Privacy Preserving Collaborative Filtering with Sharemind

Richard Sassoon
26th November 2008
Topics

• Main concepts:
  • Data Mining
  • Privacy Preserving DM
  • Collaborative Filtering
  • Association Rule Mining

• About Sharemind:
  • What is it?
  • Security guarantees

• Proposed Application:
  • Description and privacy issues

• Proposed Implementation in Sharemind:
  • Description
  • Another approach

• Conclusions
Main concepts: Data mining

• What is it?
  Statistical analysis of data warehouses in order to find patterns and/or trends that are useful for decision making processes

• How is it done?
  There are several algorithms, depending on the purpose of the mining. We focus on association rule mining.
• Privacy concerns with traditional data mining:
  • Organizations who have large data sets want to extract relevant associations from them, but without leaking any individual information
  • They should also be able to join their database with other institutions, without being able to learn anything but the output
  • Data donors are afraid that their data can be misused or simply do not want that their data are disclosed
  • There is always some information leak
Main concepts: Privacy preserving data mining

• Methods for privacy preserving data mining:
  • Randomized approach – perturbs single entries but preserves global statistical properties
  • Cryptographic methods – makes use of cryptographic primitives such as oblivious transfer, to achieve privacy

• Problems with these methods:
  • Randomized approach – no security guarantees for individual records and perturbation reduces the output's precision
  • Cryptographic methods – computation can be costly if numbers of participants is beyond a few hundreds
Main concepts: Collaborative Filtering

- Collaborative filtering is a method for generating suggestions to a user of some application, mainly e-commerce, e.g., Amazon.
- It makes use of statistical and knowledge discovery techniques.
- Identifies customers with similar taste to the interacting one and recommends items they have liked.
- Here we are concerned about similarity of items, i.e., items that are likely to appear together, and to be able to find such similarities we propose the use of association rule mining.
These recommendations are based on items you own and more.

   Average Customer Review: ★★★★★ (197)
   In Stock
   List Price: $14.95
   Price: $10.17
   90 used & new from $2.22

2. **The Salmon of Doubt: Hitchhiking the Galaxy One Last Time**
   by Douglas Adams (April 26, 2005)
   Average Customer Review: ★★★★☆ (109)
   In Stock
   Price: $7.99
   53 used & new from $2.50

3. **Dirk Gently's Holistic Detective Agency**
   by Douglas Adams (Jun 1, 1991)
   Average Customer Review: ★★★★★ (89)
   In Stock
   Price: $7.99
   67 used & new from $0.67
Music Recommended by Last.fm

**Vaiko Epnik ja Eliit**
- 99,301 plays (1,314 listeners)
- **Add to Your Library**
- indie, estonian, est, experimental, pop

Eliit is a group of contributing musicians and singers from Estonia, featuring such names as Vaiko Epnik, Tursk, Foto, Raul Saaremets, Jaan Peik, Aapo Ikäheimo, Chalice, Kildj Epnik, Evelin Pang and Taavi Tammjärvi. Material on albums is almost entirely made by Vaiko himself. [Read more]

**Leslie Da Bass**
- 9,988 plays (578 listeners)
- **Add to Your Library**
- electronica, estonian, electronic, alternative, dnb


**Steve Vai**
- 3,211,462 plays (146,663 listeners)
- **Add to Your Library**
- guitar virtuoso, instrumental rock, rock, instrumental, guitar

Steven Vai (born 6th June, 1960 in Carle Place, New York) is a Grammy Award winning guitarist, composer and record producer. [Read more]

**John Petrucci**
- 624,300 plays (53,570 listeners)
- **Add to Your Library**
- Joe Satriani (721 plays)
- Eric Johnson (858 plays)
- Dream Theater (272 plays)
- Yngwie Malmsteen (91 plays)
Main concepts: Association Rule Mining

• In order to find significant association rules, it is necessary to find frequent itemsets and then rules based on these sets. We define:

• A set of attributes $I = i_1, ..., i_n$, called items. A set $X$ containing $k$ items from $I$, called itemset

• A database of transactions $T$, where each transaction is represented by 0's and 1's, indicating if item $i_k$ of set $I$ was of interest (1) or not (0).

• The support $\text{supp}(X)$ of an itemset $X$ is defined as the number of transactions $t$ such that $X \subseteq T$
Main concepts: Association Rule Mining

In order to find significant association rules, it is necessary to find frequent itemsets and then rules based on these sets. We define:

- An association rule is of the form $X \Rightarrow Y$ such that $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The itemset $X$ is called the antecedent and the itemset $Y$ is the consequent.

- The support $supp(X, Y)$ of an association rule $X \Rightarrow Y$ is the support of $X \cup Y$.

- The confidence of an association rule $X \Rightarrow Y$ is the conditional probability of having $Y$ in a transaction, given that $X$ is also present in that transaction:
  $$confidence(X, Y) = \frac{supp(X, Y)}{supp(X)}$$

- The association rules of interest will be the ones that satisfy both support (statistical significance) and confidence (rule's strength) thresholds.
About Sharemind

- **What is it?**
  - A framework for privacy preserving computation that uses additive secret sharing scheme in a multi-party environment
  - A secret value is divided into shares, by a controller node, and sent to 3 dedicated miners that are responsible for performing operations (addition, multiplication or comparison) on these shares

- **Security guarantees:**
  - Sharemind is information-theoretically secure in an honest-but-curious model. It is assumed that the parties will not collude with each other
Proposed Application

For the analysis of a Sharemind solution to privacy preserving association rule mining we propose the following application:

- A person in a museum wants to retrieve some information about some work of art, e.g., paintings (French, Italian, Dutch, ...), drawings, sculptures, antiques, etc.

- Assume that the person has a portable device for making such a query, through a network, that is in turn sent to some service which is able to provide the requested information.
Proposed Application

• For the analysis of a Sharemind solution to privacy preserving association rule mining we propose the following application (cont.):

  • Thanks to algorithms like Apriori that generate association rules for the transactions database, it is possible to find relevant suggestions to the user, for example, "if user liked drawings and italian paintings, he will like roman antiques"

  • Sharemind already has an implementation of Apriori that can be used for this purpose.
Proposed Application

• We consider 2 options for the user:
  
  • Special purpose PDA handed out at the entrance of the museum, possibly for a small fee
  
  • Own mobile via some service interface
Proposed Application

- Special purpose PDA:
  - It is dedicated for providing information and suggestion service for the museum
  - Has a content provider embedded
  - Send the interest of the user to Sharemind that in turn sends suggestions back, at the same time that displays the information requested
  - No privacy concerns in this case
Proposed Application

- Own mobile:
  - Needs to download a special interface, that will be used to communicate with Sharemind
  - Needs an external content provider in order to receive the desired information
  - Mobile and requested information could be identified during the data gathering process
Proposed Implementation

• We describe the overall solution for the situation where the user uses his mobile to access the service:

1) We need 4 tables in a shared database:

• contents $C = \{\text{code(index)}, \text{content}\}$;
• lookup $L = \{\text{code, category}\}$;
• associations $A = \{X \Rightarrow Y\}$;
• transactions $T = \{\text{id, paintings, sculptures, ...}\}$;
Proposed Implementation

• We describe the overall solution for the situation where the user uses his mobile to access the service:

  2) Split a traditional content database into C, where the contents are stored in shares.

  3) User downloads service interface, if it's the first time using it, and can start making requests. The interface uses Sharemind's controller module.

  4) When the user makes a request, this request is split into three shares and sent to the miners. What is sent is the work of art's unique code.
Proposed Implementation

• We describe the overall solution for the situation where the user uses his mobile to access the service:

  5) The miners initialize a comparison protocol on these shares to find the requested information in the shared contents database. At the same time another comparison is instantiated on L in order to find the kind of work of art the user is interested, and with this information it's possible to start a final comparison on the association rules database with the objective of finding a set of rules that can provide some good suggestions to the user.

  6) When the set of rules has been determined, it is not necessary that all the consequents are suggested. It is possible to select, at random, k of them, find some random representatives in the contents database and use their title and location attributes as the suggestion, and keep the whole content in a shared cache in case the user accepts the suggestion, thus saving a new search in the database.
We describe the overall solution for the situation where the user uses his mobile to access the service:

7) The user interface receives the shares of the content requested and the suggestions, joins and process them, showing a human readable format to the user. It also stores the kind of work of art that was of interest, in order to compute the user's transaction and also to send it in some next request, so another set of rules can be chosen.

8) When the user finalizes the interface, his final transaction, e.g, \{0, 1, 1, 0, \ldots\} is uploaded to T.
Proposed Implementation

- Implementation Architecture:
Proposed Implementation

• Some considerations:

  • One important aspect to be considered is the latency of this approach since, as of this writing, the comparison protocol is slow.

  • After the actual implementation of this scheme and tests with real data it will be possible to verify if the time constraints are achieved.

  • The computation of the association rules is an expensive one and should be done periodically. The result should be stored in A in order to generate faster responses to user requests.
Another approach:

We could also consider a more granular approach to this application if instead of categories of work of art we consider the individual piece, e.g., 'Monalisa', 'Babel', 'The Thinker', . . . We would have then rules in the format \{Monalisa, Babel\} => \{The Thinker\}

Since this would generate too many columns we should think of a sparse matrix representation, where only non-zero entries are stored, and an association rule mining algorithm designed for this kind of representation.
Conclusions

• We have seen some basic concepts for Data Mining, Collaborative Filtering and Association Rule mining.

• The Sharemind framework was briefly revisited, as a background for our proposed application.

• It is possible to think of other applications, such as suggestion of news, that could use the same kind of implementation.

• The solution in Sharemind could not perform so well, but in order to protect its privacy one may cope with a slow down in the system.

• The framework is under development so we may expect improvements that will help its use in real world applications.
Thanks for the attention

Questions?