Business Analytics

Raghav Kumar Gautam

Department of Technology & Information Management
University of California, Santa Cruz

May 11, 2010
Outline

1. Can Analytics Make a Difference?
   - Netflix in 1997
   - Netflix’s Success

2. Business Analytics Overview

3. Data Mining Basics
   - Stages of the Data Mining Process

4. Association Rules and Frequent Item set
   - Example
   - Terminology
   - Frequent Item set Mining
Can Analytics Make a Difference?

1. Netflix in 1997
   - Netflix's Success

2. Business Analytics Overview

3. Data Mining Basics
   - Stages of the Data Mining Process

4. Association Rules and Frequent Item set
   - Example
   - Terminology
   - Frequent Item set Mining

Outline

Raghav Kumar Gautam
Idea: Flat monthly fee based video-rental with no late fees

Shortcomings

- Strong competition already existed from existing video rental chains Eg: Blockbuster
- Would people
  - Order movies online?
  - Wait for “snail mail” to deliver it?
  - Go back to mailbox to return it?

Would it get customers at the end of the day?
Netflix in 1997

- **Idea:** Flat monthly fee based video-rental with no late fees
- **Shortcomings**
  - Strong competition already existed from existing video rental chains E.g. Blockbuster
  - Would people
    - Order movies online?
    - Wait for “snail mail” to deliver it?
    - Go back to mailbox to return it?

- **Would it get customers at the end of the day?**
Outline

1. Can Analytics Make a Difference?
   - Netflix in 1997
   - Netflix’s Success

2. Business Analytics Overview

3. Data Mining Basics
   - Stages of the Data Mining Process

4. Association Rules and Frequent Item set
   - Example
   - Terminology
   - Frequent Item set Mining
<table>
<thead>
<tr>
<th>Netflix’s Success</th>
</tr>
</thead>
</table>

- $5 Million in revenue in 1999.
- $1 Billion in 2006.
- Reason for success:
  - Analytics

  - The average user rated more than 200 films
  - Netflix crunches rating of all the users to predict what they will like
Netflix’s Success

- $5 Million in revenue in 1999.
- $1 Billion in 2006.
- Reason for success: Analytics
  - The average user rates more than 200 films.
  - Netflix crunches rating of all the users to predict what they will like.
Netflix’s Success

- $5 Million in revenue in 1999.
- $1 Billion in 2006.
- Reason for success:
  - Analytics
    - The average user rates more than 200 films
    - Netflix crunches rating of all the users to predict what they will like.
Netflix’s Success

- $5 Million in revenue in 1999.
- $1 Billion in 2006.

Reason for success:

- Analytics
  - The average user rates more than 200 films
  - Netflix crunches rating of all the users to predict what they will like.
Netflix’s Success

- $5 Million in revenue in 1999.
- $1 Billion in 2006.
- Reason for success:
  - Analytics
    - The average user rates more than 200 films
    - Netflix crunches rating of all the users to predict what they will like.
Big Picture of Business Analytics

Business Analytics Software

Performance Management Tools and Applications
- Financial Performance and Strategy Management Applications
  - Budgeting, planning, consolidation, profitability mgmt./ABC, scorecards
- CRM Analytic Applications
  - Sales, customer service, contact center, marketing, Web site analytics; price optimization
- Supply Chain and Services Operations Analytic Applications
- Workforce Analytic Applications

Business Intelligence Tools
- Query, Reporting, Analysis
- Advanced Analytics (includes data mining and statistics)

Spatial Information Management Analytic Tools

Data Warehouse Platform
- (DW management and generation)
Data Mining Basics

- **Definition**: Discovery of useful summaries of data.

- **Related Areas**:
  1. Statistics
  2. Artificial Intelligence, where it is called machine learning
  3. Researchers in clustering algorithms
  4. Visualization researchers
  5. Databases: when the data is large and the computations complex
Outline

1. Can Analytics Make a Difference?
   - Netflix in 1997
   - Netflix’s Success

2. Business Analytics Overview

3. Data Mining Basics
   - Stages of the Data Mining Process

4. Association Rules and Frequent Item set
   - Example
   - Terminology
   - Frequent Item set Mining
Stages of the Data Mining Process

- **Data gathering:** e.g., from Data warehouses, Web crawling.
- **Data cleansing:** eliminate errors and/or bogus data, e.g., patient fever = 125.
- **Feature extraction:** obtaining only the interesting attributes of the data.
- **Pattern extraction and discovery:** This is the stage that is often thought of as data mining, and is where we shall concentrate our effort.
- **Visualization of the data.**
- **Evaluation of results:** not every discovered fact is useful, or even true! Judgment is necessary before following your software’s conclusions.
Stages of the Data Mining Process

- Data gathering: e.g., from Data warehouses, Web crawling.
- Data cleansing: eliminate errors and/or bogus data, e.g., patient fever = 125.
- Feature extraction: obtaining only the interesting attributes of the data
- Pattern extraction and discovery: This is the stage that is often thought of as data mining, and is where we shall concentrate our effort.
- Visualization of the data.
- Evaluation of results: not every discovered fact is useful, or even true! Judgment is necessary before following your software’s conclusions.
Stages of the Data Mining Process

- Data gathering: e.g., from Data warehouses, Web crawling.
- Data cleansing: eliminate errors and/or bogus data, e.g., patient fever = 125.
- Feature extraction: obtaining only the interesting attributes of the data
- Pattern extraction and discovery: This is the stage that is often thought of as data mining, and is where we shall concentrate our effort.
- Visualization of the data.
- Evaluation of results: not every discovered fact is useful, or even true! Judgment is necessary before following your software’s conclusions.
Stages of the Data Mining Process

- **Data gathering:** e.g., from Data warehouses, Web crawling.
- **Data cleansing:** eliminate errors and/or bogus data, e.g., patient fever = 125.
- **Feature extraction:** obtaining only the interesting attributes of the data.
- **Pattern extraction and discovery:** This is the stage that is often thought of as data mining, and is where we shall concentrate our effort.
- **Visualization of the data.**
- **Evaluation of results:** not every discovered fact is useful, or even true! Judgment is necessary before following your software’s conclusions.
Stages of the Data Mining Process

- Data gathering: e.g., from Data warehouses, Web crawling.
- Data cleansing: eliminate errors and/or bogus data, e.g., patient fever = 125.
- Feature extraction: obtaining only the interesting attributes of the data.
- Pattern extraction and discovery: This is the stage that is often thought of as data mining, and is where we shall concentrate our effort.
- Visualization of the data.
- Evaluation of results: not every discovered fact is useful, or even true! Judgment is necessary before following your software’s conclusions.
Stages of the Data Mining Process

- Data gathering: e.g., from Data warehouses, Web crawling.
- Data cleansing: eliminate errors and/or bogus data, e.g., patient fever = 125.
- Feature extraction: obtaining only the interesting attributes of the data
- Pattern extraction and discovery: This is the stage that is often thought of as data mining, and is where we shall concentrate our effort.
- Visualization of the data.
- Evaluation of results: not every discovered fact is useful, or even true! Judgment is necessary before following your software’s conclusions.
Can Analytics Make a Difference?

1. can Analytics Make a Difference?
   - Netflix in 1997
   - Netflix’s Success

2. Business Analytics Overview

3. Data Mining Basics
   - Stages of the Data Mining Process

4. Association Rules and Frequent Item set
   - Example
   - Terminology
   - Frequent Item set Mining
An Example

Customers comes to shop and buys different items in his/her basket.

- C1: {Bread, Milk, Eggs}
- C2: {Fish, Bread, Eggs}
- C3: {Carrots, Bread, Eggs}
- C4: {Eggs, Apples, Breads}

Can we do something predict something?

- Whenever a person purchases Bread, he/she will also purchase Eggs.
- How do we do this on a large scale?
An Example

- Customers comes to shop and buys different items in his/her basket.
  - C1: \{Bread, Milk, Eggs\}
  - C2: \{Fish, Bread, Eggs\}
  - C3: \{Carrots, Bread, Eggs\}
  - C4: \{Eggs, Apples, Breads\}

- Can we do something predict something?
  - Whenever a person purchases Bread, he/she will also purchase Eggs.
  - How do we do this on a large scale?
An Example

- Customers comes to shop and buys different items in his/her basket.
  - C1: \{Bread, Milk, Eggs\}
  - C2: \{Fish, Bread, Eggs\}
  - C3: \{Carrots, Bread, Eggs\}
  - C4: \{Eggs, Apples, Breads\}

- Can we do something predict something?
  - Whenever a person purchases Bread, he/she will also purchase Eggs.
  - How do we do this on a large scale?
An Example

- Customers come to shop and buy different items in his/her basket.
  - C1: \{Bread, Milk, Eggs\}
  - C2: \{Fish, Bread, Eggs\}
  - C3: \{Carrots, Bread, Eggs\}
  - C4: \{Eggs, Apples, Breads\}

- Can we do something predict something?
  - Whenever a person purchases Bread, he/she will also purchase Eggs.
  - How do we do this on a large scale?
Outline

1. Can Analytics Make a Difference?
   - Netflix in 1997
   - Netflix's Success

2. Business Analytics Overview

3. Data Mining Basics
   - Stages of the Data Mining Process

4. Association Rules and Frequent Item set
   - Example
   - Terminology
   - Frequent Item set Mining

Raghav Kumar Gautam
Business Analytics
Association Rules and Frequent Item sets

- Market-basket problem:
  - Situation: At a retail store we have some large number of items, e.g., bread, milk. Customers fill their market baskets with some subset of the items.
  - The problem is to know what items do people buy together.

Application:

- Marketers use this information to position items, and control the way a typical customer traverses the store.
- Baskets = documents; items = words. Words appearing frequently together in documents may represent phrases or linked concepts. Can be used for intelligence gathering.
- Baskets = sentences, items = documents.
- Two documents with many of the same sentences could represent plagiarism or mirror sites on the Web.
Association Rules and Frequent Item sets

- **Market-basket problem:**
  - Situatio: At a retail store we have some large number of items, e.g., bread, milk. Customers fill their market baskets with some subset of the items.
  - The problem is to know what items do people buy together.

- **Application:**
  - Marketers use this information to position items, and control the way a typical customer traverses the store.
  - Baskets = documents; items = words. Words appearing frequently together in documents may represent phrases or linked concepts. Can be used for intelligence gathering.
  - Baskets = sentences, items = documents.
  - Two documents with many of the same sentences could represent plagiarism or mirror sites on the Web.
Association Rules and Frequent Item sets

- **Market-basket problem:**
  - Situation: At a retail store we have some large number of items, e.g., bread, milk. Customers fill their market baskets with some subset of the items.
  - The problem is to know what items do people buy together.

- **Application:**
  - Marketers use this information to position items, and control the way a typical customer traverses the store.
  - Baskets = documents; items = words. Words appearing frequently together in documents may represent phrases or linked concepts. Can be used for intelligence gathering.
  - Baskets = sentences, items = documents.
  - Two documents with many of the same sentences could represent plagiarism or mirror sites on the Web.
Association Rules and Frequent Item sets

- Market-basket problem:
  - Situation: At a retail store we have some large number of items, e.g., bread, milk. Customers fill their market baskets with some subset of the items.
  - The problem is to know what items do people buy together.

- Application:
  - Marketers use this information to position items, and control the way a typical customer traverses the store.
  - Baskets = documents; items = words. Words appearing frequently together in documents may represent phrases or linked concepts. Can be used for intelligence gathering.
  - Baskets = sentences, items = documents.
  - Two documents with many of the same sentences could represent plagiarism or mirror sites on the Web.
Association Rules and Frequent Item sets

- **Market-basket problem:**
  - **Situation:** At a retail store we have some large number of items, e.g., bread, milk. Customers fill their market baskets with some subset of the items.
  - **The problem is to know what items do people buy together.**

- **Application:**
  - Marketers use this information to position items, and control the way a typical customer traverses the store.
  - Baskets = documents; items = words. Words appearing frequently together in documents may represent phrases or linked concepts. Can be used for intelligence gathering.
  - Baskets = sentences, items = documents.
  - Two documents with many of the same sentences could represent plagiarism or mirror sites on the Web.
Terminology of the Problem

- Association rules are statements of the form $X_1, X_2, \ldots, X_n \Rightarrow Y$, meaning that if we find all of $X_1, X_2, \ldots, X_n$ in the market basket, then we have a good chance of finding $Y$.

- The probability of finding $Y$ for us to accept this rule is called the confidence of the rule. For Eg. confidence of milk; butter $\Rightarrow$ bread is pretty high.

- In many (but not all) situations, we only care about association rules or causalities involving sets of items that appear frequently in baskets.

- Thus mostly we are concerned about the items with high support.
Terminology of the Problem

- Association rules are statements of the form $X_1, X_2, \ldots, X_n \Rightarrow Y$, meaning that if we find all of $X_1, X_2, \ldots, X_n$ in the market basket, then we have a good chance of finding $Y$.

- The probability of finding $Y$ for us to accept this rule is called the confidence of the rule. For Eg. confidence of $\text{milk; butter} \Rightarrow \text{bread}$ is pretty high.

- In many (but not all) situations, we only care about association rules or causalities involving sets of items that appear frequently in baskets.

- Thus mostly we are concerned about the items with high support.
Terminology of the Problem

- Association rules are statements of the form $X_1, X_2, \ldots, X_n \Rightarrow Y$, meaning that if we find all of $X_1, X_2, \ldots, X_n$ in the market basket, then we have a good chance of finding $Y$.

- The probability of finding $Y$ for us to accept this rule is called the confidence of the rule. For Eg. confidence of milk; butter $\Rightarrow$ bread is pretty high.

- In many (but not all) situations, we only care about association rules or causalities involving sets of items that appear frequently in baskets.

- Thus mostly we are concerned about the items with high support.
Outline

1. Can Analytics Make a Difference?
   - Netflix in 1997
   - Netflix’s Success

2. Business Analytics Overview

3. Data Mining Basics
   - Stages of the Data Mining Process

4. Association Rules and Frequent Item set
   - Example
   - Terminology
   - Frequent Item set Mining
Framework for Frequent Item set Mining

- We assume data is too large to be held in the main memory and is present in a database or a flat file. The cost evaluation is done on the basis of
  - Number of passes through the data
- Monotonicity of frequent item sets: If a set of items $S$ is frequent (i.e., appears in at least fraction $s$ of the baskets), then every subset of $S$ is also frequent.
Framework for Frequent Item set Mining

- We assume data is too large to fit in the main memory and is present in a database or a flat file. The cost evaluation is done on the basis of
  - Number of passes through the data
- Monotonicity of frequent item sets: If a set of items $S$ is frequent (i.e., appears in at least fraction $f$ of the baskets), then every subset of $S$ is also frequent.
Frequent Item set Mining Approach

To find frequent item sets, we can:

1. Proceed level wise, finding first the frequent items (sets of size 1), then the frequent pairs, the frequent triples, etc. In our discussion, we concentrate on finding frequent pairs because:
   - Often, pairs are enough.
   - In many data sets, the hardest part is finding the pairs; proceeding to higher levels takes less time than finding frequent pairs. Level wise algorithms use one pass per level.

2. Find all maximal frequent item sets (i.e., sets S such that no proper super set of S is frequent) in one pass or a few passes.
Frequent Item set Mining Approach

To find frequent item sets, we can:

1. Proceed level wise, finding first the frequent items (sets of size 1), then the frequent pairs, the frequent triples, etc. In our discussion, we concentrate on finding frequent pairs because:
   - Often, pairs are enough.
   - In many data sets, the hardest part is finding the pairs; proceeding to higher levels takes less time than finding frequent pairs. Level wise algorithms use one pass per level.

2. Find all maximal frequent item sets (i.e., sets $S$ such that no proper super set of $S$ is frequent) in one pass or a few passes.
Frequent Item set Mining Approach

To find frequent item sets, we can:

1. Proceed level wise, finding first the frequent items (sets of size 1), then the frequent pairs, the frequent triples, etc. In our discussion, we concentrate on finding frequent pairs because:
   - Often, pairs are enough.
   - In many data sets, the hardest part is finding the pairs; proceeding to higher levels takes less time than finding frequent pairs. Level wise algorithms use one pass per level.

2. Find all maximal frequent item sets (i.e., sets S such that no proper super set of S is frequent) in one pass or a few passes.
Frequent Item set Mining Approach

To find frequent item sets, we can:

1. Proceed level wise, finding first the frequent items (sets of size 1), then the frequent pairs, the frequent triples, etc. In our discussion, we concentrate on finding frequent pairs because:
   - Often, pairs are enough.
   - In many data sets, the hardest part is finding the pairs; proceeding to higher levels takes less time than finding frequent pairs. Level wise algorithms use one pass per level.

2. Find all maximal frequent item sets (i.e., sets $S$ such that no proper super set of $S$ is frequent) in one pass or a few passes.
Frequent Item set Mining Approach

To find frequent item sets, we can:

1. Proceed level wise, finding first the frequent items (sets of size 1), then the frequent pairs, the frequent triples, etc. In our discussion, we concentrate on finding frequent pairs because:
   - Often, pairs are enough.
   - In many data sets, the hardest part is finding the pairs; proceeding to higher levels takes less time than finding frequent pairs. Level wise algorithms use one pass per level.

2. Find all maximal frequent item sets (i.e., sets $S$ such that no proper super set of $S$ is frequent) in one pass or a few passes.
Given support threshold $s$, in the first pass we find the items that appear in at least fraction $s$ of the baskets. This set is called $L_1$, the frequent items.

1. \[
\text{SELECT } * \text{ FROM Baskets GROUP by item HAVING COUNT(*) } \geq s;
\]

2. Pairs of items in $L_1$ become the candidate pairs $C_2$ for the second pass. We hope that the size of $C_2$ is not so large that there is not room for an integer count per candidate pair. The pairs in $C_2$ whose count reaches $s$ are the frequent pairs, $L_2$.

1. \[
\text{SELECT b1.item, b2.item, COUNT(*) FROM Baskets b1, Baskets b2 WHERE b1.BID = b2.BID AND b1.item < b2.item GROUP BY b1.item, b2.item HAVING COUNT(*) } \geq s;
\]
The Algorithm -1

1. Given support threshold \( s \), in the first pass we find the items that appear in at least fraction \( s \) of the baskets. This set is called \( L_1 \), the frequent items.
   
   - SELECT * FROM Baskets GROUP by item HAVING COUNT(*) \( \geq s \);

2. Pairs of items in \( L_1 \) become the candidate pairs \( C_2 \) for the second pass. We hope that the size of \( C_2 \) is not so large that there is not room for an integer count per candidate pair. The pairs in \( C_2 \) whose count reaches \( s \) are the frequent pairs, \( L_2 \).
   
   - SELECT b1.item, b2.item, COUNT(*) FROM Baskets b1, Baskets b2 WHERE b1.BID = b2.BID AND b1.item < b2.item GROUP BY b1.item, b2.item HAVING COUNT(*) \( \geq s \);
Given support threshold $s$, in the first pass we find the items that appear in at least fraction $s$ of the baskets. This set is called $L_1$, the frequent items.

1. SELECT * FROM Baskets GROUP by item HAVING COUNT(*) $\geq s$;

Pairs of items in $L_1$ become the candidate pairs $C_2$ for the second pass. We hope that the size of $C_2$ is not so large that there is not room for an integer count per candidate pair. The pairs in $C_2$ whose count reaches $s$ are the frequent pairs, $L_2$.

2. SELECT b1.item, b2.item, COUNT(*) FROM Baskets b1, Baskets b2 WHERE b1.BID = b2.BID AND b1.item < b2.item GROUP BY b1.item, b2.item HAVING COUNT(*) $\geq s$;
The Algorithm -2

1. The candidate triples, $C_3$ are those sets $\{A;B;C\}$ such that all of $\{A;B\}, \{A;C\}$, and $\{B;C\}$ are in $L_2$. On the third pass, count the occurrences of triples in $C_3$; those with a count of at least $s$ are the frequent triples, $L_3$.

2. Proceed as far as you like (or the sets become empty). $L_i$ is the frequent sets of size $i$; $C_{i+1}$ is the set of sets of size $i+1$ such that each subset of size $i$ is in $L_i$. 

Raghav Kumar Gautam
The Algorithm -2

1. The candidate triples, $C_3$ are those sets $\{A;B;C\}$ such that all of $\{A;B\}, \{A;C\}$, and $\{B;C\}$ are in $L_2$. On the third pass, count the occurrences of triples in $C_3$; those with a count of at least $s$ are the frequent triples, $L_3$.

2. Proceed as far as you like (or the sets become empty). $L_i$ is the frequent sets of size $i$; $C_{i+1}$ is the set of sets of size $i+1$ such that each subset of size $i$ is in $L_i$. 
Summary

- Importance of Analytics
- Overview of Data Mining
- Studied a Market Basket Analysis Problem
- Looked at a technique to solve the problem
Reference

Thomas H. Davenport, Jeanne G. Harris
*Competing on analytics: the new science of winning*

Rakesh Agrawal and Ramakrishnan Srikant.
Fast algorithms for mining association rules in large databases.

http://www.oracle.com/us/corporate/analystreports/industries/05960