DIMACS Workshop on Data Processing on the Web : A Look into the Future

Organizers:

Mary Fernández	AT&T Labs - Research	`mff@research.att.com`
S. Muthukrishnan	AT&T Labs - Research	`muthu@research.att.com`
Dan Suciu	AT&T Labs - Research	`suciu@research.att.com`

Editorial Assistant:

Srikrishna Divakaran	Rutgers University	`divakara@paul.rutgers.edu`

March 24, 2000

1 Workshop Focus

Processing Web data poses new challenges for database and information-retrieval technologies. Among these challenges are: the amounts of Web data; its relevance ranging from useless to highly useful sources; its heterogeneity ranging from plain text, to structured documents, to sounds and images; the way data is generated, from static HTML pages to database queries, and its unpredictability. Despite these challenges, today we can search the entire Web, mine significant portions of it, integrate data from multiple Web data sources, and query Web data. New technologies, such as those based on the XML standard, are aimed at enterprise-wide Web applications and data integration. Future applications will probably go much further. In this two-day workshop, leading researchers at the forefront of database and Web research gave their visions of the future of data processing on the Web.

2 Search & Representation

"The Web Graph: Structure and Interpretation"

Dr. Sridhar Rajagopalan, IBM Almaden Research Center.

Author abstract: The study of the web as a graph is not only fascinating in its own right, but also yields valuable insight into web algorithms for crawling, searching and community discovery, and the sociological phenomena which characterize its evolution. We report on experiments on local and global properties of the web graph using two crawls each with over 200M pages and 1.5 billion links. Our study indicates that the macroscopic structure of the web is considerably more intricate than suggested by earlier experiments on a smaller scale.

Summary: Dr. Rajagopalan considered viewing the Web as a directed graph, with each vertex representing a page and each reference between pages representing an edge. For a long time the global topological structure of the Web graph was unknown. Dr. Rajagopalan reported a breakthrough result characterizing this structure. The graph consists of four components of about equal size: one strongly connected component (robustly identified as the largest strongly connected component in the graph), an "out" component containing all pages accessible from the connected component, an "in" component containing all pages that can access the connected component, and finally a set of pages that cannot be classified in any of the previous three categories. In addition, Dr. Rajagopalan reported several local topological properties of the Web graph:

- The probability distribution function for indegree and outdegree and reachability of the web graph is a Zipf distribution. As a consequence the Web cannot be modeled as a standard random graph (for which these distributions would be exponential), and a better model is needed.
- Web communities can be identified easily and robustly by checking the occurrence of the $K_{3,3}$ subgraph.

Finally, Dr. Rajagopalan presented a new graph theoretic model of the Web, as a graph obtained by random additions of links with a certain copy probability. The local and global topological properties of the Web graph are quite faithfully captured in the new model, showing a significant improvement over standard random graphs.

Speaker Biography: Sridhar Rajagopalan (*sridhar@almaden.ibm.com*) received a B.Tech. from the Indian Institute of Technology, Delhi, in 1989 and a Ph.D. from the University of California, Berkeley in 1994. He has been a DIMACS postdoctoral fellow between 1994 and 1996. He is now a Research Staff Member at the IBM Almaden Research Center. His research interests are algorithms and algorithm engineering, randomization, information and coding theory, and information retrieval issues on the web.

"WHIRL: A Formalism for Representing Web Data"

Dr. William Cohen, AT&T Labs - Research

Summary: Data on the Web is hard to represent with conventional knowledge-base and database formalisms, due to problems like terminological differences across sites, and the frequent interleaving of textual information with structured, data-like information. In this talk, Dr.Cohen presented a new "information representation language" called WHIRL that is used not only to represent semi-structured information but also integrate information from various places on the web.

WHIRL addresses some of the problems like representing semi structured data and name resolution by incorporating ideas ideas from both AI knowledge representation systems and statistical information retrieval. Specifically, WHIRL is a subset of Prolog that has been extended by adding special features for reasoning about the similarity of fragments of text. Rather than exploring a number of sites to get information you want, WHIRL enables you to simultaneously query a number of related sites using a database-like interface. The highlighting feature of WHIRL is a new sort of "soft" database system, which uses text-matching methods developed in information retrieval to approximately match names.

Dr. Cohen also presented some results on his work on the use of WHIRL for collecting Web data for collaborative filtering and machine learning systems.

Speaker Biography: William Cohen (`http://www.research.att.com/ wcohen/`) received his bachelor's degree in Computer Science from Duke University, and a Ph.D. in Computer Science from Rutgers University. Since 1990, Dr. Cohen has been employed at AT&T Labs. His main research area is machine learning, and his present research interests include information integration, text categorization, and learning from large datasets.

"Musings on the Extraction of Structure from the Web"

Dr. Rajeev Motwani, Stanford University

Author abstract: One of the major challenges in dealing with the web is the unstructured or semistructured nature of the data. There are major benefits in extracting the structure implicit in the web or extracting a subset of the data that can be structured easily. I will give a personal view of some attempts in this direction. The focus will be on the problems and proposals for attacking these problems.

Summary: The world wide web provides a vast source of information of almost all types of data. However, this information is distributed among various systems using many different formats. If these chunks of unstructured or semi-structured information could be extracted from the web and integrated into a structured form, then they would form an unprecedented source of information.

There has been considerable work on integrating a number of information sources using coded wrappers or filters. However, thes can be time consuming and will not scale well when the number of sources are greater than 10. In this talk, Dr. Motwani addressed the problem of extracting a relation from the various sources in the web that may hold pieces of the relation. The objective of Dr. Motwani's approach was to discover infoirmation sources and to extract the relevant information from them either entirely automatically, or with very minimal human intervention. In this talk, Dr. Motwani considered the problem of extracting a relation of books - (author,title) pairs from the web. He presented the basic idea and some theoretical and experimental results.

Speaker Biography: Rajeev Motwani (`http://Theory.Stanford.EDU/ rajeev`) is an associate professor of computer science at Stanford University, where he also serves as the director of graduate studies. He obtained his Ph.D. in computer science from University of California, Berkeley in 1988, and his BTech in computer science from Indian Institute of Technology, Kanpur in 1983. His research interests include: databases and data mining, web search and information retrieval, robotics, and theoretical computer science. He is a co-author of the book *Randomized Algorithms* published by Cambridge University Press in 1995. Motwani has received the Arthur P. Sloan Research Fellowship, the National Young Investigator Award from the National Science Foundation, the Bergmann Memorial Award from the US-Israel Binational Science Foundation, and an IBM Faculty Award.

3 Economics of Information

"A Proposal for Valuing Information and Instrumental Goods"

Dr. Marshall Van Alstyne, University of Michigan.

Author abstract: How should a firm value information capital? This essay offers one framework that combines ideas from Economics and Computer Science. Drawing a distinction between data and procedures, it augments the traditional Bayesian model that treats information as a change in uncertainty with instruments that permit information as instructions to be executed and reused. It then applies the standard hedonic methods - used in marketing to value tangible goods - to information goods. This leads to a generalized method for ascribing value that can be applied both to procedural information such as software, blueprints, and production know-how as well as to arbitrary resources that have instrumental qualities, that is, they represent tools for effecting outcomes. This approach has the added advantage of supporting efficient information transfers since consumers need not always see the information they are about to buy.

Speaker Biography: Marshall Van Alstyne (`www.si.umich.edu/mvanalst`) is an assistant professor at the University of Michigan, where he teaches information economics, electronic commerce, and computer simulation. He holds a bachelor's in computer science from Yale and MS and Ph.D. degrees in information technology from MIT. In 1999, he received an NSF Career Award and Intel Young Investigator Fellowship to explore research on the economics of information. Before returning to academia, he worked as a technology management consultant and he co-founded a software venture. Past clients include Fortune 500 companies as well as US state and federal government agencies. He has published in several journals including *Science* and *Sloan Management Review*, and his research has been the subject of radio broadcasts in the US and Canada.

4 Privacy

"Privacy Implications of Online Data Collection"

Dr. Lorrie Cranor, AT&T Labs - Research

Summary: Dr. Cranor addressed some of the privacy issues that arise out of collecting data over the web and raised the challenges involved in collecting data that can reduce privacy concerns. New Web applications are enhancing businesses' abilities to gather data about their online customers that helps them to provide customized services and targeted advertising. By learning more about their customers, online businesses can develop more personal relationships with them, allowing them to better anticipate and meet their customers' needs. However, many of the Web-based data collection systems being deployed raise many privacy concerns. First, most of these systems are being deployed silently, without notifying Web site visitors or giving them an opportunity to choose whether or not they wish to have their data collected. Second, online data is increasingly being combined with data from a variety of sources, allowing for the development of detailed individual profiles. Third, data gathered for business purposes is increasingly being subpoenaed for use in criminal investigations and civil proceedings. Even when data is stored without traditional identifiers such as name or social security number, profiles often contain enough information to uniquely identify individuals. The privacy concerns raised by online data collection are gaining increased attention from the news media, the public, and policy makers. As more advanced data collection and processing applications are deployed without addressing privacy concerns, individual privacy is slowly eroding. Individually, most of these applications do not pose major threats, but taken together they are bringing us closer to a surveillance society.

Speaker Biography: Lorrie Faith Cranor (`http://www.research.att.com/lorrie`) is a Senior Technical Staff Member in the Secure Systems Research Department at AT&T Labs-Research Shannon Laboratory in Florham Park, New Jersey. She is chair of the Platform for Privacy Preferences Project (P3P) Specification Working Group at the World Wide Web Consortium. Her research has focussed on a variety of areas where technology and policy issues interact, including online privacy, electronic voting, and spam. For more information, please see her home page at http://www.research.att.com/lorrie/http://www.research.att.com/lorrie/.

5 Revolution, not Evolution

"Online Music - The Next Big Revolution"

Dr. Narayanan Shivakumar, Gigabeat, Inc.

Author abstract: Text is passe. The growing popularity of the MP3 and RealAudio formats is leading to a new revolution on the Internet – online music. In this talk, I will discuss some challenging problems that arise in building multi-media web crawlers, and in mining the web for nuggets of text and audio information.

Summary: The growing popularity of the MP3 and RealAudio formats is leading to a new revolution on the Internet – online music. The web not only provides a new forum for emerging artists but also helps consumers to listen to music at an affordable cost and also helps in personalized music discovery. However, in order to be able to provide a good quality personalized music on the web there are various challenging problems that need to be addressed.

One of the challenges is to develop a crawler that can search for patterns in audio files. Since audio files contain unstructured data it is challenging to correlate information in two different audio files based on content or fidelity of songs or some sound patterns that are common to both the files. This problem spans several areas like pattern recognition, data mining, media processing and visualization. In addition, there are other challenges like how to provide a good quality of music when the volume of users increases (i.e., scalability issues). Finally, an exciting challenge is to represent the music discovery process visually, a user interface design question of importance.

Speaker Biography: Narayanan Shivakumar (`http://www-db.stanford.edu/shiva`) is currently the Chief Scientist of Gigabeat.com, an online music startup in Palo Alto, CA. His current research interests include data mining, multi-media, databases, and digital libraries. He has been a summer visitor at Microsoft Corp., Bell Labs, and Xerox PARC. He is a member of ACM and Tau Beta Pi. He received his computer science degrees (MS'97/PhD'99) from Stanford, and from UCLA (BS'94).

"A Petabyte in Your Pocket: Directions for Net Data Management"

Dr. David Maier, Oregon Graduate Institute

Author abstract: In 2015, for a few hundred dollars a year, you can have a personal petabyte database (PetDB) that you can access from any point of connection, with any device. It stores and organizes any kind of digital data you want to have, without losing structure or information. All this data is queryable and it is arranged by type, content, structure, association and multiple categorizations and groupings. You can also locate items by when or how you encountered them, what you have done with them, where you were when you accessed them.

What could you fit in a personal petabyte store? Some possibilities:

- The contents of every book and magazine you ever bought.
- Every web page you've ever visited.
- Every email you've ever sent or received, including all attachments.
- Every version of any piece of software you have ever used.
- Maps and images of every place you have ever been.
- Descriptions and price lists for every kind of product you might want to buy-a "universal, personal catalog."
- Portions of the web you might want to browse, including snapshots at past points of time.

Your PetDB does not appear to reside on any particular computer – you are never on the "wrong" machine to access it. More importantly, you do not have to take any explicit action to insert data into you PetDB; your PetDB does not appear to have an "outside" where data is concerned. Thus your PetDB is also your personal Internet portal: your evolving and customized view of all on-line digital data.

The PetDB is not predicated upon some massive improvement in holographic memory technology or DNA-based storage units. Rather, it is an example of what could be done with a new generation of software infrastructure we term Net Data Managers (NDMs). NDMs are a radical departure from the capabilities and structure of current database management systems. They focus on data movement, rather than data storage, working equally well with live streams of data as with files in secondary storage. They will be capable of storing data of arbitrary types, without a matching database schema having been defined previously. They will efficiently execute queries over thousands or tens of thousands of information sites. They will locate and select data items by both internal content as well as a variety of external contexts. NDMs will also support monitoring rapidly changing information sources in a way that scales to thousands or even millions of triggers.

Summary: In this talk, Dr. Maier laid out the requirements for Net Data Management, and in this context reported on the research directions being pursued by the NIAGARA project, a joint undertaking with David DeWitt and Jeffrey Naughton at the University of Wisconsin.

The objective of NIAGARA is to provide a very inexpensive personal petabyte database (PetDB) that you can access from any point of connection, with any device. It stores and organizes any kind of digital data you want to have, without losing structure or information. All this data is queryable and it is arranged by type, content, structure, association and multiple categorizations and groupings. You can also locate items

by when or how you encountered them, what you have done with them, where you were when you accessed them.

The PetDB does not reside on any particular computer and is accessible from any machine. More importantly, you do not have to take any explicit action to insert data into your PetDB; your PetDB does not appear to have an "outside" where data is concerned. Thus your PetDB is also your personal Internet portal: your evolving and customized view of all on-line digital data.

The PetDB is not predicated upon some massive improvement in holographic memory technology or DNA-based storage units. Rather, it is an example of what could be done with a new generation of software infrastructure we term Net Data Managers (NDMs). NDMs are a radical departure from the capabilities and structure of current database management systems. They focus on data movement, rather than data storage, working equally well with live streams of data as with files in secondary storage. They will be capable of storing data of arbitrary types, without a matching database schema having been defined previously. They will efficiently execute queries over thousands or tens of thousands of information sites. They will locate and select data items by both internal content as well as a variety of external contexts. NDMs will also support monitoring rapidly changing information sources in a way that scales to thousands or even millions of triggers.

Speaker Biography: David Maier (`http://www.cse.ogi.edu/maier`) is a professor of Computer Science and Engineering at Oregon Graduate Institute. His current research interests include object-oriented databases, query processing, superimposed information systems, XML and related standards, information assurance, scientific databases and net data management. He has consulted with most of the major database vendors, including Oracle, Informix, IBM and Microsoft. Maier is an ACM Fellow and a holder of the SIGMOD Innovations Award. He received his PhD from Princeton University in 1978.

6 Data on Data

"Information Access and Data Processing on the Web: Current Limitations, New Techniques, and Future Directions"

Dr. Steven Lawrence, NEC Research Institute

Author abstract: This talk describes current limitations, new techniques, and future directions for information access and data processing on the web. We describe recent studies analyzing the accessibility, distribution, and structure of information on the web, which highlight the fact that there is much room for improvement and new methods. New techniques for information access and data processing on the web are described, including two projects at NEC Research Institute: Inquirus, which is a content-based metasearch engine, and CiteSeer, which is the largest free full-text index of scientific literature in the world.

Summary: Dr. Lawrence described the study of coverage and recency of the major World wide web search engines. This study highlights the following surprising results :

- The coverage of any one single engine is significantly limited. No single engine indexes more than about 1/3rd of the "indexable web". The coverage of the six major search engines investigated varies by an order of magnitude, and combining the results of the six engines yields about 3.5 as many documents on average as compared with the results from only one engine.
- The search engines were irregular in updating information (i.e., the average age of a document ranged from 3-6 months).
- The search engines do not index sites equally, i.e., there is a bias in indexing of data. (The use of key words in web pages influences indexing.)

- Over a period of last three years the percentage of web pages covered by the top 6 search engines have decreased dramatically and the recency of the pages have increased.

These results highlight the fact that there is much room for improvement both in terms of coverage and recency. These observations about coverage and recency raise the question whether there will be increasing need for specialized web search engines in the future rather than a generalized search engines.

Finally, Dr. Lawrence described some new methods and techniques for information access and data processing on the web, including the two projects at NEC Research Institute: Inquirus, which is a content-based metasearch engine, and CiteSeer, which is the largest free full-text index of scientific literature in the world.

Speaker Biography: Steve Lawrence (`http://www.neci.nec.com/lawrence/`) is a Research Scientist at NEC Research Institute in Princeton, NJ. Dr. Lawrence has published over 50 articles in areas including information retrieval, web analysis, digital libraries, and machine learning. Dr. Lawrence has done over 100 interviews with news organizations including the New York Times, Wall Street Journal, Washington Post, Reuters, Associated Press, UPI, CNN, BBC, MSNBC, and NPR. Hundreds of articles about his research have appeared worldwide in over 10 different languages.

"On Collecting and Using Web Data"

Dr. Balachander Krishnamurthy, AT&T Labs - Research

Author abstract: Web-related data has been gathered since the inception of Web, often without the knowledge of a vast majority of Web users. The web data includes the actual raw data that the client requested, client, proxy, and server logs, and HTTP packet traces. Apart from obvious privacy issues, Web data has problems relating to gathering, storing, cleaning, and validation.

I have been involved in several aspects of collection of Web-related data from a wide variety of sources (both inside and outside AT&T) and in creating a repository in conjunction with the W3C World Wide Web consortium's Web Characterization group. The web data is used by various applications ranging from web caching, improving HTTP/1.1 protocol, testing Web software components for compliancy with the protocol, reducing validation traffic, to predicting future access.

Speaker Biography: Balachander Krishnamurthy (`http://www.research.att.com/bala/papers`) has been with AT&T Labs-Research since receiving his PhD in Computer Science from Purdue University. He has written and edited a book called *Practical Reusable UNIX Software* (John Wiley, 1995) and was the series editor of *Trends in Software* (John Wiley) consisting of 8 books published over a period of five years. He has several patents, published over thirty five technical papers, given invited lectures in over twenty countries, and presented tutorials on aspects related to the Web. He is the area editor for Web related issues for ACM SIGCOMM's 'Computer Communications Review' and is currently working on a book that will provide a technical overview of the World Wide Web.

7 XML

"XML + Databases = ?"

Dr. Michael Carey, IBM Almaden Research Center

Author abstract: In the first half of this talk, I will share some of my thoughts on semistructured databases, object-relational databases, XML, web querying, and how they are all related (or not). Using one of my favorite queries ("find U.S.-made Fender Jazz Bass or Precision Bass guitars available for under $700 within 50 miles of my home in San Jose, California"), I will talk about what one can and cannot do on the web today and how the database community can hopefully help change that. I will also talk about the pros and cons of the aforementioned technologies, in terms of making my query answerable, and I will propose a possible XML-based research agenda that might help us get there from here.

In the second half of this talk, I will discuss a new project - called Xperanto (Xml Publishing of Entities, Relationships, ANd Typed Objects) - that we have initiated at the IBM Almaden Research Center. The goal of this project is to provide facilities to enable "XML people" (as opposed to "SQL people") to conveniently publish content from relational and object-relational databases on the web in queryable XML form. I will outline the approach that we are taking, including the architecture of the system, the roles of the various Xperanto components, and some of the technical issues and challenges involved in the project.

Speaker Biography: Michael J. Carey (`http://www.almaden.ibm.com/cs/people/carey`) received the Ph.D. degree from UC Berkeley in 1983. He spent 13 years on the faculty at the University of Wisconsin-Madison,where he conducted research on DBMS performance, transaction processing, distributed and parallel database systems, extensibility, and object-oriented (O-O) databases. He co-directed the EXODUS and SHORE projects while at Wisconsin. In mid-1995, Carey joined the staff of the IBM Almaden Research Center, where he has worked on the Garlic heterogeneous information system project and more recently on object-relational (O-R) database system technology for DB2. Inspired by a semester spent as Stonebraker Visiting Fellow at UC Berkeley in 1999, he has also begun to explore the intersection of XML and object-relational database system technology. His current interests include O-R DBMS implementation techniques, the use of XML to publish databases' contents on the web, and the ongoing evolution of the SQL standard.

"The Next 700 Markup Languages"

Dr. Philip Wadler, Lucent Technologies - Bell Labs

Author abstract: XML (eXtensible Markup Language) is a magnet for hype: the successor to HTML for Web publishing, electronic data interchange, and e-commerce. In fact, XML is little more than a notation for trees and for tree grammars, a verbose variant of Lisp S-expressions coupled with a poor man's BNF (Backus-Naur form). Yet this simple basis has spawned scores of specialized sublanguages: for airlines, banks, and cell phones; for astronomy, biology, and chemistry; for the DOD and the IRS. Domain-specific languages indeed! There is much for the language designer to contribute here. In particular, as all this is based on a sort of S-expression, is there a role for a sort of Lisp?

Summary: Dr. Wadler analyzed the features of XML and contrasted it with languages like LISP and also highlighted the serious limitations and drawbacks in XML and the need to come up with framework-like query algebra for XML and the need for designing a language for the web that can support polymorphism and strong typing.

Speaker Biography: Philip Wadler (`http://www.cs.bell-labs.com/ wadler`) is a researcher at Bell Labs, Lucent Technologies, and codesigner of the languages Haskell and GJ. He spends his time on the border between theory and practice, seeking ways one may inform the other. He helped turn monads from a concept in algebraic topology into a way to structure programs in Haskell, and his work on GJ may help turn quantifiers in second-order logic into a feature of the Java programming language. He edits the Journal of Functional Programming for Cambridge University Press, and writes a column for SIGPLAN Notices. He was an ACM distinguished lecturer 1989–1993, and has been an invited speaker in Amsterdam, Austin, Boulder, Brest, Gdansk, London, Montreal, New Haven, Portland, Santa Fe, Sydney, and Victoria.

8 Ubiquity

"ObjectGlobe: Ubiquitous Query Processing on the Internet"

Dr. Alfons Kemper, Universitat Passau

Author Abstract: We present the design of ObjectGlobe, a distributed and open query processor. Today, data is published on the Internet via Web servers which have, if at all, very localized query processing capabilities. The goal of the ObjectGlobe project is to establish an open market place in which data and query processing capabilities can be distributed and used by any kind of Internet application. The goal of the ObjectGlobe project is twofold. First, we would like to create an infrastructure that makes it as easy to distribute query processing capabilities (i.e., query operators) as it is to publish data and documents on the Web today. Second, we would like to enable clients to execute complex queries which involve the execution of operators from multiple providers at different sites and the retrieval of data and documents from multiple data sources. All query operators should be able to interact in a distributed query plan and it should be possible to move query operators to arbitrary sites, including sites which are near the data. The only requirement we make is that all query operators must be written in Java and conform to the secure interfaces of ObjectGlobe. One of the main challenges in the design of such an open system is to ensure security. We discuss the ObjectGlobe security requirements, show how basic components such as the optimizer and runtime system need to be extended. Finally, we present the results of some performance experiments that assess the benefits of placing query operators close to the Internet data sources and the additional cost for ensuring security in such an open system.

Speaker Biography: Alfons Kemper (`http://www.db.fmi.uni-passau.de/`) received his Bachelor degree in Computer Science from the University of Dortmund (Germany) in 1979. He then moved to the University of Southern California where he obtained the Masters degree and the Ph. D. degree in Computer Science in 1981 and 1984, respectively. From 1984 until 1991 he was an Assistant Professor of Computer Science at the University of Karlsruhe, Germany. He spent two years (from 1991 until 1993) as an Associate Professor at the Technical University (RWTH) of Aachen, Germany. He is currently a Full Professor of Computer Science at the University of Passau, Germany. His research interests center around the design and realization of advanced database technology. His main research focus was on indexing and query processing techniques for object-oriented and object-relational database systems and performance issues related to complex database application systems (such as decision support systems and SAP R/3). In his recent work he concentrates on distributed database implementation techniques and distributed query processing over Internet data sources.

"Data Management for Ubiquitous Computing"

Dr. Alon Levy, University of Washington

Summary: In the not too distant future, many devices (e.g., common household appliances, PDAs, cellphones, cars) will contain computer chips that will enable them to exhibit more sophisticated behavior and interact with other devices. For example, refrigerators will be able to monitor their contents and automatically order supplies. The heating system of a house will monitor the alarm clocks and calendars of their owners to set the temperature in an optimal fashion. In the context of ubiquitous computing, data exchange and computation occur in the background in response to cues from users. Computing is centered around the network and data, rather than around the computing devices or disks as it is today. Devices are added and removed from the network on a regular basis, and they must be able to interoperate with little human intervention.

In the first part of the talk, Dr. Levy raised the following question: What is the appropriate programming abstraction for specifying for ubiquitous computing? In the second half of the talk he addressed this question in the context of the Sagres Project being conducted at the University of Washington, which addresses the data management issues that arise in the context of ubiquitous computing. In particular, the issues of modeling a wide variety of devices, describing the complex interactions between devices and the easy addition and removal of devices.

Speaker Biography: Alon Levy(`http://www.cs.washington.edu/homes/alon`) joined the faculty of the Computer Science and Engineering Department of the University of Washington in January, 1998. Before joining the U. of Washington, he was a principal member of technical staff at AT&T (previously, Bell) Laboratories. He received his Ph.D in Computer Science from Stanford University in 1993. Alon's interests are in data integration, web-site management, semi-structured data, database aspects of ubiquitous computing, query optimization, and interactions between Databases and Artificial Intelligence. In June, 1999 he co-founded Nimble.com, a company that builds tools for query processing and data integration for XML.

9 Open Problems

9.1 Research Areas

A variety of research issues were raised during the open-problems sesssion.

Centralized computation in a massively distributed environment. The data in the web is distributed across various locations/sites. So, it seems natural that the web search engines are good target applications for distributed computing. However, the existing web search engines tend to centralize the computation rather than distribute the computation among the various sources of data. This raises the following issues that are potential topics for further research.

- Can we provide a framework that can exploit the distributed nature of the web by distributing the computation performed by the search engines among the various data sources? If so, what will be the its impact on the network bandwidth?
- What are the costs/benefits of distributing the computation versus centralized computation?
- Is the cost model for consumer applications different from that of the business applications?

Some of the above issues were also discussed during the talk by Dr. Kemper.

Techniques for searching special types of unstructured data. The Web, in addition to textual data, also contains other forms of data like Protein Data Bank, DNA Data Bank, audio files and images. The textual data is highly structured whereas the other forms of data are unstructured. This raises the following research problem:

- How do we search for patterns within specialized data like images, protein data, DNA sequences, etc.?

Quality of service. There is an increasing need to provide good quality of service for the various web applications. This raises the following questions:

- What are the measures/metrics that can be used to characterize Quality of Service?
- How do we come up with tools and techniques to measure QoS?
- How do we design applications that can provide good QoS?

Processing data in wide area networks. In the context of wide area networks the following issues were raised:

- How do you do wide area processing?
- Where do you get the data?
- Where do you obtain web statistics?
- How do you address performance issues, QoS, etc.?

Defining and adapting to standards.

- How do you come up with procedures and tools so that it makes it easy for users to adapt to new standards?
- What is a good model for coming up with consensus on what should be a standard and what should not be a standard?

Economic feasibility of distributing the Web. The services in the web are increasingly getting centralized and due to competing business interests there seem to be a lot of barriers in sharing data across businesses. So this raises the following issues:

- Do the economic models point to a centralized model for the web?
- Are the cost models for the consumer applications different from the business to business applications?

Role of emerging network infrastructure and devices on data processing on the Web. The network infrastructure is diversifying with high speed optical, cable and wireless networks, and the network bandwidth and reach is dramatically increasing. Also, devices such as PDAs are becoming commonplace and the trend is to look upon them as extensions of the network itself. How does this affect the model of data processing, querying, and visualizing answers?

9.2 Lessons Learned and Issues for Future Research

Many of the speakers addressed related topics from different perspectives.

- Both experimental measurements and analytical models of the Web are necessary to understand the Web's actual graph structure. This knowledge can have a significant impact on how pages are indexed and made available to users at large. These issues were addressed by Drs. Rajagopolan and Lawrence.
- Despite the growth of XML, HTML is not going away and therefore the demand for "HTML wrappers", which can extract data from HTML pages, will continue to exist. Given this demand, it is important to study the problems of how to automate generation of wrappers; how to extract content "in the small", i.e., using statistical models to determine the relevance of neighboring terms on a page; and how to make meta data more useful since it is currently used rarely or inaccurately. These issues were raised by Drs. Motwani, Lawrence, and Cohen.
- XML is emerging as the defacto standard for data exchange in business-to-business (B2B) electronic commerce, e.g., in virtual marketplaces. In this application domain, XML is expected to be used as a message format, i.e., to represent small, asynchronous segments of data. Existing query languages and systems for XML are oriented towards transactional, legacy database systems. An open question is what are the basic functions of a query language for B2B messaging systems. These issues were raised by Drs. Maier, Kemper and Levy.
- Already a great deal of information about an individual Web user can be extracted from logged trace data, which can possibly compromise the user's privacy. As more structure is added to Web data, e.g., in the form of tagged XML, a user's privacy will be potentially even harder to protect. Concurrent with technological advances, the sociological and legal dimensions of privacy must be considered. Drs. Krishnamurthy and Cranor addressed some of these issues.

Acknowledgement: DIMACS thanks the National Security Agency, the New Jersey Commission on Science and Technology, and the Office of Naval Research for support of this workshop.