Web Mining

Data mining provides the information that hundreds of businesses rely on, from international banks to communication conglomerates and online archives. The technologies behind data mining are maturing rapidly and giving increasingly accurate results.
We need to explore the technical, ethical and social implications of data mining, especially as web consumers release increasing amounts of personal information on to the web through "Web 2.0" sites such as Facebook and Twitter.

Legal Issues

The 2005 scandal where U.S. President Bush admitted that the government was using data mining on a massive scale to monitor American citizens brought the issue of data mining to public attention. Data mining itself may not be illegal, nor collecting large amounts of data, but the knowledge that can be discovered by analysing all of a persons' online activity may reveal a large amount of personal information which was not explicitly put on the web; the use of this extracted information can be considered questionable from a legal point of view.

In America, the Federal Agency Data Mining Reporting Act of 2007 states that the Department of Homeland Security must report to congress all the details of any data mining activity undertaken. This is a special case as the DHS can gain access to both public and private databases, but web search companies and banks do not need permission to perform data mining or inform their customers, nor are they accountable for their data mining activities.

Corporate Responsibility

Many companies use data mining to improve the service to customers. For example search engines mine user activity to provide more appropriate search results, but when you perform a search on Google you are not informed that all of your search queries and the websites you visit from the results are being stored and analysed. Similarly social networking sites mine your personal information to serve you "relevant" advertisements and suggest friends.

What responsibilities do companies have when mining their customers' data? i.e. Should companies have to inform customers about data mining practices. And do companies have a responsibility to report suspicious, possibly criminal behaviour to the authorities?

Ethical Implications

The great power of data mining applications comes from the ability to extract previously unknown knowledge from large sets of data. This knowledge could be information on a person which they have not explicitly stated, for example it is possible to determine criminal activity by mining social network behaviour and search queries. So to what extent should rules learned by data miners be trusted, and should they be admissible in a court? It has been shown that people act differently in an online space to how they might in the offline world, exploring fantasies that are only made possible by the web.

So how much does someone's behaviour online relate to their behaviour in the 'real world'?

The inherent personal security of your data being spread thinly in multiple places does not exist any more, thanks to the internet and data mining software. So how should we use this new resource?

Why We Need It

Data mining can give insights into large data sets which would be impossible to study manually. It can detect patterns and irregularities in bank transactions and travel plans allowing criminals to be caught.

Web mining is the process of analysing data posted online. Some of the biggest applications include social network analysis and search engine optimisation. Web mining is invaluable for online archives and most search engines. It can also be used to create personalised services such as Facebook's recommended friends and targeted advertising functions.

Technical Considerations

There is an active area of research within data mining regarding privacy protection on the technical level. There are two generalised data mining situations to be considered:

Mining information about individuals	Mining to find trends in a population
Examples are targeted advertising and friend suggestions. Protecting privacy in this situation is obviously important. The most basic method to protect user privacy is to anonymize the data, although this system is not infallible as recently demonstrated in the national press: a woman was identified personally by her search queries on aol.com. Fule and Roddick have proposed a system which automatically detects rules about sensitive data and presents a dialog to an administrator to decide whether to accept or ignore that rule.	Examples are decisions on e-commerce site layout and advert placement. In situations where rules learned by a data miner are not specifically targeted at individuals, but are used for a more general analysis, the privacy of the individuals still needs to be preserved. Some sensitive trends may be discovered relating to criminal intent or religious persuasion which could be used against the individual, for example 'GayDar' recently in the press which uses information mined from Facebook to determine someone's sexual orientation. Many methods of protecting privacy have been suggested, such as perturbing individual records in a database or injecting false records, both of which can give good results while preserving the privacy of individuals.

Conclusions

The practice of mining data on the web is very common, even routine, but the public are left relatively in the dark about who has access to their information and what it is being used for, continuing to post increasing amounts of personal data online. The law regarding web mining is as immature as the law regarding the web itself, and desperately needs to be reviewed. Regulations regarding web mining practises, privacy, identity and legal responsibilities need to be created to protect the web user from identity theft, discrimination and possible incrimination.

Web Science
Research

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Simon Hearne MEng
University of Southampton
sjmh105@ecs.soton.ac.uk