

---

# Mining Knowledge from Text Collections Using Automatically Generated Metadata

John M. Pierre  
jpierre@interwoven.com  
Interwoven, Inc.  
San Francisco, CA, USA

# Outline

---

## **Project overview**

## **Methodology**

- Faceted Metadata
- Automated text categorization
- Data mining

## **Example**

- Text collection
- Metadata schema
- Results

## **Conclusions**

# Project Overview

---

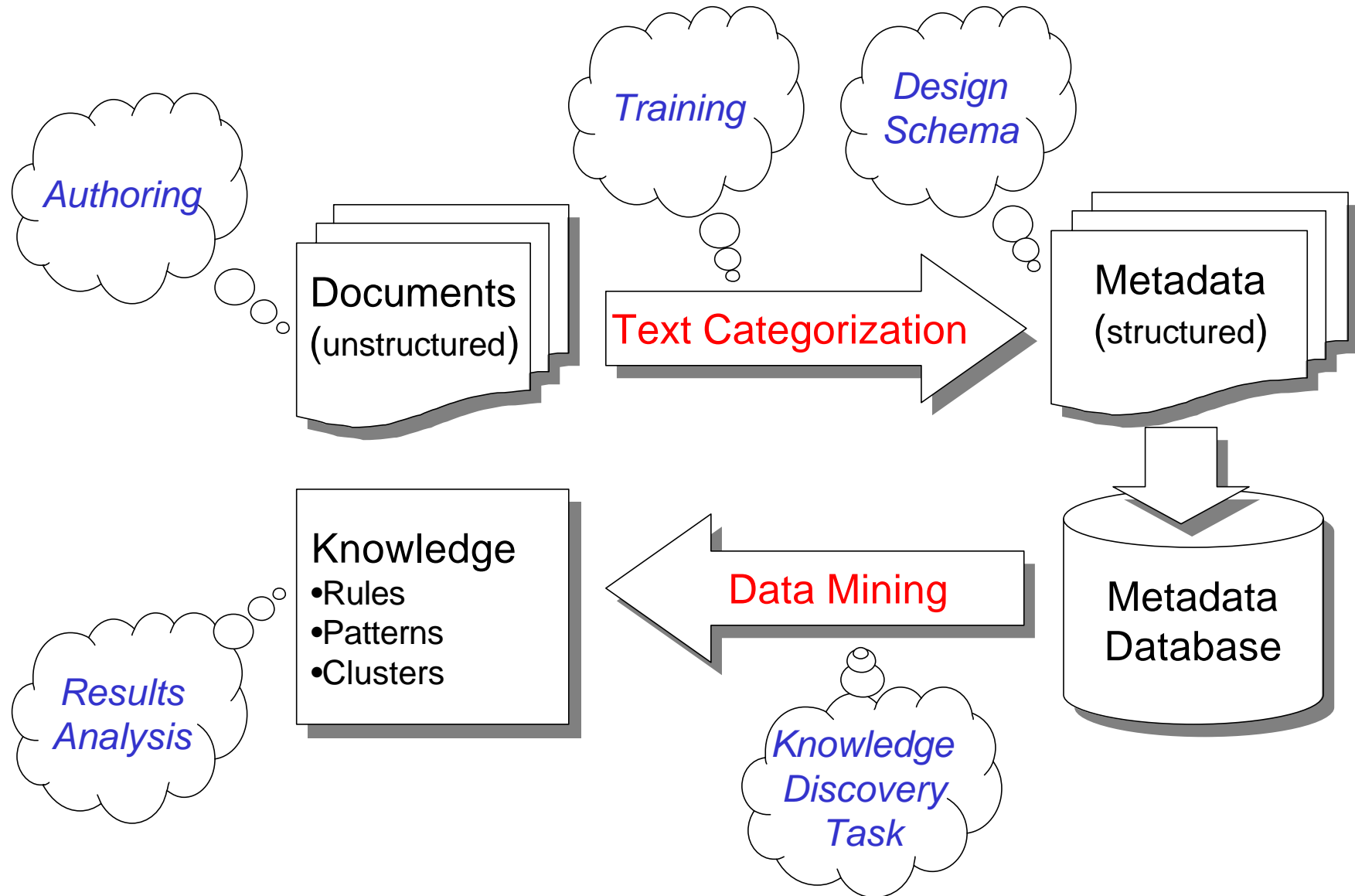
## **Problem Statement:**

- Text collections store vast amounts of collective knowledge in businesses and organizations
- File systems, email, document management, portals, etc.
- Knowledge discovery and re-use is difficult

## **Solution:**

- Automatically generate a database of faceted metadata
- Apply data mining to discover new knowledge

# Project Overview



# Methodology

---

## **Document Selection**

- Covers the domain of interest
- Statistical sample

## **Document Segmentation**

- Appropriate set of transactions

## **Metadata Schema**

- Concepts (vocabulary problem)
- Facets

## **Text Categorization**

- Automatically generate a metadata database
- Requires training examples

## **Data Mining**

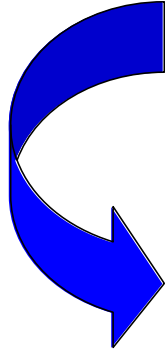
- Discover knowledge from metadata database
- Associations between concepts, clustering, visualization, etc.

# Example: Product Reviews

---

## Documents:

- 47,923 product reviews
- 24147 for training
- 23776 for mining



The DIA 100MK2 is better than anything close to its price. Built like a tank, with plenty of very clean, power that comes out of a totally black, noise free background, this design really proves that simpler is better! For those with ears and good associated equipment, a rare deal, in fact a steal.

## Metadata:

- 4 facets

```
<metadata>
  <category>Amplification</category>
  <subcategory>Amplifiers</subcategory>
  <products>Acurus DIA/100</products>
  <rating>GOOD</rating>
</metadata>
```

# Example: Text Categorization

---

<b>Metadata Facet</b>	<b>Number of Concepts</b>	<b>Classifier</b>	<b>Precision</b>	<b>Recall</b>
Category	11	Naïve Bayes	0.79	0.79
Subcategory	49	Naïve Bayes	0.54	0.54
Products	1610	Boolean	0.31	0.18
Rating	2	Naïve Bayes	0.91	0.91

Precision = # correctly assigned / # assigned

Recall = # correctly assigned / # total correct

# Example: Data Mining

## Association Rules:

- $Support(A) = (|A| / |T|) = P(A)$
- $Confidence(A \rightarrow B) = Support(A,B) / Support(A) = P(B|A)$
- Support thresholds 0.1%, 0.05%, 0.01%
- Confidence threshold 60%
- Used **Apriori** algorithm

Rule Type	Examples
$Subcategory(A) \rightarrow Category(B)$	A / V Receivers ? Amplification DVD Players ? Home Video Main Speaker ? Speakers
$Product(A) \rightarrow Category(B)$	Yamaha RX / V795 ? Amplification Samsung DVD/611 ? Home Video Paradigm Atom ? Speakers
$Product(A) \rightarrow Subcategory(B)$	Yamaha RX / V795 ? A / V Receivers Samsung DVD/611 ? DVD Players Paradigm Atom ? Main Speaker
$Product(A) \rightarrow Rating(B)$	Yamaha RX / V795 ? GOOD Samsung DVD/611 ? BAD Paradigm Atom ? GOOD

# Example: Knowledge Discovery Results

Precision = # correct rules mined / # total rules mined

Recall = # correct rules mined / # correct rules expected

Rule Type	Support	# mined rules	Precision	Recall
Subcategory(A) ? Category(B)	0.1%	16	1.0	0.33
	0.05%	20	1.0	0.41
	0.01%	26	1.0	0.53
Product(A) ? Category(B)	0.1%	168	0.88	0.10
	0.05%	341	0.84	0.18
	0.01%	475	0.74	0.29
Product(A) ? Subcategory(B)	0.1%	86	0.74	0.05
	0.05%	210	0.60	0.08
	0.01%	443	0.39	0.11
Product(A) ? Rating(B)	0.1%	259	0.91	0.17
	0.05%	474	0.92	0.28
	0.01%	881	0.89	0.5

# Conclusions

---

## Summary

- Possible to mine knowledge from text collections using faceted metadata
- Automated methods can be used to efficiently assign metadata
- Presented an example with good results

## Applications

- Marketing research
- Discover trends and patterns
- Analyze internet resources (blogs, news groups, email)