Is There A Grand Challenge or X-Prize for Data Mining?

Gregory Piatetsky-Shapiro
KDnuggets
gps@acm.org

Chabane Djeraba
University of Lille
Chabane.Djeraba@lifl.fr

Lise Getoor
University of Maryland
getoor@cs.umd.edu

Robert Grossman
UIC & Open Data Group
rlg@opendatagroup.com

Ronen Feldman
University of Bar-Ilan & ClearForest
ronenf@gmail.com

Mohammed Zaki
RPI
zaki@cs.rpi.edu

ABSTRACT

This panel will discuss possible exciting and motivating Grand Challenge problems for Data Mining, focusing on bioinformatics, multimedia mining, link mining, text mining, and web mining.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining

General Terms

Measurement, Performance, Experimentation.

Keywords

Data mining, bioinformatics, multimedia mining, image mining, video mining, link mining, text mining, web mining, grand challenge, X-prize.

1. INTRODUCTION

Recently we saw several major scientific and engineering advances that were stimulated by a grand challenge/prize [1, 2]. The DARPA Grand Challenge produced great advances in robotic car navigation in 2005; X-prize led to the first successful commercial spaceflight; and RoboCup, whose goal is to developing a team of humanoid robots that can win against the human world soccer champion team by 2050, has greatly advanced robotic performance and created many enthusiasts.

Fig. 1 Robotic Grand Challenge

??? Data Mining Grand Challenge

Is there a grand challenge problem for data mining ?

This question is timely -- X-prize foundation is looking for additional fields where the prize can be created.

A good grand challenge problem should satisfy several criteria:

1) It should be relevant to data mining and knowledge discovery and be based on analysis of large volumes of data, preferably publicly available data.

2) It should be sufficiently important and difficult so that its solution will advance the field and benefit the society at large.

3) It should be interesting and exciting to attract researchers, public and press attention, and funding. This requires a simple and concise problem statement – one or two sentences.

4) The required domain knowledge should be relatively accessible.

5) Other groups are not actively working on this problem already.

Some potential ideas for a grand challenge include:

- Automatic tagging and classification of digital photos and images on the web
- Identifying all genes and potential therapeutic targets for cancer
- A text-mining and understanding system that can use the web to pass standard tests like SAT
- Discovering how, when and where the genes are expressed

We examine several current hot research areas including web mining, text mining, link analysis, video and image mining, and bioinformatics, and discuss possible proposals for an exciting and worthwhile grand challenge / X-prize for data mining.

2. MULTIMEDIA MINING (Djeraba)

Multimedia data is growing at enormous rate. While it contains much valuable information, because it is mostly un-structured or semi-structured, it is difficult (if not impossible) for people to extract the information without powerful tools.

Multimedia mining systems could automatically extract semantically meaningful information (knowledge) from multimedia data. This research overlaps several area: multimedia indexing and retrievals, multimedia semantics, pattern recognition, and multimedia usage mining.

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.
ACM 1-59593-339-5/06/0008.

The common points of these sub-area of research concern a large number of techniques that have been proposed ranging from simple features (e.g. color histogram for images, energy estimates for audio signal, texture, shapes) to more sophisticated systems like user multimodality (hand, body, eye) capturing and mining, speaker emotion recognition in audio, automatic summarization of TV programs, extracting relationships between video events, extracting eye-tracking pattern from several user scan paths, classifying images based on their content, extracting patterns in sound, categorizing speech and music, and recognizing and tracking objects in video streams, etc.

Multimedia mining involves several layers of transformations of features from low level to semantics, consisting of :

- Selecting multimedia data (e.g., football videos of good quality, during 2006 in Europe),

-Preprocessing data: data cleaning, normalization, transformation, feature selection,

- Mining data. The mining consists of identifying events (e.g., shooting the ball, goal, running) or objects (ball, football player: Zidane, football player: Dugarie), and it consists of discovering knowledge, such as relationships between objects or segments within multimedia (e.g., Zidane passing to Dugarie).

- Evaluating and interpreting knowledge in order to obtain the final application's knowledge.

Multimedia usage mining and multimedia semantics are two promising aspects of multimedia mining.

2.1 Multimedia usage mining

Multimedia usage mining place the user in the center of multimedia documents. It consists of extracting knowledge involving the usage of the user of multimedia data. The problem concerns:

- making multimedia trackable, saving all user operations (e.g., play, pause, visualize, eye fixation) and multimedia pieces concerned by these operations,

- mining user actions, considering for example, intra/inter video actions, the user is included in the loop of multimedia semantic.

2.2 Multimedia semantics

Multimedia semantics has been studied for quite some time. What is now needed is for researchers to develop approaches to extract semantics from multimedia documents so that retrievals using concept-based queries can be tailored to individual users. The semantic gap, or, as others put it, the semantic chasm, must be crossed. Multimedia usage mining coupled with domain ontology may be a revolutionary way to deal with the lack of semantics in multimedia information, and will certainly contribute to the hot domain of multimedia semantics.

2.3 Grand challenges

Two categories of challenges turn around usage and data.

1) Mining user behaviors in interactions with multimedia data to anticipate future behaviors or to diagnose medical or psychological conditions of the users. The challenge is to mine not only explicit actions (interactions, navigation), but also implicit reactions such as eye/gaze fixation, emotions (70% of people is emotion), heartbeat, respiration rate, stress, etc. The challenge is also to use non-intrusive sensors (e.g., cameras), rather than intrusive sensors.

2) Crossing the semantic gap between multimedia data and semantics. The challenge is to extract automatically the meaning of multimedia content so that exploitation (e.g., retrievals) using semantic information can be tailored to individual applications (security, marketing, business, etc.). The is a very difficult process considering the high volume, the complexity and the heterogeneity of multimedia data. Multimedia data is the most natural information-conveying vehicle but also the most complex to index and mine. The challenge is to generate metadata that describe the content and that may be exploitable in applications.

3. GRAND CHALLENGES FOR LINK MINING (Getoor)

Many data mining domains of interest today are best described as networks or graphs. Common examples include social networks, biological networks and communication networks. There are interesting practical and theoretical issues in mining such data; in some cases there are also difficult legal and privacy issues. While certain link mining problems such as mining call records for terrorist activity have received a lot of recent media attention, they do not meet the requirements of a grand challenge as proposed by this panel. In this presentation, I will propose other potential link mining problems that make appropriate grand challenges.

4. TECHNICAL & PRAGMATIC GRAND CHALLENGES (Grossman)

Following [3], we divide data mining grand challenges into three categories:

a) Technical Grand Challenges: the algorithmic, systems, and application challenges related to important technical problems, such as scaling data mining algorithms, mining complex data types, mining mixtures of data, etc.

b) Theoretical Grand Challenges: developing the theoretical foundations for working with large, complex data.

c) Pragmatic Grand Challenges: concerned with data preparation and integration, and developing, deploying and embedding statistical and data mining models.

For comparison, let's consider technical, theoretical and pragmatic challenges from related computer science fields.

For example, technical challenges in computing include building a petaflow computer, developing a terabit network, or building an autonomous vehicle that can cross a desert. Theoretical challenges in computing include proving P != NP or developing new algorithms for factoring integers with a thousand digits. Finally, pragmatic challenges in computing include building databases that can manage trillions of rows or ones that can break the current TPC-C benchmark.

We discuss technical challenges involving the Sloan Digital Sky Survey (SDSS), a multi-terabyte data set of astronomical data, focusing on problems related to the classification of objects in the SDSS, to analyzing SDSS data remotely, and to the distributed integration of SDSS data with other astronomical data sets, such as the 2 Micron All Sky Survey Data (2MASS).

I will also discuss several pragmatic challenges involving deploying and embedding analytics. In particular, I'll discuss the development of a standard-based infrastructure that can meet the requirements needs of those deploying and embedding analytic models, but is also powerful enough to capture the data preparation and data integration that is also required in applications.

Finally, I'll discuss the development of community benchmarks so that we can measure our progress in developing scalable algorithms for clustering, regression and other standard problems.

5. GRAND CHALLENGES FOR TEXT MINING (Feldman)

Text Mining is an exciting research area that tries to solve the information overload problem by using techniques from data mining, machine learning, NLP, IR and knowledge management. Text Mining involves the preprocessing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate representations, the techniques to analyze these intermediate representations (distribution analysis, clustering, trend analysis, association rules etc) and visualization of the results.

Here are some of the grand challenges that are facing the text mining research area:

1. Most text analytics systems rely on accurate extraction of entities and relations from the documents. However, the accuracy of the entity extraction systems in some of the domains reaches only 70-80% and creates a noise level which prohibits the adaptation of text mining systems by a wider audience. We are seeking domain independent ad language independent NER (named entity recognition) systems that will be able to reach an accuracy of 99-100%. Based on such system, we are seeking domain independent ad language independent relation extraction systems that will be able to reach precision of 98-100% and recall of 95-100%. Since the systems should work in any domain they must be totally autonomous and require no human intervention.
2. Based on systems developed in step 1, we would like to have text mining systems that will be able to pass standard reading comprehension tests such as SAT, GRE, GMAT etc. Systems that will be able to pass the average scores will win the grand challenge. The systems can utilize the web when answering the test questions.

3. Text Analytics systems today are pretty much user guided, and they enable users to view various aspects of the corpus. The grand challenge is to have a text analytics system which is totally autonomous and will analyze huge corpuses and come up with truly interesting findings that are no captured by any single document in the corpus and are not known before. The system can utilize the internet to filter findings that are already known. The "interest" measure which is totally subjective will be defined by a committee of experts in each domain. Such systems can then be used for alerting purposes in the financial domain, the anti-terror domain, the biomedical domain and many other commercial domains. The system will get streams of documents from a variety of sources and send emails to relevant people if an "interesting" finding is detected.

6. DATA MINING GRAND CHALLENGES IN BIOINFORMATICS (Zaki)

Large-scale databases from sequencing projects, microarray studies, gene-function studies, protein-protein interactions, comparative genomics, structural biology, and open source journal articles, are growing at rapid rates. The challenge in systems biology is to connect all the dots from the diverse molecular, cellular, organism and environmental data sources to deduce how sub-systems and whole organisms work. We need to decipher the language of life – the language of the genome, protein folding, developmental pathways, and much more. Some of the scientific challenges include how to infer the complete network that controls genes, to infer the complex signaling pathways that regulate the health of the cell, to infer how cell specialization occurs during development, and so on.

There are numerous computational challenges in collecting, indexing, searching and mining these vast data sources. I will try to outline a few data mining challenges in bioinformatics such as in gene regulation, protein folding, and so on. For example, discovering how, when and where the genes are expressed is one such grand challenge. The goal is to infer the transcription machinery and the gene regulatory circuitry. The central actors include, proteins that shut down/activate genes, transcription factors that bind to specific regions, chromatin proteins that may expose different genes, microRNA and other small RNA, etc. Mining the diverse sources of public data pertaining to these actors will be a crucial component in piecing together the bigger picture.

7. REFERENCES

[1] "Grand challenges spur grand results - Private groups are offering big cash prizes to anyone who can solve a range of daunting problems". The Christian Science Monitor, January 12, 2006 http://www.csmonitor.com/2006/0112/p13s01-stss.html

[2] Prize for DNA Decoding Aims to Fuel Innovation, Wall Street Journal, Jan 27, 2006

[3] Usama Fayyad and Robert L. Grossman, Grand Challenges for Data Mining: Technical, Theoretical, and Pragmatic, submitted for publication.