2007,
search_monthstart = 1, search_monthend=12):
      return
["http://blogsearch.google.com/blogsearch?as_q="+search_term+"&num=10&hl=en&ct
z=240&c2coff=1&btnG=Search+Blogs&as_epq=&as_oq=&as_eq=&bl_pt=&bl_bt=&bl_url=&b
l_auth=&as_qdr=a&as_drrb=b&as_mind=1&as_minm="+str(search_monthstart)+"&as_min
y="+str(search_year)+"&as_maxd=31&as_maxm="+str(search_monthend)+"&as_maxy="+s
tr(search_year)+"&lr=&safe=active",lambda x:re_func(x,"of about <b>(\S+?)</b>
for")]
```

In a future step, this code may be extended to one level of web crawling, to collect the blog post title, abstract, source blog, source author, and date for the most prevalent hits. Other web wrapping technologies such as Cameleon have been developed to specialize in the extraction of content and will be investigated[27]. In the abstract of the blog entry, it may be possible to extract any other terms in the ontology, and populate a database with this information to inform future searches.

Using this code, searches were conducted for clusters of keywords in the renewable energy domain, which were extracted from a study of emerging technology trends in the energy domain[28]. Since the blog data was only available on Google for the past four years, the search was altered to search for monthly results, to provide more resolution to the data. The search result data was fit into three growth curves to speculate on the relative growth patterns of each of the technology clusters. Results from this curve fitting exercise are shown below in a table – in the table, the variable 'a' represents the rate of growth when each cluster's data is fit to each type of curve.

CLUSTER NAME	Exponential			Logistic			Gompertz		
	a	Y_0	t_0	L	a	b	L	a	b
ENGINE	0.15	0	2	1.2	5.3	0.13	1.5	-2.3	-0.07
SOLAR+CELL	0.15	0	1	1.4	13	0.16	2.5	-3.5	-0.06
COMBUSTION	0.13	0.02	4	1.7	3.5	0.09	1.5	-1.7	-0.07
POWER+SYSTEM	0.12	0.01	1	1	12	0.19	2	-3.2	-0.07
BATTERY	-0.03	0.29	-7	0.45	2.8	14	2	-2.9	-0.06
HEAT+PUMP	0.16	0	0	6.5	13	0.03	0.89	-1	-0.05
PETROLEUM	0.17	0	1	2.9	9	0.08	2	-2.6	-0.06
COAL	0.16	0	2	1.1	4.5	0.13	1.5	-2.8	-0.07
FUEL+CELL	0.12	0	0	1.1	3	0.11	1.5	-2.3	-0.06
WASTEWATER	0.16	0	0	0.49	1.7	13	2	-2.7	-0.06

Table 1. Parameter estimates for each keyword cluster, based on three different growth curves.

Based on the data in the table above, each type of curve yielded a different ordering of rate of growth for the various technologies, as listed below.

- Exponential: petroleum, heat+pump, wastewater, coal, engine, solar+cell, combustion, fuel+cell, power+system, battery
- Logistic: battery, wastewater, power+system, solar+cell, engine, coal, fuel+cell, combustion, petroleum, heat+pump
- Gompertz: coal, power+system, combustion, engine, solar+cell, fuel+cell, wastewater, petroleum, battery, heat+pump

This initial attempt to rather grossly adapt bibliometric code meant for academic journal articles to internet blogs has shown that additional work is needed to yield more consistent and more useful results. In the analysis done using published academic literature, the three growth curves yielded much more consistent results[29].

The next step for the pilot study was to adjust the code for unique aspects of blogs. Blogs are known for a very high frequency of posts – unlike an academic journal which publishes every month or quarter, blog entries may come every hour or day. Thus, there are much shorter intervals between blog posts. Blog posts also have the potential to contain a much more varied semantic context than academic journals in a particular domain. Wheras a bibliometric analysis of journal articles in the renewable energy domain may not yield data points concerning topics such as fire trucks, cell phone batteries and diabetes treatments, we found these and many other topics included in our search of blogs and included them as 'different context' terms paired with their respective keywords.

Cluster Name	Keyword	Different Context	Cluster Name	Keyword	Different Context
battery	electrochemistry	phone	petroleum	asphaltene	
battery	lithium+ion	price	petroleum	resin	sell
battery	capacitor	support	petroleum	pyrolysis	
battery	electrode	price	petroleum	combustion	carbohydrates
coal	liquefaction	democrats	power+system	synchronous+machine	software
coal	gasification	trading	power+system	circuit	race
coal	coal+char	restaurant	power+system	lord	god
coal	combustion		power+system	motor	race

combustion	flame	software	solar+cell	photovoltaic	
combustion	turbulent	economics	solar+cell	silicon	implant
combustion	reaction+mechanism	kit	solar+cell	thin+film	movie
combustion	soot	fireplace	solar+cell	organic	food
combustion	kinetics		wastewater	pollution	news
engine	carnot+engine		wastewater	waste+disposal	news
engine	heat+engine	fire+truck	wastewater	textile+dye	
engine	thermoeconomics		wastewater	biomass	jobs
fuel+cell	proton+exchange		heat+pump	heat+pumping	baby
fuel+cell	membrane	diabetes	heat+pump	heat+transfer	recipe
fuel+cell	crossover		heat+pump	hysteresis	
fuel+cell	methanol		heat+pump	absorption	coffee

Table 2: Keyword clusters, and negative context terms, used in the searches

Thus, two factors were introduced: (a) Search type –searching just the title, searching the entire body, and searching the entire body but removing results which also contain a 'different context' term, and (b) time interval – weekly, monthly, quarterly and yearly. After running each of these 12 searches against each of the keywords, and then grouping results by each of the 10 clusters (for a total of 120 search replicates), a linear regression was done against the exponential model and the growth parameter as well as the adjusted R^2 value was calculated (based on the different N based on different time intervals). The growth parameter allowed us to assess the relative growth between clusters for a given search constraint, and the R^2 value allowed us to assess the relative impact our changing treatment of the search would have on the nature of the search results. Each of the 120 search replicates produced a graph similar to the two shown below, basically demonstrating the rising number of hits for a particular keyword cluster and a particular search type and time interval.

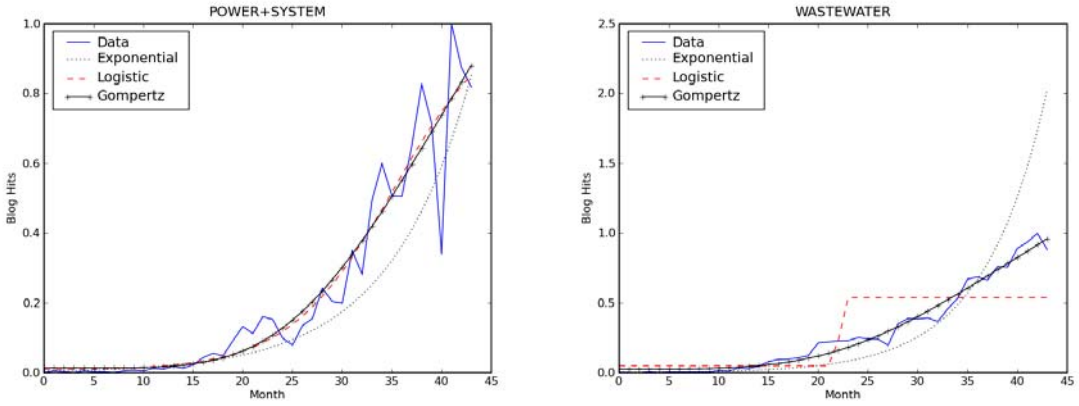

Figure 2: Blog hits by month curves for each cluster, compared to fitted growth curves

The next step in the analysis was to ascertain the impact of the two factors, search type and time interval, on the R^2 value of the various searches which were done, using a two factor ANOVA with 10 replicates per trial, where the 10 replicates are the searches done for the 10 different keyword clusters. The initial ANOVA analysis, using an alpha value of 0.05, suggested that both factors had a significant impact. However, visual inspection of the data showed erratic behavior that led us to re-run the ANOVA analysis with the two factors each constrained to the two most promising levels – weekly and monthly time interval, and body and context search type. The summarized results are shown in the table below and suggest that the time interval factor does have an impact on R^2 but the search type did not. This makes

intuitive sense because it is conceivable that searching by weeks instead of months would yield data that varied more significantly and did not allow for as nice of a curve fit, wheras the context treatment we applied in the search type was likely not significant enough to have an impact. In the planned study we will include more context constraint, with the involvement of stakeholders in the domain who can inspect frequent terms and remove if necessary.

ANOVA						
Source	*SS*	*DoF*	*MS*	*F*	***P-value***	*F crit*
Search Type	0.000255	1	0.000255	0.008355	**0.927676**	4.113161
Time Interval	0.377153	1	0.377153	12.33808	**0.001216**	4.113161
Interaction	0.000357	1	0.000357	0.011684	**0.914524**	4.113161
Within	1.100455	36	0.030568			
Total	1.478221	39				

Table 3. Summarized ANOVA results for two factors with only two levels.

Manual inspection of the content reveals an additional set of unique characteristics within blogs which need to be incorporated into the bibliometric analysis method. Blogs contain a less formal use of language, and thus terms require a higher degree of neighboring contextual terms to be included in the search. Blogs also include a higher degree of noise through repeated results and results from a different context, so negative filtering of results will be required. Finally, blogs do not require a peer review or high cost of publishing, and thus the value of the information in blogs is more a gauge of public interest than true acceptance by a technical community within the domain, unless the notion of blog author context is introduced into the methods.

More advanced treatment of the blog information will result in altered code being written to produce different sets of data which will be more useful in representing the level of interest. For example, we can consider only searching for blogs which include the term multiple times, thereby indicating an actual interest in the topic rather than a passing reference. The blog comments can be searched indepedently of the blog content, as another means for assessing whether the topic of the blog is the actual term being referenced.

Data splitting will be utilized to further validate the results. We have the opportunity, given clusters of keywords representing a particular topic, to split the data in creative ways to validate the data. First, data can be split by removing every other month in the dataset and assessing to what degree the regression built on this set predicts the other set, and vice versa. We can also take a random set of the keywords in each cluster, remove data for these particular keywords, and then assess whether the regression built for the overall set predicts the rate of growth for the removed keywords.

From a prediction standpoint, the goal is not to predict the actual future occurrences of the keywords in future blog posts. Instead, it is to predict a relative rate of growth, in order to inform a relative shift in investment from the stakeholder. The data collected in this study is simply one input into the decision making process of the stakeholder. The contribution of this work is in the information extraction and analysis techniques which allow a significant volume of new information to be utilized in the decision making process.

The next steps also include introducing a ‘control’ to the experiment, to be able to benchmark the growth information against the entire blog database or at least a subset related to the particular domain or the particular area of professional blogs. A logical next step in the technology forecasting domain would be to develop a probabilistic model which would model how likely a technology cluster is to grow based on

predictions from the blog data. Further consideration is also required in the area of context awareness, to consider aspects such as professional vs personal, semantic context, and post length as determinants of context. Stakeholders must also be involved to refine keywords used for clusters, and to discuss if this is even the right way to conduct the searches. Finally, future consideration must be provided towards the impact of adding or deleting keywords in a cluster, and handling duplicate search results when the goal is studying relative rates of growth.

Long tail effects are another consideration. While it is interesting to see aggregate characteristics and identify relative weight between clusters, it is also relevant in an environment such as blogs to identify particular areas which are receiving a small but growing amount of interest – analytical methods such as what was described here could serve as early warning indicators for growing interest in a given area. Based on the results of the study, it may be the case that advanced visualization rather than detailed quantitative analysis may be the more useful output of this work. For example, the ManyEyes research project has a set of advanced visualizations which take data from an internet accessible feed and provide a graphic representation which can be modified by the user[30]. The Many Eyes system also allows multiple users to set different perspectives on the data and hold a discussion with each state of the chart saved. Visualization can also aid in identifying outliers, which may be the result of particular news events that impacted blog interest – unlike traditional experimental data which often disregards outliers, we expect the inspection of outliers to be a particularly interesting aspect of this analysis. The original goal is to provide an automated way of leveraging information within a large scale information system, and this may still be met without stringent quantitative analysis being performed. The survey will be used to assess whether either the statistical representation or the visual representation were useful in adjusting the system representation in a way that influenced technology strategy choice.

Already the work we have described represents a significant body of research. However, it is also the intention of this project to examine the broader implications of our techniques when integrated into the over-arching information and innovation system of the organization. This issue will be addressed in the following pilot study.

Phase II – Dynamic System Re-Representation

Phase II System Rep: Key Points
- Method: CLIOS system representation
- Justification: articulates nested complexity
- Contributes new method for re-representation based on new information
- Preliminary manual representation demonstrates potential of automation
- Input: Identified trends from phase I
- Output: Re-representation based on trends

The CLIOS Process, developed by Prof. Joe Sussman and his team at MIT, is a process for studying Complex, Large, Interconnected, Open, Sociotechnical Systems[31]. A number of MIT Engineering Systems Division thesis have been based on this process and it provides a useful framework for dealing with systems such as the ones which this study will inform. The process focuses on systems with the nested complexity of a physical domain set within a sphere of institutions. Systems such as this align very well with the goal of this research, since we are looking not only at the information content itself but the social context of the information – who posted in the blog, and which institution they are representing. As described in the teaching note from Sussman et. al., the CLIOS process provides a framework for interested parties to see their viewpoint of a system in the context of the entire system. This is another point of alignment with the research proposed here, as a key characteristic of blogs is that the cost of publishing is low so posts are often made without context.

In this study, the system representation phase of the CLIOS process is being used simply because it is the most geared towards articulating nested complexity and understanding various strategic alternatives. We are not using the entire CLIOS process and would not require the surveyed stakeholders to use the entire process to select the strategic alternative their division must invest in - they may do so if they find it

useful, but for the scope of this work, an actionable representation is the key requirement. An actionable representation is a representation which leads to a particular action which would not have occurred without the representation exercise. CLIOS system representation splits technological subsystems from the institutional sphere, and this is particularly useful because the core work in new methods of information extraction and analysis depend on the carefully constructed choice of keywords related to various alternatives in a given domain, and a representation which aligns with this construction is particularly useful. The institutions in question, and the various policy levers which are used by the institutions to affect various technologies, are important to separate in the representation because this lends itself well to attacking the new information source of blogs -- we can search for blogs from a particular institution, look for references to particular institutions related to particular keywords, and in blogs which have very succinct and focused units of information posts, the discussion of policy levers is often distinct from a discussion of elements of the technological subsystem. Thus, the nature of our information source and our core contribution around new methods of information extraction are closely aligned with the core tenants of the system representation phase of CLIOS.

Another important consideration is the audience we are addressing in this research. Since we are applying these methods in the context of a technology firm with surveys to 1700 employees across the organization, we need to consider how these individuals make decisions in their day to day job. At this level, individuals often spend a significant portion of focused effort on the technological subsystems (with some consideration of the policy drivers coming from institutional actors such as regulatory bodies), and then only at a periodic basis and in certain roles do they take time to look at the complexity introduced by the institutions which are affecting their particular technology domain. If we were dealing with government stakeholders who are policy makers, the requirement would be different. Therefore it is necessary to utilize a system representation framework which separates out the various levels of complexity in a manner which matches the audience and the domain for the research.

It is important to note that as we are purely focused on the system representation, and not a model, this endeavor is by nature descriptive, and not normative or prescriptive as defined in the literature [32]. In this research, the goal is to produce a descriptive system representation - using information extraction and analysis to describe the state of the world as it is, not as it should be. There is a significant volume of information being produced on the Internet and our goal is simply to make that volume of information accessible to decision makers by describing it in a way that it can be acted upon. It is beyond the scope of this research to speculate on the normative aspects of the representation (i.e., what topics a standard set of blog posts should cover), or a prescriptive aspect of the representation (which would potentially lead to a model being used to prescribe action or behavior).

The first phase of this process is system representation, developed by a long and careful interview process with many stakeholders involved in the CLIOS system. The key attributes of a CLIOS system are technological subsystems made up of components with common drivers which go between the subsystems, and an organizational or institutional sphere which provides a context for the technological subsystems. Policy drivers and external factors from the institutional sphere impact the behavior of the physical subsystem. The value of the CLIOS process is that it provides a means for selecting between various strategic alternatives by modeling the impact each alternative would have on the system. In this research, the representation of the system will be informed by the analysis of information contained in blogs. New topics being associated with components already in the system may be considered for whether they represent new components which should be in the system. If a significant increase in activity is seen in a certain topic area, this may also indicate that a new external factor should be introduced, or that a policy decision has been made which has created a new policy driver for the system.

A pattern exists in Carlos Osorio-Urzua's dissertation for extending the CLIOS system representation phase, and building a derivative model by utilizing a complementary method which provides a more

refined representation [33]. The study being reported here will use this dissertation as a pattern for how to augment the CLIOS system representation phase. Specifically, the output of the quantitative analyses being done in this study would be aggregated into an effectively qualitative stakeholder which could be used just as other human stakeholders are used in the system representation phase. Just as human stakeholders are informed by numerical data, this virtual stakeholder would provide a daily update to those representing the system, and identify key trends, new topics, individuals or institutions which may be considered in the system representation. The value of this new stakeholder is the ability to have the system representation adapt to change on a much more continuous basis, by using data collected by blogs to suggest new relationships between technological subsystems, or new institutions which may be considered in the institutional sphere. The reach of this stakeholder is far more global, comprehensive and more frequently up to date. The visualization techniques mentioned earlier will help better articulate the results, just as a human stakeholder would provide a prose delivered context to quantitative data.

This augmentation to the CLIOS process would yield both a methodological and domain contribution, just as the Osorio dissertation had done. The opportunity exists to make a methodological contribution by extending bibliometrics to understand the blogs system and using this to extend the CLIOS process. The concept of dynamic system representation would be introduced, as discussed below. The opportunity also exists to make a domain contribution – this study can focus examples of the above method and supporting code around a particular domain such as renewable energy, and reveal specific insights about this domain.

The concept of "dynamic system representation" is that a CLIOS system representation which normally takes significant stakeholder interaction to produce may be informed hourly or daily by new information sources. The goal is not to actually change the system representation on a dynamic basis, but instead to treat the output of the quantitative analysis as a qualitative stakeholder, similar to how any other qualitative stakeholder would themselves use quantitative data to provide input. Certain domains are moving at such a rapid pace that incorporating new sources of information dynamically will be very beneficial. Information sources could reveal new institutions, new individuals, new geographies, which have interest in the particular CLIOS.

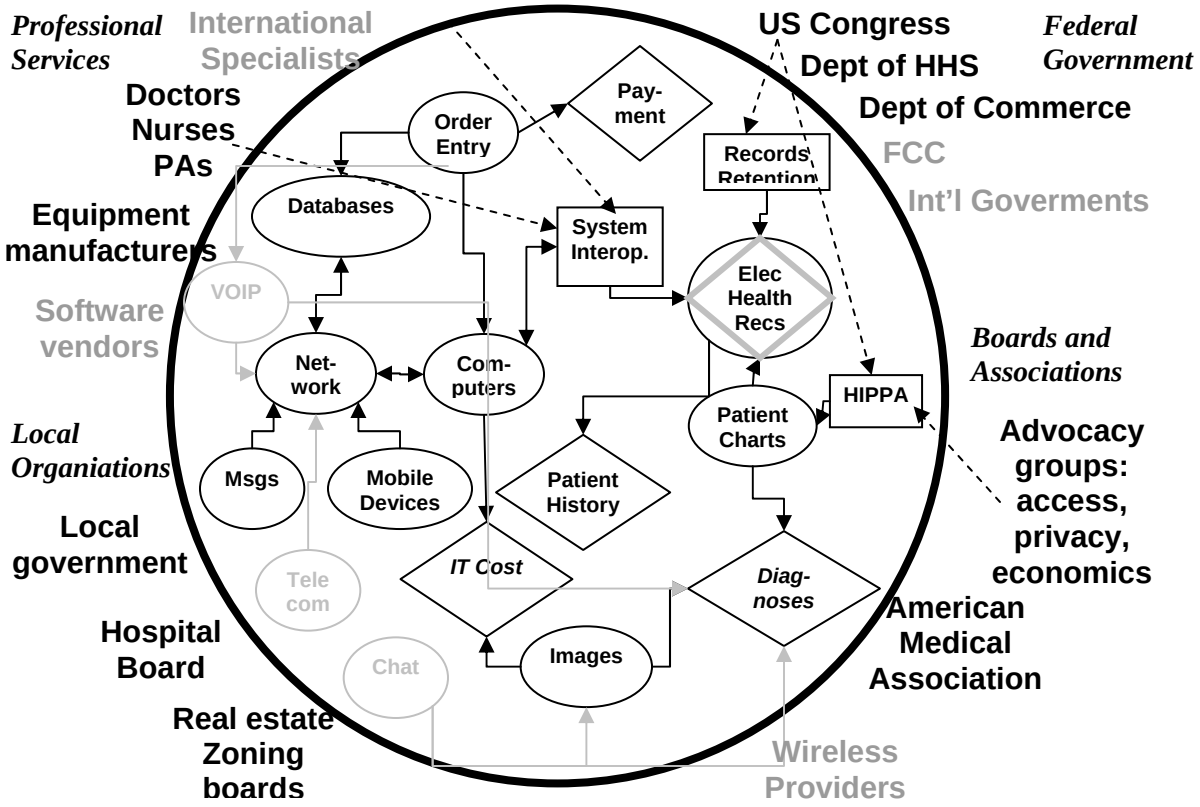

Figure 3: Dynamic re-representation of a CLIOS system.

The above figure is the Information Technology physical subsystem which is one of three subsystems of the Healthcare IT system (the other two are economic cost structure and healthcare delivery), with actors in the institutional sphere shown outside the circle. For details on the CLIOS process and nomenclature, the reader is referred to the CLIOS teaching note referenced in this paper. For the scope of this report, the reader must understand simply that ovals represent regular components, rectangles are policy levers used by institutional actors, and diamonds represent drivers common to multiple physical subsystems. Elements in dark color represent an initial system representation possible by existing methods. Elements shown in lighter color are those which could be added based on insight gained through an analysis of blogs within this domain (in this case demonstrating the impact of increased focus on emerging IT opportunities within the healthcare field). Note that certain new institutions are found, along with new technological components, and the role of existing components in the system can also change (in this case electronic health records become a common driver between multiple subsystems).

Phase III – Using 'Social Context' to Apply Findings

Phase III Social Context: Key Points
• Method: Context interchange
• Justification: aggregate heterogeneous info
• Contributes new method for extending the notion of context to include info. contributors
• Preliminary manual walkthrough using IBM data set shows the potential
• Input: Re-representation based on trends
• Output: Organization's capacity to react

The next natural step in analyzing information to inform the organization is to understand the information within the organization. Organizations such as IBM have begun to introduce blogging as an internal mechanism for sharing ideas and information. However, within the organization, the notion of which people are the sources of information is as important as the information itself. The organization is attempting to understand their own capacity for addressing the various strategic alternatives, and must understand the individuals present within the organization. Blogging provides a

unique linkage between people and information because it may reveal sources of interest and expertise which are not represented in the formal job descriptions stored in a company's employee directory.

We forge this link by introducing a methodological advancement to the notion of "context" developed at MIT Sloan [34]. Context refers to the definition of how particular data should be interpreted, based on a set of semantic definitions in a particular data model. The definitions include the semantic type of particular information units and can mediate across different use of the same vocabulary (for example, the notion of "tall" or "cheap" or "price") and different numerical parameters (for example the underlying currency of financial data, or the units for financial data). Context is associated with both the information source (such as a financial database), and the information receiver (such as an accountant, performing a query against the financial database). Context can come from the geography, company, function or any other environmental factor which describes where the information source or request comes from. The concept of context is most useful when attempting to aggregate information across sources which each use their own context and require mediation to effectively execute a query which extracts information across these sources.

The same method described earlier will be applied to blogs inside the organization. However, an additional step will be introduced, to incorporate "social context" by identifying the role and position of the individuals who are highly active in a certain topic area. This can be done provided the organization has a standard employee directory which can be queried, for example using the Lightweight Directory Access Protocol (LDAP) to gather specific information. This provides information on the individual's group within the organization, their position in the hierarchy, and the location. Prototype software to navigate the employee directory of IBM has been developed in conjunction with a class project done in the Advanced Systems Architecture class within MIT's Engineering Systems Division.

This project will extend the notion of context to include social context as a primary aspect of the machine processing of information. Social context is defined simply as knowledge related to the individual or organization producing the information as well as the individual or organization requesting the information. This matches the established notion that context is something that must be understood both from the information source as well as the information requestor. Examples of social context would include the professional title, expertise, interests and connections of the individual. For example, if an engineer in an organization performs a query for a specific topic against a knowledge repository, the social context of the requestor includes the fact that they are technical and they perform product development for the organization. Sources are then sorted in relation to whether the individual contributing the source is themselves technical, and involved in product development. Additional weight may be placed to sources which come from individuals who are ranked higher in the organization, and who have (in the case of blogs) many comments associated with their posting. If it were a marketing individual making the same query, or even a technical individual who was in the research side of the organization, the same sources would be processed in a very different manner. It is entirely conceivable that these assumptions about the requestor's goals are wrong (the engineer may very well want to see posts of a marketing nature), but by introducing this notion we allow the requestor a mechanism for articulating these preferences.

This notion diverges from the traditional view of context because it allows the same data within the same data source to have different context. In the traditional view, if a database exists with a particular semantic context defined for particular fields, if two individuals enter the same data values into that database, a query for that data (no matter what the requestor's context is) should behave the same. Furthermore, over time the context of the same data can change as the contributor changes their social context. An individual who is promoted, or switches from an industry executive to an academic professor, would cause the data that individual has entered to be processed with a different context. Another point of potential divergence is that the traditional form of context is commonly (but not exclusively) used for

aggregating numerical information, where data formats and scale are highly relevant, while in this case the primary form of data to be consumed is textual in nature. That being said, the ability to understand formats such as dates can be relevant to social context as well – for example, a researcher looking for "recent" posts may mean posts over the last ten years while a sales individual with the same query may only be interested in posts from the last month.

While appearing divergent to some degree, I propose this as an extension because the same principles and mechanisms for context interchange apply here. Context must be well defined, both from the source and requestor standpoint. Context will allow information from multiple sources to be aggregated – such as blogs, press releases and patents – with an increased understanding of the social context underlying the information. Intermediate data sources are likely to be required, to mediate context (for example, an employee directory could be our "currency converter" in this case). Another form of intermediate data source to mediate would be a database which mapped terms in a domain to context. For example, a biologist searching for "cell membrane" expects very different results than a mechanical engineer, and a database which contained a mapping to the term "fuel" (which could be added to the query to become "fuel cell membrane") would help constrain the query to the appropriate context. Associated technologies such as web wrappers for semi-structured web sources also remain entirely relevant to the process of making blogs and other social information machine processible.

Initial Walkthrough Using the IBM Data Set

An initial walkthrough of the use of social context is provided below, for the data set within IBM. Three scenarios were investigated from the research, marketing and product development perspective and due to space considerations, only one is provided here, from the research perspective. A manual walkthrough demonstrates how this research would affect the analysis of information within an organization.

Social Context Demonstration based on IBM data set

Data

- Profiles: Name, role, organization, division, team, tags, connections
- Bookmarks: URL, name, description, tags, date
- Blogs: Name, post subject, post content, tags, date, location, comments, commenting individuals

Current State Description	**Desired State Description**
Query for 'social network research' • Must be done in multiple databases • Results not sorted by any useful means • Unclear whose posts should be trusted more • Multiple clicks to find relevant assistance • Multiple search terms attempted Result: Inefficient use of disparate sources with no context for pieces of information	Single query for 'social network research' • Profiles system searched first • Users with social-network tag • Users with seniority in the organization • Users with high density of blog posts tagged with social-network • Use # comments to validate self-made experts • 'Related People' as another option to validate • Bookmarked URLs and blog postings tagged with tags of person contributing the posting • Results ranked by tags first, then by person • Trusted individuals shown first and experts quickly identified

Manual Walkthrough Simulating Desired State

Question: Find three references which IBMers have found useful in conducting social network research?

- Go to Communities, search for 'social network research'
- Receive 21 results. Analyze.
  - First three are related to Lotus Connections product
  - Fourth is in German language
  - Fifth, sixth are research based but have 7 and 2 members, and 3 and 1 posting (not useful)
  - Seventh and eighth are focused on marketing
  - Ninth is cross-functional, includes research, but is related to different topic
  - Others are ruled out by visual inspection
- Go to Profiles, search for 'social network research'
- Receive 83 results. Analyze based on job role and organization
- Visual inspection reveals 7 of top 10 are sales. Others are 2 IT specialists and 1 strategy.
- Continue to inspect profile results beyond top 10
  - Result 14 is Peter Schuett (*Leader for Knowledge Mgmt and Social Networking Solutions* in Germany
  - Result 17 is Luis Suarez (*Knowledge manager, community builder, social computing evangelist)*
  - Result 28 is Carol Jones, IBM Fellow. Selected b/c previous communications w/her, known expert
  - Result 30 is Jeff Schick, VP of Social Software, selected for level in hierarchy, although not researcher
  - Result 33 is Eric Wilcox, researcher in social software – first official researcher found.
  - Result 36 is Larry Proctor, researcher, but rejected because in services not social networking
- Results up to 83 do not yield anything more – curious why no one in Lotus Research showed up, so traversed organizational chart starting with Irene Grief (known to be IBM Fellow and director of research) and found Kate Ehrlich, profile was empty
- Go to Dogear (Bookmarks), search for 'social network research'
- 1 result – RIT Lab for Social Computing
- Review tags starting with social: social, social-software, socialnetworking, social-networking, social_networking, social-computing, socialsoftware, socialbookmarking, socialcomputing, social_bookmarking, social-network-analysis, social-network-anlysis
- Search for social-network-analysis, 5 results, 1 from research lab, paper on Analysis of Social Networks
- Search by person, using list of people generated from profiles list
- Visually inspect tags for each of the 6 individuals
  - Peter Schuett uses "socialnetworking" – 2 bookmarks, one is Irene Greif on Social Networking
  - Luis Suarez uses "social-networking" (248 hits) and social_networking (95 hits) – too many to inspect

Result

When Manual walkthrough halted, three references found, total time spent = 3 hours

- If further walkthrough was done, additional references could have been found
- One ref per hour = only 40 references in a work week

Validating the Method

Validation via Survey: Key Points
• Method: tailored survey, internet delivered
• Justification: validates utility of the research
• Preliminary survey design has detailed plan
• Input: Results of research, for ten domains
• Output: Usefulness of research for IBM

Quantiative analysis produced by this pilot study will be validated by a stakeholder survey with members of IBM's Horzonwatch program as described above, along the lines of the survey validation conducted in Courseault's thesis. Historical evaluation has been proposed as another evaluation – running the analytical tools on past data to reproduce a known outcome – and this method may be considered but is not fully developed by the researcher and thus not presented in this report. Frey's notion of an "animal test" analogy applies to some degree here, where mice are chosen to represent humans in validation[35]. The choice of IBM as the survey target is in part due to accessibility and to the predisposition this community has to using blogs as an information source.

Survey Design

The survey will be based on Dillman's tailored survey design method [36]. .The proposed project calls for a two part survey, administered first to gain an understanding of the current state and needs and then to assess the merits of the tools and data analysis provided. The respondents will be split among two groups – half will be from a set of individuals self-selected into the IBM Horizon Watch population, and half will be selected purposefully from the broader population, based on a two-stage survey to assess those exhibiting characteristics of decision making influence and interest. Both groups will be stratified between the ten domains that were chosen earlier. We also considered stratification at the geography and job role level. Since the goal is to tie use of blogs to the overall technology strategy, we need to consider the variations among different business units which each have a particular technology strategy rather than the variations between other attributes of individuals within the process.

The survey mode to be used is an embedded internet survey within an electronic mail, with followup provided by instant message (analogous to phone followup). Rich electronic mail clients such as Microsoft Outlook and Lotus Notes allow for electronic mails to be more than text and include interactive content. A set of visual representations of the survey replies will be updated automatically, so that respondents can immediately see what others have said in the same topic once they have replied or chosen not to reply. The researcher will have a continuous visual view of data to modify respondent samples.

The unit of analysis in these surveys will be the division. The goal is to determine the usefulness of various data analyzed from public blogs in making strategic decisions on the division level. This unit was chosen because examining the unit of analysis at the firm level would have been too broad and incorporated too many factors to have a single survey capture the true impact of the small factor of blogs. The unit of an individual would have been too fine grained – individuals do not make strategic decisions for themselves, they simply provide input for decisions made at the division level. The other unit considered was the product team unit, and this was determined to be too fine grained also, because a strategic shift from one technology to another has the potential to discontinue one team and create another. Thus, divisions with a pure focus in one given area of the industry were chosen as the unit of analysis. It is certainly possible that the data collected may not be truly representative of the division's view on the matter, especially since this is an unknown area and thus personal opinions may dominate results, however the core technology decisions in a division are ultimately made by individuals so the survey method matches the decision making process.

A major piece of our conceptual model is the notion of a decision maker and a decision. At the divisional level, there are often many decision makers at varying levels of influence. The goal of this study is not to focus too greatly on identifying exactly when a decision is made, and who made the decision. Instead, in this study we recognize that many incremental decision are made which overall influence higher level

choice of technology strategy. With this mindset, our decision maker can be any self identified individual in the company who feels they acquire knowledge from a variety of sources and use that knowledge to make a selection of which technologies among a given set of technologies they will invest in, or advise investment in. It is critical as part of our instrument to help our respondents identify themselves as decision makers and understand this conceptual model which we have developed.

After building the analytical system and testing it via dynamic system representation in a variety of domains, the information will be available publicly and a survey will be administered to gather feedback on the results. In this survey we have considered whether to ask about specific analysis produced for a specific individual, or whether we should present all results and ask for general feedback. We chose the latter approach because we do not want the specific results in a domain to influence the respondent's thoughts on whether this type of system is generally applicable. Other questions will relate to the specific information sources currently used to make technical decisions, the specific individuals or institutions consulted, and the timeframe in which decisions are made.

Project Milestones and Timeline

Date	Milestone	Details
1/08-5/08	Prototype software for blog analysis	Use specific domain from Masdar project (energy) Search Google blog database by month Output: statistical measures for key terms' growth and noise
06/08-08/08	Identify ten domains of focus	Interview / survey 1700 IBM Horzonwatch members and identify ten domains Document current processes used and gaps in current knowledge
05/08-08/08	Write / revise introductory chapters	Goals, Motivation, Hypothesis, Literature review, Methodology
07/08-09/08	Software analysis of blog data for domain	Identify keyword clusters for each domain, execute searches Develop and refine method, compare to published papers
10/08-12/08	Begin data analysis chapter	Focus on external data first
06/08-12/08	Manual CLIOS representation of ten domains	Use domain literature and stakeholder interviews to represent system as CLIOS Validate initial representations with ten domain stakeholders
01/09-04/09	Develop dynamic system representation	For each system, map terms to policy drivers, functional components, institutions Apply results for five year period, produce qualitative output
04/09-05/09	Write system representation method chapter	Document case study and describe as generalized method
05/09-08/09	Validation of external data with stakeholders	Set of qualitative interviews, survey to 1700 members of group Provide real-time representation of the system, document insights gained
07/09-09/09	Analysis of internal profiles / capacity to react	Use role & expertise to rerank results, identify relationships Analyze relation of individuals to use of information sources
09/09-11/09	Validation of internal data with stakeholders	Discover whether new relationships were found
11/09-1/10	Write/revise data analysis and case studies	Complete documentation of all data collected, software used, methods implemented and domain specific lessons
1/10-2/10	Write contributions, future work, conclusions	Document contributions to the field of engineering systems
2/10	Complete first draft of dissertation for revision	Based in part on feedback from various stakeholders and committee
5/10	Defense	

References

1 Lenhart, A. and Fox, S. “Bloggers.” Pew Internet & American Life Project, Jul. 2006
2 Woon, W. L. and Madnick, S. <Insert reference to working paper or project proposal?>. 2008.
3 Vantage Point, http://www.thevantagepoint.com
4 “Thomson Scientific Announces Strategic Alliance with Collexis to Develop Custom Data Mining Solution for Web of Science Users,” Thompson Reuters press release, Feb. 21, 2008
5 Roberts, E.B. “Exploratory and Normative Technological Forecasting: A Critical Appraisal.” *Technology Forecasting*, 1, pp113-127, 1969.
6 Koh, H., Magee, C. “A functional approach for studying technological progress: Application to information technology.” *Technological Forecasting and Social Change*. 73, pp 1061-1083, 2006.
7 Makridakis, S. Hibon, M. Moser, C. “Accuracy of Forecasting: An Empirical Investigation.” *Journal of the Royal Statistical Society*. 142 (2), 1979.
8 Hauptman, O., Pope, S.L. “The Process of Applied Technology Forecasting.” *Technological Forecasting and Social Change*. 42, pp 193-211, 1992.
9 Daim, T. U. et. al. “Forecasting emerging technologies: Use of bibliometrics and patent analysis.” *Technological Forecasting and Social Change.* 73, pp 981-1012, 2006.
10 Morris, S. et. al. “DIVA: A Visualization System for Emploring Document Databases for Technology Forecasting.” *Computers and Industrial Engineering*. 43, pp 842-862, 2002.
11 Mann, D.L. “Better technology forecasting using systematic innovation methods.” *Technological Forecasting and Social Change*. 70, pp 779-795, 2003.
12 De Miranda Santo, M., dos Santos, D. M. “Text Mining as a valuable tool in foresight exercises: A study on nanatechnology.” *Technological Forecasting and Social Change*. 73, pp 1013-1027, 2006.
13 Courseault, C. “A Text Mining Framework Linking Technical Intelligence from Publication Databases to Strategic Technology Decisions”. PhD Thesis, Alan Porter (Advisor), Georgia Institute of Technology, May 2004.
14 Cachia, R. et. al. “Grasping the potential of online social networks for foresight.” *Technological Forecasting and Social Change*. 74, pp 1179-1203, 2007.
15 Hookway, N. “Entering the blogosphere: some strategies for using blogs in social research.” *Qualitative Research*, 8 (91), 2008.
16 Kumar, R., et.al. “On the Bursty Evolution of Blogspace.” *World Wide Web*. 8(2), 2005.
17 Glance, N.S., Hurst, M. Tomokiyo, T. “BlogPulse: Automated Trend Discovery for Weblogs,” In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004, at WWW’04: the 13th international conference on World Wide Web*, 2004.
18 Kaye, B.K. “It’s a blog, blog, blog world: Users and users of weblogs.” *Atlantic Journal of Communication*, 13(2), pp 73-95. 2005
19 Mishne, G., Glance, N. “Leave a Reply: An Analysis of Weblog Comments.” ,” In *WWW 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006, at WWW’06: the 15th international conference on World Wide Web*, 2006.
20 Chen, Y. Tsai, F.S., and Chan, K.L. “Blog Search and Mining in the Business Domain.” In the *2007 ACM SIGKDD Workshop on Domain Driven Data Mining*. ACM, 2007.
21 Anjewierden, A. and Efimova, L. “Understanding Weblog Communities Through Digital Traces: A Framework, a Tool and an Example.” In Proc. of the OTM 2006 Workshops, LNCS 4277, pp 279–289. Springer, 2006.
22 Breslin, J.G. et. Al. “Towards Semantically-Interlinked Online Communities.” *The Semantic Web: Research and Applications*. Springer, 2005, pp 500-514.
23 Groves, R. M., et. al.*Survey Methodology*. Hoboken, N.J.: John Wiley and Sons. 2004.
24 Interview with William Chamberlain, Director, IBM Horzonwatch, January 2008.
25 “Engineering Systems Research and Practice.” ESD Symposium Committee, 2002.
26 Woon, W. L. <Reference technical paper from March 2008?>
27 Firat, A., Madnick, S., Siegel, M. “The Cameleon Web Wrapper Engine”, Proceedings of the VLDB Workshop on Technologies for E-Services, Cairo, Egypt. 2000.
28 Kajikawa, Y., et. al.. “Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy.” *Technological Forecasting and Social Change*, In Press, Corrected Proof.

29 <Technical report>
30 Viegas, F.B., Wattenberg, M., van Ham, F., Kriss, J., McKeon, M., "ManyEyes: a Site for Visualization at Internet Scale," *IEEE Transactions on Visualization and Computer Graphics*, 13(6) 2007.
31 Sussman, et. al. "The CLIOS Process: A User's Guide." MIT Teaching Note, April 26, 2007.
32 Valerdi, R., Ross, A., Rhodes, D. "A Framework for Evolving System of Systems Engineering." *CrossTalk*. Oct. 2007.
33 Osorio-Urzua, C. "Architectural Innovation, Functional Emergence and Diversification in Engineering Systems." Ph.D. Thesis, MIT Engineering Systems Division, 2007.
34 Bressan, S., Goh, C., Levina, N., Madnick, S., Shah, A., Siegel, M. "Context Knowledge Representation and Reasoning in the Context Interchange System." *Applied Intelligence*. 13, pp 165-180, 2000.
35 Frey, D.D., Dym, C.L. "Validation of design methods: lessons from medicine." *Research in Engineering Design*. 17 pp 45-57, 2006.
36 Dillman, D. A. *Mail and Internet Surveys; The Tailored Design Method* (Second Edition) Hoboken, N.J.: John Wiley and Sons. 2007.