Research Statement
Sofus A. Macskassy
December 2004

Research Overview

My long-term vision is to build a personal information management and access (PIMA) system which filters and provides ubiquitous access to information that is up-to-date, relevant and of interest to its user. This vision is broad and has led me to investigate and use methods from areas such as relational learning, text classification, machine learning, information retrieval, data mining, wireless information access and human computer interaction.

My core research is focused on applying machine learning and information filtering techniques in order to assemble some of the core building blocks for a PIMA system. I have focused both on the long-term challenge of integrating and exploiting many sources of information in performing PIMA filtering tasks, as well as developing algorithms and building small-scale systems and toolkits that can be used in the shorter term.

My work at Rutgers, some of which became the core of my dissertation, focused on email filtering [7], wireless information access [8], business news filtering [11] and creating a new technique for using numerical values in text classification [10]. I joined NYU as an Associate Research Scientist after completing my Ph.D., and began working with Foster Provost in the area of statistical relational learning (SRL). This has been exciting because this new field has many open research questions which relate directly to problems which need to be solved for the general PIMA system. For example, most information we deal with on a daily basis is composed not of isolated objects but rather a complex network of interconnected pieces of information. My current work is focused on classification in networked data [12, 13].

I have ordered the remaining part of this document into the following five categories:

1. Classification in Networked Data: Networked data consist of objects interconnected in complex information networks. Methods using only the network and class-labels of neighbors turn out to perform remarkably well.
2. Wireless Information Access: I built a system which enabled ubiquitous information access using multiple wireless devices such as PDAs and email pagers.
3. Information Filtering: I built an email filtering system to forward important emails to a pager. I also built a system to filter financial news stories such as identifying stories that were precursors to large stock movements.
4. Adding Numbers to Text Classification: text classifiers generally do not handle numerical features in a meaningful way. I created a technique to inject numerical features into a text classification system and showed that it improved the performance of the classifiers.
5. Future Directions: Combining text classification and network learning, active acquisition of information, minimizing the need for explicit user feedback, explaining the model, handling dynamic information.

Classification in Networked Data

Most data are relational in nature and a PIMA system can use such information in order to improve its performance. For example, emails share senders and receivers, financial news stories mention people and companies, research papers cite each other and share authors, web-pages link to each other, etc. Relational data can be thought of as a graph, where the nodes represent information objects such as emails or companies, and the edges represent the relations between them, such as email having the same sender or two companies being competitors. Most evaluations of statistical relational models (SRL) (cf. PRMs [17]) have been done by comparing the SRL models to
traditional non-relational machine learning methods. My research shows that looking at relational-only models is another dimension of the baseline comparison that should not be ignored and that network-only methods perform surprisingly well [12, 13]. In order to perform this research, I developed a powerful Network Learning Toolkit (NetKit), which is available as open-source on my website. The toolkit, written in Java 1.5, is modular and allows for easy plug-and-play of its various components enabling it to instantiates and mimic many existing network learning algorithms from prior work (e.g., [2, 6, 12].) It also opens up the possible design space of relational learners by allowing combinations of components that could until now not be realized. Being open-source and publicly available will also make it easier for the statistical relational learning community to do comparative studies, something which has until now not been possible.

My work differs from most work in relational learning in that I use, and get much leverage from, the fact that often labeled instances are interconnected with the instances which need to have their class labels estimated. Such a setting is seen in many real-world applications, such as telephone fraud, counterterrorism, classifying publicly traded companies, patent searches, etc. I have also found that many funding agencies are very interested in applying methods such as these to counterterrorism-related tasks and much of my funding is directly tied to such tasks. The current climate suggests that this situation is likely to continue.

Wireless Information Access

Ubiquitous access to information is an important aspect of a PIMA system. I built a general testbed that supported the use of multiple wired and wireless devices (including RIM two-way pagers and Palm VII wireless PDAs), which handled access to email, the Web, and personal files [8]. The system further tailored information presentation to each device and allowed its users to send, or forward, information between the devices. The system tracked the use of these devices to learn not only user interests but also on which device the user was accessing the information.

The system was based on placing a PIMA agent at a central location where the user profile is kept. The agent would filter, proactively get information from the internet, and possibly push this information to the user (to an appropriate device) if it was deemed interesting or important enough.

Information Filtering

Another important aspect of a PIMA system is information filtering. To filter information, you need to have a profile of the user. This is at best a difficult task for various reasons. There is a huge amount of information available, the number of information sources available is continuously growing, and users generally have difficulty in precisely stating their own interests. Finally, it is unreasonable to expect a user to spend a lot of time maintaining and building a profile. I am using machine learning and information retrieval techniques to help alleviate this burden for the user.

My work on information filtering spans both email [7] and financial business news [11], each using a different technique to build a model of the user’s interest. In the email domain, I focused on the problem of learning to predict whether a user was going to read an email on his or her pager. I built a tracking agent for gathering implicit user feedback (watching which emails were actually being read). I gathered usage data over a period of more than one year for three different users. I showed, using this data, for each user separately, that it was possible to learn to predict which emails to forward to the pager [7].

In the domain of filtering financial news stories, I focused on the case where a user is able to specify a prospective interestingness criterion but cannot articulate what it is about a story that would make it likely to have this characteristic [11]. For example, a financial news story is interesting if the stock price of a company mentioned in the story moves significantly in the hour following the story being published. In fact, the user might be just as interested in being told what characterizes such a story as he or she is in the story itself. This work led to a new methodology of getting user interest and learning from it [11]. This work required being able to integrate multiple sources
of information (such as news stories and stock data), as well as using large amounts of historical data in an effective way. It also proved necessary to analyze the final learned model in detail to explain what was learned to the user (a financial analyst) both for validation as well as for getting more insight into the domain [11].

**Adding numbers to Text Classification**

The information filtering work described above made use of commonly used text classification algorithms such as Naive Bayes [3], Rocchio [16, 5] and Maximum Entropy [14]. However, these methods do not handle numerical features in a meaningful way—they either ignore them or consider them as generic tokens. Such an approach ignores the rich information in numerical features such as length of an email message, number of seconds since last email from a sender or stock prices of a company mentioned in a news story. I therefore developed a new technique for using numerical features with text-classifiers, and showed that I could get significant improvements in performance by using this technique [10, 9].

**Future Directions**

My network learning toolkit is maturing to the point where I can look at broader questions of information integration. I have until now looked at text-only and network-only learning. I now want to integrate the two and investigate techniques for more advanced information integration by actively gathering information and populating the network.

Such an environment is interesting from a PIMA point of view because there are many PIMA questions we can now focus on. Some of the interesting ones I plan to investigate are: (1) How much information needs to be labeled to perform well? We do not want to bother the user more than necessary, and so there is a “cost” associated with labeling information, and therefore we want to minimize this cost. (2) If we do need to get information labeled, which information should that be? This has interesting ties to active learning. (3) How do we intelligently acquire more, and hopefully new, information? By actively seeking new and interesting information, we get an information network tailored to the user’s interest. (4) The information is dynamic as new instances and links are being generated constantly. How does this effect classification? If there is no activity between two linked vertices in a graph, should the link be removed, or should its importance decay?

There are also other aspects of PIMA which I plan to investigate in greater depth. (1) Transparency of the model. How does the system explain to the user why it performed a certain action? The system needs to either have a transparent model or be able to explain itself. There has been some work in this area, but not enough (cf. [1]). (2) Ubiquitous information access has come a long way in recent years, and the power of mobile devices is growing rapidly. How do we take advantage of this to let the users have ubiquitous information access at their fingertips, maybe bringing part of their profile with them, while also preserving privacy.

In the near-term, I plan to follow up on my univariate case study in network learning [13] and investigate: (1) How much the characteristics of the graph can tell us about the expected performance of a network learning system. For example, do degree distribution, topology, or assortativity (cf. [15]) correlate well with expected performance of all learners or, failing that, any specific learner? (2) Are there methods which perform better when few instances are labeled vs. when many instances are labeled? (3) How much predictive power is there in the various pieces of information in the graph, such as intrinsic variables, class labels, class labels of neighbors, intrinsic variables of neighbors, etc. (4) What kind of interaction is there between the three components of a network learning system in the multivariate case, and are there combinations that work well together? There has been a preliminary study on this using synthetic data [4] and in the univariate case [13]. However, it remains to be seen how much of this carries over to real-world data with multivariate models. I plan to perform an in-depth systematic study of these to set the foundation for my PIMA work described above.
References


