

Interactively Building Agents for Consumer-side Data Mining

Rattapoom Tuchinda and Craig A. Knoblock

Information Sciences Institute
University of Southern California
Marina Del Rey, CA 90292 USA
+1 310 448 8786

{pipet, knoblock}@isi.edu

ABSTRACT

Integrating and mining data from different web sources can make end-users well-informed when they make decisions. One of many limitations that bars end-users from taking advantages of such process is the complexity in each of the steps required to gather, integrate, monitor, and mine data from different websites. We present the idea of combining the data integration, monitoring, and mining as one single process in the form of an intelligent assistant that guides end-users to specify their mining tasks by just answering questions. This easy-to-use approach, which trades off complexity in terms of available operations with the ease of use, has the ability to provide interesting insight into the data that would require days of human effort to gather, combine, and mine manually from the web.

Categories and Subject Descriptors

H.5.2 [User Interfaces]: *Graphical user interfaces (GUI), Interaction styles*

General Terms

Algorithms, Design, Human Factors

Keywords

Information Integration, Information Agent, User Interface

1. INTRODUCTION

Any kind of data imaginable can be found on the Internet today: pricing and reviews of goods and service from multiple vendors, maps, and statistic about events around us - just to name a few. This data, when combined and mined the right way, can give consumers interesting insights. Hamlet [1] is one such example where airline ticket prices can be predicted from mining ticket price data listed on major travel websites.

However, the mining process often requires extracting and integrating data from multiple web sources, monitoring to collect the data over time, and then mining the result. Such a task is often too time consuming for end-users (i.e., consumers) to perform manually. A wide range of data mining software exists

but it either requires a steep learning curve or addresses only a specific business problem.

We present the idea of combining the data integration, monitoring and mining as one single process that can be done by end-users. Our implementation, the Consumer-side Data Mining Assistant (CDMA), lets everyday users build information agents from multiple sources using easy-to-use operations and visualizations. The agent can also be re-executed, reconfigured with different parameters, and reused as a part of other agents.

2. MOTIVATING EXAMPLE

Peter plans to buy a house in California. His main considerations are reasonable price, low crime rate, and decent high schools for his teenage kids. He is willing to make some trade-offs between these factors within some reasonable limits (i.e., getting a little more expensive house just to be in a lower crime area). There are various factors that contribute to housing price, but he wants to get a feel for what range of options he can expect. Here are some questions to which Peter might want to know answers to narrow down his choice of which city to consider: Is there any relationship between the average price per square-foot and the crime rate? The average high school ranking versus the crime rate? The average high school ranking versus the average price per square-foot?

The statistics for each city exist in different places on the web. To answer those questions, Peter needs to be able to pull information from different websites, integrate different information together, and use data mining techniques to derive the answers. Figure 1 shows how data from multiple sources can be combined to be in a form that is suitable for mining. We need to manipulate and integrate three data sources (by taking the average and joining). This integration step, if done by a person manually, could take days - not to mention the mining part to get the answers.

3. APPROACH

The example above illustrates one of a myriad of tasks where consumers have access to the necessary data but do not have a system that relieves them from manually manipulating the data in one simple session. We believe that it is possible to create an integrated system that combines data integration, monitoring, and mining as one process.

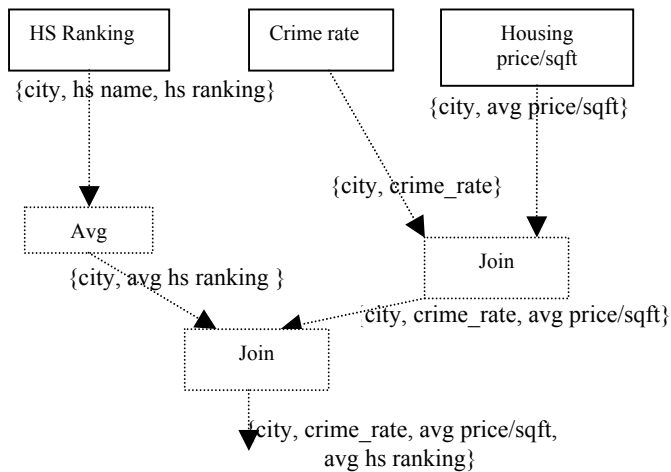


Figure 1: The overall operations for the data integration part of the task. Each data source is a table with its attribute names shown in {}.

By selecting only a necessary subset of operations, the system is simple enough for end- users to use while still allowing them to execute complex tasks. We show how CDMA fulfills this requirement using Peter’s example as a demonstration. CDMA can be divided into three parts. The Data Integration part retrieves and integrates data similar to what is shown in Figure 1. The Data Monitoring allows end-users to specify how the data should be collected. The Data Mining part specifies what mining operations to be performed on the data. Inspired by easy-to-use guiding steps like the Excel chart wizard, CDMA guides users through each stage by using the question-answering techniques described in our previous work on the Agent Wizard [2]. CDMA presents a user with a sequence of questions and creates an agent capable of manipulating data for the user’s task based on the user’s answers.

3.1 Data Integration

The data integration part of CDMA guides users to select which data source to use, and to transform and integrate all the data sources together as shown in Figure 1. This is done by asking users a series of questions. This approach is possible in the data integration phrase because we can impose a structure of how to integrate the data. At any point during the data integration phrase, we impose limited choices, so users can only choose one of the following operations: select which data source to use, manipulate the data source (i.e., filtering and averaging an attribute), integrate two data sources together, and output the final integrated data to the next stage.

For Peter’s example, he would start using CDMA by selecting three different data sources (housing price data from statistic data provider, high school ranking data from the department of education, and crime rate data from the disaster center). In CDMA, we provide data sources using wrappers [3]. A wrapper is a software agent that extracts data from a website. While CDMA does not focus on how to create a wrapper, Fetch Technologies (www.fetch.com) offers software that requires minimal training to build a wrapper. Wrappers differ from traditional database because these data sources can change frequently depending on websites.

After selecting the data sources, CDMA will guide Peter through the process of manipulating the data (i.e., taking an average of school ranking), and integrating the data from different data sources together. Figure 2 shows how CDMA asks Peter to join two data sources together. The output from this phrase would be an integrated table with city name, avg high school ranking, and crime rate. In our example, the attributes “city” from HS Ranking, Crime rate, and Housing price/sqft are the same. In the case where attributes with the same meaning might have different names, users can choose “compare over multiple attributes” and specify how each attribute from each source should be joined.

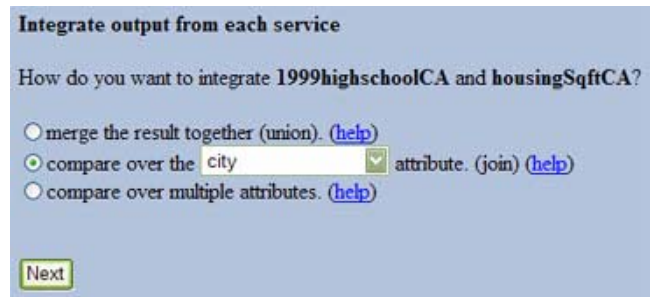


Figure 2: guiding users to join the result from two data sources together.

3.2 Data Monitoring

This part of CDMA focuses on the case where data is dynamic (i.e., stock market, hotel ticket price, weather). The user will be asked when to start retrieving the data, when to stop, how often should the data be sampled (i.e., once a day, twice a day, every 30 minutes), and when to start employing the mining algorithm. An addition to Peter’s example that involves data monitoring would be to monitor the average driving time from a particular set of cities to the place he works through a site such as Etak (www.etak.com).

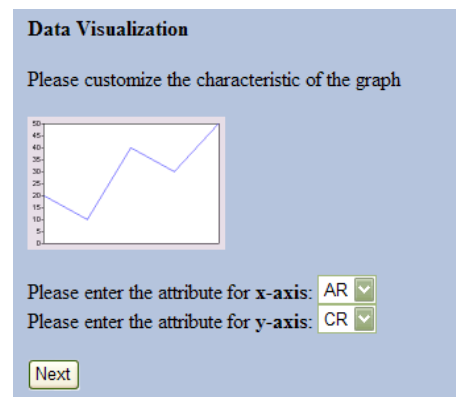


Figure 3: Creating a graph in CDMA

3.3 Data Mining

Data mining is the process of analyzing data from different perspectives to gain new insights -- relationships between attributes in data. Data Mining can become very complicated and we provide only a subset of possible operations as well as ways to visualize data that end-users understand and use. CDMA provides arithmetic operations across attributes (i.e., add, and multiply), basic statistic operations (mean, median, summation, and correlation), and ways to visualize data using graphs. CDMA

would guide users through these options to transform the data into the format that would give them different perspectives. In the case where monitoring is involved, the user can also select how to be notified when the result is ready (e.g., by email).

In Peter's case, to find out the relationships between the average price per square-foot (AP), the average high school rankings (AR),

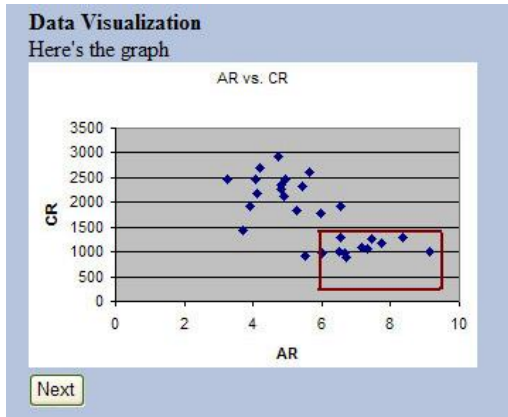


Figure 4: the output graph for AR vs. CR

and the crime rate (CR) of each city, he might choose to create three XY graphs (AR vs. AP, AP vs. CR, and AR vs. CR). Figure 3 shows how a user can create a graph and Figure 4 shows the output graph of AR vs. CR.

By viewing data in a graph instead of a table, Peter can see how each city fares in terms of AR and CR; there seems to be an inverse relationship between AR and CR. Peter can now decide what range of the parameters would be acceptable (i.e., $CR < 1500$ and $AR > 6$ region)

3.4 Output Agent

The final output of CDMA is an agent containing parameters that specify which data sources to use, how to integrate the data together, how the data should be monitored and mined, and how to notify the user with the results. By abstracting these parameters, the resulting agent can be reused (with different parameters), reconfigured (i.e., a user might want to change how often the data is monitored), or used as a part of a new agent; the agent's output from the data integration part can also be treated a data source input when building another agent.

4. RELATED WORK

In the business area, existing software either covers some subsets of functionalities or provides an integrated solution to a specific problem. SAS (www.sas.com), the market leader in data mining tools, offers a general purpose data mining tool that allows users to monitor and mine data from traditional databases. However, the software requires extensive training before someone can use it. Spss's Clementine offers a visual data mining tool that allows users to build a data mining plan by composing various types of nodes similar to that of Figure 1. While Clementine's visual interface provides the freedom for users to customize a plan that can be more complicated than CDMA, it also requires users to know machine learning algorithms and its language to compose expressions to derive new attributes. Furthermore, it is not clear how data monitoring might be easily configured in Clementine.

Instead of focusing on business problems, CDMA focuses on the consumer by providing an easy way to setup data monitoring and to build an expression that derives new attributes. CDMA also uses an integrated ontology to suggest users how to combine similar concepts across related attributes. DeepAnalysis (www.hammertap.com) has an integrated approach similar to CDMA, but it only focuses on helping Ebay sellers analyze auctions.

In the research area, Lixto [4] provides a visual user interface for users to build wrappers, integrate information agents, and monitoring data. However, Lixto does not support mining functionality and requires expert users to utilize the system. CAT [5] guides users through the process of building workflows but does not have the monitoring and mining capabilities. Personal Choice Point [6] integrates users' preference and provides ways to visualize trade-offs, but it only address a specific problem of buying a car, while CDMA can be applied to any task depending on data sources.

5. CONCLUSION

In this paper, we argue that there is untapped potential of how end-users can gain new knowledge from available data on the web. We address this problem by viewing data integration, monitoring, and mining as a single process. While trading off complexity in terms of functionalities with the ease of use, it is still possible to gain interesting insights when putting different data sources together.

6. ACKNOWLEDGMENTS

This material is based upon work supported in part by DARPA under Contract No. NBCHD030010, in part by AFOSR under grant numbers FA9550-04-1-0105.

7. REFERENCES

1. Etzioni, O., Knoblock, C.A., Tuchinda, R., and Yates, A., *To Buy or Not to Buy: Mining Airline Fare Data to Minimize Ticket Purchase Price*. In *Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2003. Washington, DC.
2. Tuchinda, R., and Knoblock, C.A. *Agent Wizard: Building Information Agents by Answering Questions*, In *Proc. of the IUI 2004* January 13-16, 2004, Funchal, Madeira, Portugal
3. Knoblock, C.A., Lerman, K., Minton, S., Muslea, I., *Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach*. IEEE Data Engineering Bulletin, 2000. 23(4): p. 33-41.
4. Baumgartner, R., Flesca, S., Gottlob, G., *Visual Web Information Extraction with Lixto*. In *VLDB*. 2001.
5. Kim, J., Spraragen, M., Gil, Y., *An Intelligent Assistant for Interactive Workflow Composition*, In *Proc. of the IUI 2004* January 13-16, 2004, Funchal, Madeira, Portugal
6. Fano, A., and Kurth, S.W., *Personal Choice Point: Helping Users Visualize What It Means to Buy a BMW*, In the *Proceeding of the IUI*. January 2003