Data Exploration and Knowledge Discovery in a Patient Wellness Tracking (PWT) System at a Nurse-Managed Health Services Center

Yuan An
The iSchool at Drexel
University, USA
yan@ischool.drexel.edu

Ritu Khare
The iSchool at Drexel
University, USA
ritu@ischool.drexel.edu

Il-Yeol Song
The iSchool at Drexel
University, USA
isong@ischool.drexel.edu

Xiaohua Hu
The iSchool at Drexel
University, USA
thu@ischool.drexel.edu

ABSTRACT

This paper describes our ongoing research on data exploration and knowledge discovery in a patient wellness tracking (PWT) information system developed for a nurse-managed community health center. The center employs an innovative and transdisciplinary care model that fully integrates behavioral and various wellness services into primary care to form a team approach. We have developed the PWT system that integrates clinical data collected in an electronic medical record (EMR) system with the data generated by a spectrum of healthy living programs and wellness services. While data is being collected rapidly in large volumes, it is imperative to develop effective tools in helping clinicians explore data and discover knowledge. In this paper, we present (1) an exploratory data browser based on information content in information theory for searching granularity patient data, and (2) a knowledge discovery component based on probabilistic graphical models for diagnosis, prognosis, and revealing clinical cause-effect interactions.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Management—*Database Applications*; J.3 [**Computer Applications**]: Life and Medical Systems

General Terms

Algorithms, Experimentation, Performance

Keywords

Health Knowledge Discovery, Probabilistic Reasoning for Health Decision Making, Health Data Management

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
IHI'12, January 28–30, 2012, Miami, Florida, USA.
Copyright 2012 ACM 978-1-4503-0781-9/12/01 ...$10.00.

1. INTRODUCTION

"*We are drowning in data but starving for knowledge.*" This problem has been exacerbated in the health care domain ever since electronic medical/health record (EMR/EHR) systems were deployed to document individual encounters, store test results, and exchange treatment plans. It is imperative to develop effective tools that can assist clinicians to make decisions in various health care processes. This paper describes our ongoing effort on developing the data exploration and knowledge discovery functionality in a patient wellness tracking (PWT) information system. The system has been developed in a nurse-managed community health center which serves primarily low-income and underserved population. The goal of this paper is to describe the desired functionality that would help clinicians at the center monitor and manage health issues among the local residents.

It is well recognized that health information technologies have a great potential in helping healthcare professionals reduce costs and improve outcomes [7, 5, 6]. The nurse-managed health center (hereafter, the center) employs an innovative and transdisciplinary care model that fully integrates behavioral and various wellness services into primary care to form a team approach. Although the center implemented an electronic medical record (EMR) system, many aspects of the transdisciplinary care model were not covered by the EMR system. To address the issue, we have developed the PWT system, a comprehensive health information system that extends the EMR system and integrates clinical data with various data generated by a spectrum of healthy living programs and wellness services. While data is being collected rapidly in large volumes, it is important and necessary to develop effective tools in helping clinicians explore data and discover knowledge. In this paper, we present our efforts to data exploration and knowledge discovery by applying the state-of-the-art information theory techniques and probabilistic graphical models.

2. DATA COLLECTION IN THE PATIENT WELLNESS TRACKING (PWT) SYSTEM

The PWT system aims to improve data gathering, analysis, and sharing related to a wide variety of healthy living programs and wellness services provided by the center

[2]. These programs and services include physical exams, diagnosis and treatment of illness, family planning, health maintenance/disease prevention services, behavioral health services, physical fitness programs, dental services, nutrition services, and chronic disease management programs. The system is used by multiple health professionals who together make up a healthcare team. Team members include nurse practitioners, behavioral health specialists, social workers, wellness coordinators, nutritionists, physical therapists, and dentists.

On a daily basis, team members are faced with an array of important clinical decisions (e.g., what type of guidance to provide to their patients; what clinical assessments and interventions are necessary). Having a complete view of patient health status including medical history, current treatments, daily activities, and social interactions is critical to making decisions. The current PWT system has implemented the following additional functionalities: (1) patient self-reporting data related to behavioral health, (2) creating healthy living programs and recording patient attendance, and (3) linking and integrating the clinical data extracted from the electronic medical record (EMR) system.

3. DATA EXPLORATION

While data is accumulated into several relational databases with more than 120 relational tables, browsing and querying the databases is constantly requested by the clinicians. Although form-based query interfaces have been built, the ability of navigating the databases is limited by the fixed number of forms. In this section, we describe a multifaceted data exploration system that supports users' serendipitous and diverse search in the databases.

A typical multifaceted system uses hierarchical categories for browsing and navigation [13, 8]. A hierarchical category describes metadata of an underlying database. Users can expand a category and follow the parent-child links to navigate the hierarchical category and browse the underlying database. Traditional library catalogue systems and existing Web directories have demonstrated the efficacy of hierarchical categories. In these systems, no explicit query languages are required for searching the information spaces. However, building such a system requires substantial human effort in constructing a hierarchical category structure and associating information items with categories in advance. Navigation and search is constrained by the pre-defined paths.

We consider a dynamic multifaceted data exploration system. The system uses a set of category hierarchies each of which corresponds to a different concept or entity relevant to the collection of data. A category hierarchy allows the navigation from a concept to attributes of the concept and to values of an attribute. Concepts, attributes, and values constitute the multi-facets which should be presented dynamically and "optimally" at each turning point of the user search process.

Figure 1 illustrates an example showing the entities managed in an underlying database and the corresponding faceted categories for navigation. The faceted categories on the left-hand side will be displayed on the graphical user interface. Underlined terms are facets that are clickable. When a facet is clicked, the system will discover the semantic relationships between the currently clicked facet and the previously selected ones. In the meantime, facets are reorganized and displayed based on the "interestingness" with regard to the user's query intention.

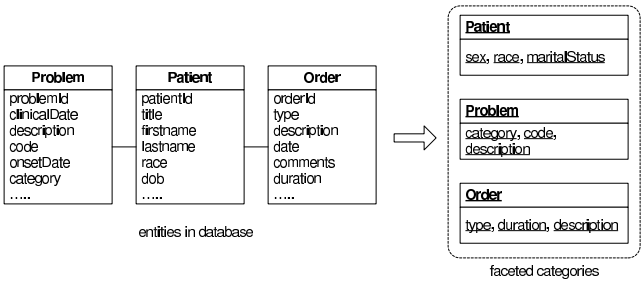

Figure 1: From Database Entities to Faceted Categories

The architecture of the data exploration system is shown in Figure 2. The system uses a conceptual model as the source of faceted categories. The conceptual model is extracted from the underlying relational schemas by existing reverse-engineering or mapping tools [3, 1]. The conceptual model provides user-friendly category and subcategory names and serves as a global schema for data integration.

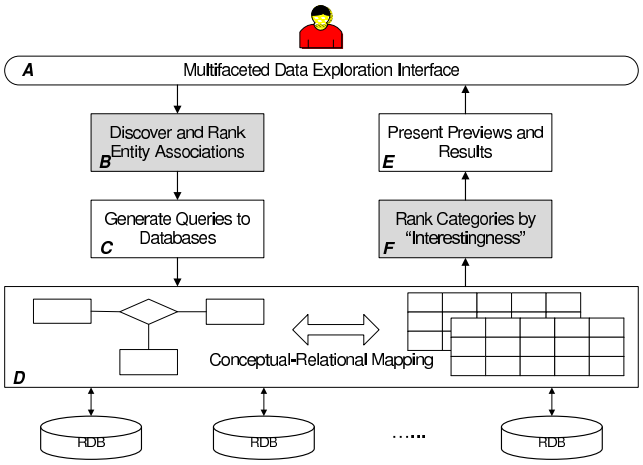

Figure 2: The Architecture of the Data Exploration System

There are two key challenges in developing the dynamic multifaceted data exploration system. The first is *how to rank the facets at each turning point of the user search process and present the most "interesting" categories?* (module **F** in Figure 2). The second challenge is *how to infer the most likely relationship between two selected categories/concepts for querying the databases and generating results?* (module **B** in Figure 2).

Ranking Facets. The PWT interface displays concepts, attributes, and values as facets. A typical conceptual model may contain dozens of concepts and each concept may have dozens of attributes. A challenge of presenting the facets is to select a reasonable number of most "interesting" facets and judiciously organize them to help users browse and search. As opposed to the previous work [4], our goal here is to support the clinicians in exploring data residing in different ta-

bles (possibly from different sources), instead of distinguishing tuples in a single table. To achieve the goal, we rank the facets based on some sense of "interestingness" and adopt the well-known 7-item principle about the capacity limits of human cognition [14], i.e., only present the top-7 items in each category and allow the user to expand on demand.

We develop three different techniques for ranking the three different types of facets. We rank values simply by their frequencies under the attribute in the database. We rank attributes based on the entropies of the corresponding columns in the databases. Entropy is used to measure the information content of an attribute [16]. The entropy of an attribute is computed as follows. For an attribute (corresponding to a column) c in a table T with size $|T|$, let $A = \{a_1, a_2, ..., a_m\}$ be all the values under the column $T.c$. Let n_{a_i} be the number of tuples which have the value a_i on column c, then the entropy of the column c is

$$H(c) = \sum_{i=1}^{m} p_i log(1/p_i)$$

where $p_i = n_{a_i}/|T|$ is the fraction of the tuples in T which have the value a_i on c. While entropy provides a numerical measure of the information content for attributes, an intuitive approach of using it, such as, ranking attributes with larger entropy higher, may not be the best strategy. We are experimenting different thresholds and strategies in our methods.

For ranking a concept, we exploit the information in several sources: the connectivity of the conceptual model, the data instances in the underlying databases, and the user's query preference. We use $Pr(C_j|C_i)$ to denote the probability of selecting concept C_j given that the concept C_i is selected. We assume that the concept selection process is a first-order Markov chain, i.e., the currently selected concept depends only on the previously selected concept (excluding any earlier concepts selected). Given a set of concepts $\mathcal{C} = \{C_i | i = 1..n\}$, at each step when a concept C_i is selected, we compute $Pr(C_j|C_i)$ for $\forall j \in \{1..n\}$ and $j \neq i$. We then present the top-k concept with highest k $Pr(C_j|C_i)$ values on the multi-faceted interface. The process is repeated until the user stops the exploration. The probability $Pr(C_j|C_i)$ is computed using the following formula:

$$Pr(C_j|C_i) = w_1 * Imp_G(C_j|C_i) + w_2 * Imp_D(C_j|C_i) +$$

$$w_3 * Pr_U(C_j|C_i)$$

where $Imp_G(C_j|C_i)$ is the importance of concept C_j with respect to the concept C_i in a graph G, $Imp_D(C_j|C_i)$ is the importance of concept C_j with respect to the concept C_i in the underlying database, and $Pr_U(C_j|C_i)$ is the probability of selecting the concept C_j if the user has selected the concept C_i. The coefficients w_1, w_2, and w_3 are weights for combining the three different quantities. The final probability is normalized into a value in the range [0...1].

Each item in the formula is computed as follows. We treat the conceptual model as a graph $G = (V, E)$. The connectivity of a graph refers to how the nodes are connected in the graph. A static random walk algorithm [16] computes the importance of nodes for a given graph. We develop a novel algorithm to compute the *relative importance* of nodes with respect to a given node. In particular, the algorithm takes the distance from all the other nodes to the given node into consideration. The farther the other node is from the given node, the less likely the importance of the given node is propagated to the other node. The relative importance of a node is computed with respect to a given node in the graph corresponding to the conceptual model. The same algorithm is applied to the underlying data instances as well. Specifically, data instances of an underlying database are connected based on the foreign key referential relationships between different relational tables. In addition to the connectivity, the importance of a concept also takes the information content of a table into consideration. The notion of information content is based on the entropy of attributes. For computing the probability $Pr_U(C_j|C_i)$ of selecting concept C_j given the concept C_i, we analyze logs containing query histories. The more frequent a user queries the PWT system, the more likely the query logs provide patterns of the user's preference.

Inferrring Semantic Relationships between Concepts. A conceptual model describes concepts and relationships in a domain. While the multifaceted interface displays concepts and their attributes as clickable facets, the relationships between concepts are hidden from users. Let $\mathsf{C}_i.\mathsf{a}_j$ denote the attribute a_j of the concept C_i. Let $G' = (V', E')$ be the subgraph of the conceptual graph the user has already navigated. If the next facet being selected is $\mathsf{C}_i.\mathsf{a}_j$, we need to infer the semantic relationship between the concept C_i and a node in the graph $G' = (V', E')$.

A similar problem has been studied in universal relation [12] where a domain is described by a set of attributes and the actual data is stored in decomposed relations. When a query only refers to the attributes, the system has to infer the most reasonable join paths between the decomposed relations. The key idea behind the universal relation is the notion of lossless join and maximal objects [12]. We need to infer the most reasonable relationship between a selected concept and a given subgraph. We develop a technique taking the following information into consideration: length of a relationship (path), functional (many-to-1) or non-functional relationships (a functional relationship gives rise to lossless join), and the evidence in the data instances.

Let C be the next concept selected and $V' = \{C_1, C_2, ..., C_n\}$ be the set of concepts in the subgraph $G' = (V', E')$. We infer and rank the connections between C and a concept $C_i \in V'$ in the original conceptual model graph as follows. We first select the shortest paths between C and $C_i \in V'$. We rank all the shortest paths among all the nodes $C_i \in V', i = 1...n$ by their distances; the shorter, the better. Next, we rank the functional relationship higher than non-functional relationships. A functional relationship could be from C to $C_i \in V'$ or vice versa. Finally, we rank the relationships by the importance calculated using the data instances as follows.

Let $p =< C_1, C_2, ..., C_m >$ be a path between two concept nodes, C_1 and C_m, in a conceptual graph, where a node C_i corresponds to a concept. We rank the importance of the path based on the instantiations of the path in the data instances. We can get the number of instantiations of a path and the number of instances of a concept from the underlying databases. Let $Inst(C)$ be the number of instances of a concept C. For an instance o_i of C, let $N_p(o_i)$ be the number of instantiated paths p the instance o_i participates in. We call an instance o_i *a linked object* if $N_p(o_i) > 0$. Let $M_p(C)$ be the number of linked objects of C participating in the path p. The quantity $Frac(C) = \frac{M_p(C)}{Inst(C)}$ is the fraction

of the instances of C that participate in some instantiated paths p. The quantity

$$AvgP(C) = \frac{\sum_{i=1}^{Inst(C)} N_p(o_i)}{M_p(C)}$$

is the average number of instantiated path p a linked object of C participates in. The importance of the path $p =< C_1, C_2, ..., C_m >$ is computed as:

$$Imp(P) = \sum_{j=1}^{m} w_j \times AvgP(C_j) \times Frac(C_j)$$

where w_j is the weight measuring the contribution of a concept to the importance.

4. KNOWLEDGE DISCOVERY

In this section, we describe data analysis and knowledge discovery in the PWT system in helping the clinician make decisions. As the PWT collects and integrates various kinds of medical, social, behavioral, and spiritual data, we aim to analyze the data to help the clinicians answer the following three kinds of questions: (1) given the screening results and vital signs reading, what would be the possible health status changes since the patient's last visit? (**diagnosis**), (2) if the patient is referred to the specific program and service X, what would be the possible outcomes? (**prognosis**), and (3) what would be the possible cause of the particular change in the patient's health status since the patient's last visit? (**cause-effect relationship**).

Bayesian networks with their associated methods are especially suited for capturing and reasoning with uncertainty [10]. They have been around in biomedicine and health care for many years and have become increasingly popular for handling diagnosis, prognosis, and discovering cause-effect relationships [11]. We investigate the problems of learning and using Bayesian networks in the PWT system for knowledge discovery and decision making.

A Bayesian network captures the joint distributions and conditional independence over a set of random variables. For a set of random variables $\mathbf{X} = \{X_1, X_2, ..., X_n\}$, a Bayesian network consists of (1) a network structure S that encodes a set of conditional independence assertions about the variables in $\mathbf{X}$, and (2) a conditional probability table (CPT) associated with each variables. The network structure S is a directed acyclic graph. The nodes in S are in one-to-one correspondence with the variables $\mathbf{X}$. For monitoring and managing health issues, the clinicians at the center may have interests in many variables. For example, the following is a list of some interesting variables:
Smoking, *Age*, *Sex*, *Weight*, *Attending_Yoga_Class*, *Attending_Cooking_Class*, *Using_Fitness_Center*, *Blood_Pressure*, *Zipcode*, *SF-36_Physical_Score*, *SF-36_Mental_Score*, *Type-II_Diabetes*, *Behavioral_Problem*,..., etc,
where the *SF-36_Physical_Score* and *SF-36_Mental_Score* are the screening scores of the 36 health survey questions [15]. Figure 3 shows an example Bayesian network describing the states of some above variables. Let PS denote the variable *SF36_Physical_Score*, let CC denote the variable *Attending_Cooking_Class*, and let FC denote the variable *Using_Fitness_Center*. Table 1 shows the conditional probability table (CPT) for the variable *SF36_Physical_Score*. The structure of the Bayesian network indicates that the variable *SF36_Physical_Score* is conditionally independent of the rest of the variables given its parents, *Attending_Cooking_Class* and *Using_Fitness_Center*.

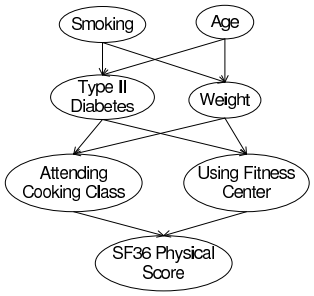

Figure 3: A Bayesian Network Example

$Pr(PS = \text{“}down\text{”} \mid CC = \text{“}yes\text{”}, FC = \text{“}yes\text{”})$
$Pr(PS = \text{“}down\text{”} \mid CC = \text{“}yes\text{”}, FC = \text{“}no\text{”})$
$Pr(PS = \text{“}down\text{”} \mid CC = \text{“}no\text{”}, FC = \text{“}yes\text{”})$
$Pr(PS = \text{“}down\text{”} \mid CC = \text{“}no\text{”}, FC = \text{“}no\text{”})$
$Pr(PS = \text{“}up\text{”} \mid CC = \text{“}yes\text{”}, FC = \text{“}yes\text{”})$
$Pr(PS = \text{“}up\text{”} \mid CC = \text{“}yes\text{”}, FC = \text{“}no\text{”})$
$Pr(PS = \text{“}up\text{”} \mid CC = \text{“}no\text{”}, FC = \text{“}yes\text{”})$
$Pr(PS = \text{“}up\text{”} \mid CC = \text{“}no\text{”}, FC = \text{“}no\text{”})$

Table 1: The Conditional Probability Table (CPT) of the Variable *SF36_Physical_Score*

In general, a Bayesian network provides a complete description of a domain captured by the set of variable $\mathbf{X} = \{X_1, X_2, ..., X_n\}$. Every entry in the full joint probability distribution $P(x_1, x_2, .., x_n)$ can be calculated from the information in the network. In particular, the value of a joint probability is given by the formula

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i)),$$

where $parents(X_i)$ denotes the specific values of the variables in the parents of the variable X_i. Thus, each entry of the joint probability distribution of the domain is calculated by the product of the appropriate elements of the conditional probability tables (CPTs) in the Bayesian network.

Once we have constructed a Bayesian network (from expert knowledge, data, or combination), the network allows us to make probabilistic inferences, i.e., to determine various probabilities of interest from the model. For example, we want to know the probability of *Using_Fitness_Center* given the observations of other variables. The real model for the health center includes many variables. In particular, the current PWT system provides 16 screening surveys related to behavioral health with total 370 questions, and records the attendance information about 54 healthy living programs and wellness services. There are about 3400 patients participating in various programs. Each of the 370 survey questions could be considered as a variable modeled by the Bayesian network. In addition, the EMR system records the clinical data related to patient visit, individual encounter, test results, order, medication, and treatment

plan. We have plenty of data for developing Bayesian networks for knowledge discovery and decision making.

Because building Bayesian networks is a creative process requiring substantial involvement of domain experts, we work closely with the clinicians at the health center to develop the networks in an iterative fashion. We first ask them to list the survey questions, diseases, treatments, programs, etc., of their interests. We then use the data to construct a Bayesian network for these variables. After the clinicians examine the network and make modifications, we use the data to enhance and verify the model again. This process is iterated many times. In each iteration, several knowledge discovery tasks are fulfilled: evaluating the performance of various healthy living programs and wellness services to patient health status, discovering the possible causes to a specific observation, predicting the outcomes of a particular service, and providing evidence and guidelines to patients for managing chronic diseases. There are several potential computational challenges in the application. First, we will investigate effective learning algorithms for discovering the structures and parameters of Bayesian networks from the combination of expert knowledge and data. Second, we need to consider to search for models with hidden variables when the data is incomplete. The resultant models should be able to help the clinicians discover hidden causes. Third, with uncertainty of domain knowledge (even the clinicians are not sure about the interactions between various treatments and outcomes) and the large scale of networks, answering queries through efficient inference algorithms is challenging. We may need to apply approximate methods to the inference problem.

5. CONCLUDING REMARKS

This is a short paper describing the ongoing research on developing the decision-support functionality for the patient wellness tracking (PWT) system in a nurse-managed health services center. The entire study is a collaborative work between the College of Information Science and Technology and the College of Nursing and Health Professions at Drexel University. The earlier work has been focused on understanding the information needs of various healthcare professionals at the center [2], and assisting the users to collect dynamic data through a flexible interface [9]. The recent focus has been shifted to using the accumulated data for knowledge discovery and decision making. This paper describes the effort of applying sophisticated database, information theory, and machine learning techniques in the tasks. More mature ideas, substantial amount of experimentation, system evaluation, and data analysis will be presented in future publications.

6. ACKNOWLEDGMENTS

This research is partially supported by NSF CCF 0905291, NSF CCF 1049864, and NSFC 90920005.

7. REFERENCES

[1] Y. An, A. Borgida, and J. Mylopoulos. Inferring Complex Semantic Mappings between Relational Tables and Ontologies from Simple Correspondences. In *Proceedings of International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE)*, pages 1152–1169, 2005.

[2] Y. An, P. W. Dalrymple, M. Rogers, P. Gerrity, J. Horkoff, and E. Yu. Collaborative social modeling for designing a patient wellness tracking system in a nurse-managed health care center. In *DESRIST '09: Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, pages 1–14, Philadelphia, PA, USA, 2009. ACM.

[3] M. Andersson. Extracting an entity relationship schema from a relational database through reverse engineering. In *Proceedings of International Conference on Conceptual Modeling (ER)*, 1994.

[4] S. Basu Roy, H. Wang, G. Das, U. Nambiar, and M. Mohania. Minimum-effort driven dynamic faceted search in structured databases. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 13–22, New York, NY, USA, 2008. ACM.

[5] J. d. Ralston, D. Revere, L. S. Robins, and H. I. Goldberg. Patients' experience with a diabetes support programme based on an interactive electronic medical record: Qualitative study. *BMJ*, 328:1159, 2004.

[6] G. Demiris, L. B. Afrin, S. Speedie, et al. Patient-centered applications: Use of inforamtion technology to promote disease management and wellness: A white paper by the amia knowledge in motion working goup. *J Am Med Inform Assoc*, 15:8–13, 2007.

[7] D. Dorr, L. M. Bonner, A. N. Cohen, R. S. Shoai, R. Perrin, E. Chaney, and A. S. Young. Informatics systems to promote improved care for chronic illness: A literature review. *J Am Med Inform Assoc*, 14:156–163, 2007.

[8] M. Hearst. Next generation web search: Setting our sites. *IEEE DATA ENGINEERING BULLETIN*, 23, 2000.

[9] R. Khare, Y. An, X. Hu, and I.-Y. Song. Can clinician create high-quality databases? a study on a flexible electronic health record (fehr) system. In *The Proceedings of the 1st ACM Health Informatics Symposium (IHI'10)*, Washington, DC, USA, 2010.

[10] P. Lucas. Bayesian analysis, pattern analysis and data mining in health care. *Current Opinion in Critical Care*, 10(399-403), 2004.

[11] P. Lucas, L. C. G. van der, and A. Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 30:201–214, 2004.

[12] D. Maier and J. Ullman. Maximal Objects and the Semantics of Universal Relation Databases. *ACM TODS*, 8(1):1–14, March 1983.

[13] G. Marchionini. Exploratory search: From finding to understanding. *Communication of the ACM*, 49(4), 2006.

[14] G. Miller. The magical number seven, plus or minus two. *The Psychological Review*, 63(2):81–97, 1956.

[15] J. Ware, M. Kosinski, and B. Gandek. *SF-36 Health Survey: manual and interpretation guide*. QualityMetric Incorporated, 2005.

[16] X. Yang, C. Procopiuc, and D. Srivastava. Summarizing relational databases. In *Proceedings of Very Large Databases (VLDB'09)*, 2009.