Content-based Recommender Systems
problems, challenges and research directions

Giovanni Semeraro & the SWAP group
http://www.di.uniba.it/~swap/
semeraro@di.uniba.it

Department of Computer Science
University of Bari “Aldo Moro”

UMAP 2010 – 8° Workshop on
INTELLIGENT TECHNIQUES FOR WEB PERSONALIZATION
& RECOMMENDER SYSTEMS (ITWP 2010)
BIG ISLAND OF HAWAII, JUNE 20 2010
Outline

1. Content-based Recommender Systems (CBRS)
   - Basics
   - Advantages & Drawbacks

2. Drawback 1: Limited content analysis
   - Beyond keywords: Semantics into CBRS
   - Taking advantage of Web 2.0: Folksonomy-based CBRS

3. Drawback 2: Overspecialization
   - Strategies for diversification of recommendations
Content-based Recommender Systems (CBRS)

1. **Recommend an item to a user based upon a description of the item and a profile of the user’s interests**

2. Implement strategies for:
   - representing items
   - creating a user profile that describes the types of items the user likes/dislikes
   - comparing the user profile to some reference characteristics (with the aim to predict whether the user is interested in an unseen item)

Content-based Filtering

Information Source

User profile compared against items for *relevance* computation

Items recommended to the user

Target User

User Profile
Content-based Filtering

1. Each user is assumed to operate independently

2. Items are represented by some features
   - Movies: actors, director, plot, ...

3. The profile is often created and updated automatically in response to feedback on the desirability of items that have been presented to the user
   - Machine Learning for automated inference
   - Relevance judgment on items, e.g. ratings
   - Training on rated items → user profile

4. Filtering based on the comparison between the content (features) of the items and the user preferences as defined in the user profile
   - Keyword-based representation for content and profiles → string matching or text similarity
General Architecture of CBRS

represented items

structured item representation

content analyzer

new items

information source

user u<sub>a</sub> training examples

user u<sub>a</sub> profile

profiles

active user u<sub>a</sub>

filtering component

user u<sub>a</sub> feedback

feedback

list of recommendations

profile learner

user u<sub>a</sub> profile

user u<sub>a</sub> feedback

item descriptions

active user u<sub>a</sub>
Advantages of CBRS

**1. USER INDEPENDENCE**
- CBRS exploit solely ratings provided by the active user to build her own profile
- No need for data on other users

**2. TRANSPARENCY**
- CBRS can provide explanations for recommended items by listing content-features that caused an item to be recommended

**3. NEW ITEM** (Item not yet rated by any user)
- CBRS are capable of recommending new and unknown items
- No first-rater problem
Drawbacks of CBRS: LIMITED CONTENT ANALYSIS

1. No suitable suggestions if the analyzed content does not contain enough information to discriminate items the user likes from items the user does not like.

2. Content must be encoded as meaningful features:
   - automatic/manually assignment of features to items might be insufficient to define distinguishing aspects of items necessary for the elicitation of user interests.
   - keywords not appropriate for representing content, due to polysemy, synonymy, multi-word concepts (*homography, homophony,...*) – “Sator arepo eccetera” [Eco07]
Keyword-based Profiles

AI is a branch of computer science

the 2011 International Joint Conference on Artificial Intelligence will be held in Spain

apple launches a new product...

USER PROFILE
artificial 0.02
intelligence 0.01
apple 0.13
AI 0.15
...

MULTI-WORD CONCEPTS
Keyword-based Profiles

**doc1**

AI is a branch of computer science

**doc2**

the 2011 International Joint Conference on Artificial Intelligence will be held in Spain

**doc3**

apple launches a new product...

---

**USER PROFILE**

<table>
<thead>
<tr>
<th>term</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>artificial</td>
<td>0.02</td>
</tr>
<tr>
<td>intelligence</td>
<td>0.01</td>
</tr>
<tr>
<td>apple</td>
<td>0.13</td>
</tr>
<tr>
<td>AI</td>
<td>0.15</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

---

SYNONYMY
AI is a branch of computer science.

The 2011 International Joint Conference on Artificial Intelligence will be held in Spain.

Apple launches a new product...

Keyword-based Profiles

<table>
<thead>
<tr>
<th>USER PROFILE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>artificial</td>
<td>0.02</td>
</tr>
<tr>
<td>intelligence</td>
<td>0.01</td>
</tr>
<tr>
<td>apple</td>
<td>0.13</td>
</tr>
<tr>
<td>AI</td>
<td>0.15</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

NLP methods are needed for the elicitation of user interests.
Drawbacks of CBRS: OVERSPECIALIZATION

1. CBRS suggest items whose scores are high when matched against the user profile
   - the user is going to be recommended items similar to those already rated

2. No inherent method for finding something unexpected

3. Obviousness in recommendations
   - suggesting “STAR TREK” to a science-fiction fan: accurate but not useful
   - users don’t want algorithms that produce better ratings, but sensible recommendations

4. The Serendipity Problem

The serendipity problem: mind cages

- Homophily: the tendency to surround ourselves by like-minded people
  - opinions taken to extremes
  - cultural impoverishment
  - biodiversity?
The homophily trap

- Does homophily hurt RS?
  - try to tell Amazon that you liked the movie “War Games”…

www.ethanzuckerman.com/blog/2008/04/25/homophily-serendipity-xenophilia/
The homophily trap

Customers Who Bought This Item Also Bought

- *WarGames: The Dead Code* DVD by Colm Feore
  - Rating: ★★★☆☆ (22)
  - Price: $13.49

- *Hackers* DVD by Jonny Lee Miller
  - Rating: ★★★☆☆ (313)
  - Price: $6.99

- *The Last Starfighter 25th Anniversary Edition* DVD by Lance Guest
  - Rating: ★★★☆☆ (179)
  - Price: $11.49

Looking for "wargames" Products?

Other customers suggested these items:

- *Watch Your Back* by 1 customer
  - Price: $24.82

- *Field of Glory: Ancient and Medieval Wargaming Rules* by Richard Bodley-Scott
  - Price: $23.07

- *CIT Introductory Box Set (Classic Battletaci)* by Catalyst Game Labs
  - Price: $19.95

- *Legions Triumphant: Field of Glory Imperial Rome Army List* by Richard Bodley-Scott
  - Price: $29.95

- *Going to War: Create War Games* by Jaso
  - Price: $26.22

Recommendations by other (ageing?) COMPUTER GEEKS!
“Item-to-Item” homophily...
Harry Potter for ever?
Novelty vs Serendipity

- **Novelty**: A novel recommendation helps the user find a surprisingly interesting item she might have autonomously discovered.

- **Serendipity**: A serendipitous recommendation helps the user find a surprisingly interesting item she might not have otherwise discovered.

- How to introduce serendipity in (CB)RS?

“Computational” serendipity? A motivating example

for Star Trek fans: Did you try “Star Trek – The experience” in Las Vegas?
<table>
<thead>
<tr>
<th>PROBLEMS</th>
<th>CHALLENGES</th>
<th>RESEARCH DIRECTIONS</th>
</tr>
</thead>
</table>
| Limited Content Analysis| Beyond keywords: novel strategies for the representation of items and profiles | ▪ Semantic analysis of content by means of external knowledge sources  
▪ Language-independent CBRS |
|                         | Taking advantage of Web 2.0 for collecting User Generated Content | Folksonomy-based CBRS                                        |
| Overspecialization      | Defeating homophily: recommendation diversification           | ▪ “computational” serendipity → programming for serendipity  
▪ Knowledge Infusion      |
<table>
<thead>
<tr>
<th>PROBLEMS</th>
<th>CHALLENGES</th>
<th>RESEARCH DIRECTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited Content Analysis</td>
<td>Beyond keywords: novel strategies for the representation of items and profiles</td>
<td>▪ <strong>Semantic analysis of content by means of external knowledge sources</strong>&lt;br&gt;▪ <strong>Language-independent CBRS</strong></td>
</tr>
<tr>
<td></td>
<td>Taking advantage of Web 2.0 for collecting User Generated Content</td>
<td>Folksonomy-based CBRS</td>
</tr>
<tr>
<td>Overspecialization</td>
<td>Defeating homophily: recommendation diversification</td>
<td>▪ “computational” serendipity → programming for serendipity&lt;br&gt;▪ <strong>Knowledge Infusion</strong></td>
</tr>
</tbody>
</table>
Semantic Analysis: beyond keywords

Semantic Analysis =

1. Semantics: concept identification in text-based representations through advanced NLP techniques → “beyond keywords”

2. Personalization: representation of user information needs in an effective way → “deep (high-accuracy) user profiles”
Beyond keywords: Word Sense Disambiguation (WSD) - from words to meanings

- WSD selects the proper meaning (sense) for a word in a text by taking into account the context in which that word occurs.

Dictionaries, Ontologies, e.g. WordNet

AI is a branch of computer science

the 2011 International Joint Conference on Artificial Intelligence will be held in Spain

apple launches a new product...

USER PROFILE

<table>
<thead>
<tr>
<th>Concept</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>#12387</td>
<td>0.03</td>
</tr>
<tr>
<td>apple</td>
<td>0.13</td>
</tr>
<tr>
<td>AI</td>
<td>0.15</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

MULTI-WORD CONCEPTS
AI is a branch of computer science.

The 2011 International Joint Conference on Artificial Intelligence will be held in Spain.

Apple launches a new product...

**USER PROFILE**

<table>
<thead>
<tr>
<th>Term</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>#12387</td>
<td>0.18</td>
</tr>
<tr>
<td>apple</td>
<td>0.13</td>
</tr>
<tr>
<td>#12387</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**SYNONYMY**
AI is a branch of computer science

the 2011 International Joint Conference on Artificial Intelligence will be held in Spain

apple launches a new product...

---

SEMANTIC USER PROFILE
sense identifiers rather than keywords

USER PROFILE

#12387 0.18

#12567 0.13

...


Advantages of Sense-based Representations

1. Semantic matching between items and profiles
   - computing semantic relatedness [Pedersen04] rather than string matching (e.g., by using similarity measures between WordNet synsets)

2. Senses are inherently multilingual
   - Concepts remain the same across different languages, while terms used for describing them in each specific language change

3. Improving transparency
   - matched concepts can be used to justify suggestions

4. Collaborative Filtering could benefit too
   - finding better neighbors: similar users discovered by looking at profile overlap even if they did not rate the same items
   - semantic profiles succeed where Pearson’s correlation coefficient fail

Sense-based profiles in a hybrid CB-CF recommender

- Sense-based profiles obtained by applying WSD on textual description of items
  - WordNet as sense repository
  - Synset-based user profiles

- Hybrid CB-CF RS

Clustering of sense-based profiles

User profiles

Clusters of profiles

Profiles in the cluster used as neighbors

Active user
Experimental Evaluation on EachMovie dataset

- 835 users selected from EachMovie dataset*
  - 1,613 movies grouped into 10 categories,
    180,356 ratings, user-item matrix 87% sparse
  - Each user rated between 30 and 100 movies
  - Discrete ratings between 0 and 5
  - Movie content crawled from the Internet Movie Database (IMDb)

- CF algorithm using Pearson’s correlation coefficient vs. CF algorithm integrating clusters of semantic user profiles

*2,811,983 ratings entered by 72,916 users for 1628 different movies. As of October, 2004, HP/Compaq Research (formerly DEC Research) retired the EachMovie dataset. It is no longer available for download
Sense-based profiles improve recommendations

Rating scale: 0-5

Mean Absolute Error (MAE)

Max Neighborhood Size

Pearson

Hybrid
## Semantic Analysis: Ontologies in CBRS

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>DESCRIPTION</th>
</tr>
</thead>
</table>
| SEWeP (Semantic Enhancement for Web Personalization) [Eirinaki03] | Manually built domain-specific taxonomy of categories for the automated annotation of Web pages  
WordNet-based word similarity used to map keywords to categories  
Categories of interest discovered from navigational history of the user |
| Quickstep & Foxtrot [Middleton04]            | Recommendation of on-line academic research papers  
Research paper topic ontology based on the computer science classification of the DMOZ open directory project  
K-NN classification used to associate classes to previously browsed papers |

## Semantic Analysis: Ontologies in CBRS

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>DESCRIPTION</th>
</tr>
</thead>
</table>
| Informed Recommender [Aciar07]            | Consumer product reviews to make recommendations  
                                         | Ontology used to convert consumers’ opinions into a structured form  
                                         | Text-mining for mapping sentences in the reviews into the ontology information structure  
                                         | Search-based recommendations                                                                                                                   |
| RS for Interactive Digital Television     | OWL ontology for representing TV programs and user profiles  
                                         | OWL representation allows reasoning on preferences and discovering new knowledge  
                                         | Spreading activation for matching items and preferences                                                                                     |
| [Blanco-Fernandez08]                     |                                                                                                                                                                                                             |
| News@hand [Cantador08]                    | Ontology-based news recommender  
                                         | 17 ontologies adapted from the IPTC ontology (http://nets.ii.uam.es/neptuno/iptc/)  
                                         | Items and user profiles represented as vectors in the space of concepts defined by the ontologies                                          |

- Do we really need only ontologies?
  - What about encyclopedic knowledge sources available on the Web?
- Is Wikipedia potentially useful for CBRS? How?
  - It is free
  - It covers many domains
  - It is under constant development by the community
  - It can be seen as a multilingual corpus
  - Its accuracy rivals that of Encyclopaedia Britannica [Giles05]

Explicit Semantic Analysis (ESA)

Technique able to provide a fine-grained semantic representation of natural language texts in a high-dimensional space of comprehensible concepts derived from Wikipedia [Gabri06]

Wikipedia viewed as an ontology = a collection of \( \sim 1M \) concepts


Wikipedia is viewed as an **ontology** - a collection of ~1M concepts

Every Wikipedia article represents a **concept**

---

**Panthera**

*From Wikipedia, the free encyclopedia*

Panthera is a genus of the family Felidae that contains four well-known living species: the lion, tiger, jaguar, and leopard. The genus comprises about half of the big cats. One meaning of the word panther is to designate cats of this family. Only these four cat species have the anatomical changes enabling them to roar. The primary reason for this was assumed to be the incomplete ossification of the hyoid bone. However, new studies show that the ability to roar is due to other morphological features, especially of the larynx. The snow leopard, *Uncia uncia*, which is sometimes included within Panthera, does not roar. Although it has an incomplete ossification of the hyoid bone, it lacks the special morphology of the larynx, which is typical for lions, tigers, jaguars and leopards.

**Species and subspecies**

---

Article **words** are associated with the **concept** (TF-IDF)
Wikipedia is viewed as an **ontology** - a collection of ~**1M** concepts

Every Wikipedia article represents a **concept**

Article words are associated with the **concept** (TF-IDF)
Wikipedia is viewed as an **ontology** - a collection of ~**1M** concepts.

Every Wikipedia article represents a **concept**

Article words are associated with the **concept** (TF-IDF)

The **semantics** of a word is the **vector** of its **associations** with Wikipedia concepts.
Explicit Semantic Analysis (ESA)

The **semantics of a text fragment** is the average vector (centroid) of the semantics of its words.

In practice – WSD...
$D_1 = 2C_1 + 3C_2 + 5C_3$

$D_2 = 3C_1 + 7C_2 + 1C_3$

$C_i = \text{Wikipedia article}$

Wikipedia and CBRS: recent ideas

- Wikipedia used for computing the similarity between movie descriptions for the Netflix prize competition [Lees08]
- ESA used for user profiling, spam detection and RSS filtering [Smirnov08]
- Wikipedia included in a Knowledge Infusion process for recommendation diversification [Semeraro09a]


<table>
<thead>
<tr>
<th>PROBLEMS</th>
<th>CHALLENGES</th>
<th>RESEARCH DIRECTIONS</th>
</tr>
</thead>
</table>
| Limited Content Analysis | Beyond keywords: novel strategies for the representation of items and profiles | ▪ Semantic analysis of content by means of external knowledge sources  
▪ **Language-independent CBRS** |
|                          | Taking advantage of Web 2.0 for collecting User Generated Content | Folksonomy-based CBRS                                    |
| Overspecialization       | Defeating homophily: recommendation diversification | ▪ “computational” serendipity → programming for serendipity  
▪ Knowledge Infusion      |
MARS (MultilAnguage Recommender System) cross-language user profiles

- WSD for building language-independent user profiles
- MultiWordNet as sense repository
  - Multilingual lexical database that supports English, Italian, Spanish, Portuguese, Hebrew, Romanian, Latin
  - Alignment between synsets in the different languages
    - Semantic relations imported and preserved

<table>
<thead>
<tr>
<th>Language</th>
<th>Synset</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td><em>world, human race, humanity, humankind, human beings, humans, mankind, man</em></td>
<td><em>all of the inhabitants of the earth</em></td>
</tr>
<tr>
<td>Italian</td>
<td><em>mondo, umanità, uomo, genere umano, terra</em></td>
<td><em>insieme degli abitanti della terra, il complesso di tutti gli esseri umani</em></td>
</tr>
</tbody>
</table>
MARS (MultilAnguage Recommender System)
cross-language user profiles

**ENGLISH description**
CLOCKWORK ORANGE
Being the adventures of a young man whose principal interests are rape, ultra-violence and Beethoven

**ITALIAN description**
ARANCIA MECCANICA
Le avventure di un giovane i cui principali interessi sono lo stupro, l’ultra-violenza e Beethoven

**Bag of Synset**
“a12889641” “n5477412”
“n3652872” “a2584413”
“n3255687” “a3225896”
“n32256325” “n225784”
“n255632” “Beethoven”

**WordNet**
"n5477412" “a1744532”
“a2584413” “n3652872”
“a3225722” “n32256325”
“n225784” “n255632”
“Beethoven”
MARS (Multilingual Recommender System)

cross-language user profiles

Target User

Bag of Synset from ITALIAN

"n5477412" "a1744532" "a2584413" "n3652072"
"a3225722" "n32256325" "n225784" "n255631"
"Beethoven" ...

User PROFILE

"n5477412" "a1744532" "a2584413" "n3652072"
"a3225722" "n32256325" "n225784" "n255631"
"Beethoven" "a5547652" "n632258" "n11052255"
"n777412" "a95525" ...

ENGLISH description

A rape victim, enraged at the light sentence her attackers received on account that she was of "questionable character" goads a female prosecutor to charge the men who literally cheered the attack on.

Bag of Synset from ENGLISH

"n5477412" "a34225" "n63325" "n52223665"
"a2584413" "n3652072" "n32256325" "n225784"
"n255632" ...

SUGGESTED
MARS (MultilAnguage Recommender System)
preliminary results

- MovieLens 100k ratings dataset
- 613 users with ≥ 20 ratings selected from 943 different users
  - 520 movies and 40,717 ratings
  - movie content crawled from Wikipedia (English and Italian)
  - same movie - different descriptions in English and Italian

- Results in terms of $F_{ß=0.5}$ measure
  - no statistically significant difference wrt the baselines

- Neither content translations nor profile translations achieve the same effectiveness (they cannot avoid the negative impact of polysemy and lack of context)

<table>
<thead>
<tr>
<th>Recommendations</th>
<th>Profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="https://example.com/flag.png" alt="UK" /></td>
<td><img src="https://example.com/flag.png" alt="UK" /></td>
</tr>
<tr>
<td><img src="https://example.com/flag.png" alt="UK" /></td>
<td><img src="https://example.com/flag.png" alt="UK" /></td>
</tr>
<tr>
<td><img src="https://example.com/flag.png" alt="Italian" /></td>
<td><img src="https://example.com/flag.png" alt="Italian" /></td>
</tr>
<tr>
<td><img src="https://example.com/flag.png" alt="Italian" /></td>
<td><img src="https://example.com/flag.png" alt="Italian" /></td>
</tr>
<tr>
<td>64.91</td>
<td>63.98</td>
</tr>
<tr>
<td>63.70</td>
<td>63.71</td>
</tr>
</tbody>
</table>
# Putting Intelligence into CBRS: Challenges & Research Directions

<table>
<thead>
<tr>
<th>PROBLEMS</th>
<th>CHALLENGES</th>
<th>RESEARCH DIRECTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited Content Analysis</td>
<td>Beyond keywords: novel strategies for the</td>
<td>▪ Semantic analysis of content by means of external knowledge sources</td>
</tr>
<tr>
<td></td>
<td>representation of items and profiles</td>
<td>▪ Language-independent CBRS</td>
</tr>
<tr>
<td></td>
<td><strong>Taking advantage of Web 2.0 for collecting</strong></td>
<td><strong>Folksonomy-based CBRS</strong></td>
</tr>
<tr>
<td></td>
<td><strong>User Generated Content</strong></td>
<td></td>
</tr>
<tr>
<td>Overspecialization</td>
<td>Defeating homophily: recommendation diversification</td>
<td>▪ “computational” serendipity → programming for serendipity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>▪ Knowledge Infusion</td>
</tr>
</tbody>
</table>
Web 2.0 & User-Generated Content (UGC)
Social Tagging & Folksonomies

- Users annotate resources of interests with free keywords, called **tags**
- Social tagging activity builds a bottom-up classification schema, called a **folksonomy**
  - Folksonomy: “Folks” + “Taxonomy”
- How to exploit folksonomies for advanced user profiling in CBRS?
In the context of cultural heritage personalization, does the integration of UGC and textual description of artwork collections cause an increase of the prediction accuracy in the process of recommending artifacts to users?
FIRSt: **Folksonomy-based Item Recommender System**

- Artwork representation
  - Artist
  - Title
  - Description
  - Tags

- Semantic Indexing
  - Change of text representation from vectors of words (BOW) into vectors of WordNet synsets (BOS)
  - From tags to semantic tags

- Supervised Learning
  - Bayesian Classifier learned from artworks labeled with user ratings and tags
27) Caravaggio - Deposition from the Cross

Descrizione dell'opera

The Deposition, considered one of Caravaggio’s greatest masterpieces, was commissioned by Girolamo Vittrice for his family chapel in S. Maria in Vallicella (Chiesa Nuova) in Rome. In 1797 it was included in the group of works transferred to Paris in execution of the Treaty of Tolentino. After its return in 1817 it became part of Pius VII’s Pinacoteca. Caravaggio did not really portray the Burial or the Deposition in the traditional way, inasmuch as Christ is not shown at the moment when he is laid in the tomb, but rather when, in the presence of the holy women, he is laid by Nicodemus and John on the Anointing Stone, that is, stone with which the sepulchre will be closed. Around the body of Christ are the Virgin, Mary Magdalene, John, Nicodemus and Mary of Cleophas, who raises her arms and eyes to heaven in a gesture of high dramatic tension. Caravaggio, who arrived in Rome towards 1592-93, was the protagonist of a real artistic revolution as regards the way of treating subjects and the use of colour and light, and was certainly the most important personage of the "realist" trend of seventeenth century painting.

Social Tags (from other users): caravaggio, deposition, christ, cross, suffering, religion

5-point rating scale

Personal Tags: passion
The Deposition, considered one of Caravaggio’s greatest masterpieces, was commissioned by Girolamo Vittoria for his family chapel in S. Maria in Vallicella (Chiesa Nuova) in Rome. In 1797 it was included in a group of works transferred to Paris by the execution of the Treaty of Tolentino. After its return in 1817 it became part of Pius VII’s Pinacoteca. Caravaggio did not try to portray the Burial or the Deposition in the traditional way, inasmuch as Christ is not shown at the moment when he is laid in the tomb, but rather when, in the presence of the holy women, he is laid by Nicodemus and John on the Anointing Stone, that is stone with which the sepulchre will be closed. Around the body of Christ are the Virgin, Mary Magdalene, John, Nicodemus and Mary of Cleophas, who raises her arms and eyes to heaven in a gesture of high dramatic tension. Caravaggio, who arrived in Rome towards 1592-93, was the protagonist of a real artistic revolution as regards the way of treating subjects and the use of colour and light, and was certainly the most important protagonist of the “realist” trend of seventeenth-century painting.
Experimental Evaluation

Goal: Compare predictive accuracy of FIRSt when user profiles are learned from:

- Static content only, i.e., textual descriptions of artifacts (content-based profiles)

- both Static and Dynamic UGC (tag-based profiles). UGC can be:
  - Personal Tags, entered by a user for an artifact, i.e., the user’s contribution to the whole folksonomy
  - Social Tags, i.e., the whole folksonomy of tags added by all visitors
Experimental Setup

Dataset
1. 45 paintings from the Vatican picture-gallery
2. Static content (i.e., title, artist and description) captured using screenscraping bots

Subjects
1. 30 volunteers
2. average age ≈ 25
3. none reported to be an art expert
Experimental Design

- 5 experiments designed
  - EXP#1: Static Content
  - EXP#2: Personal Tags
  - EXP#3: Social Tags
  - EXP#4: Static Content + Personal Tags
  - EXP#5: Static Content + Social Tags

- 5-fold cross validation

- One run for each user:
  1. Select the appropriate content depending on the experiment
  2. Split the selected data into a training set $Tr$ and a test set $Ts$
  3. Use $Tr$ for learning the corresponding user profile
  4. Evaluate the predictive accuracy of the induced profile on $Ts$

- Evaluation Metrics: Precision ($Pr$), Recall ($Re$), F1 measure
## Analysis of Precision

<table>
<thead>
<tr>
<th>Type of Content</th>
<th>Precision*</th>
<th>Recall*</th>
<th>F1*</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP#1: Static Content</td>
<td>75.86</td>
<td>94.27</td>
<td>84.07</td>
</tr>
<tr>
<td>EXP#2: Personal Tags</td>
<td>75.96</td>
<td>92.65</td>
<td>83.48</td>
</tr>
<tr>
<td>EXP#3: Social Tags</td>
<td>75.59</td>
<td>90.50</td>
<td>82.37</td>
</tr>
<tr>
<td>EXP#4: Static Content + Personal Tags</td>
<td>78.04</td>
<td>93.60</td>
<td>85.11</td>
</tr>
<tr>
<td>EXP#5: Static Content + Social Tags</td>
<td>78.01</td>
<td>93.19</td>
<td>84.93</td>
</tr>
</tbody>
</table>

* Results averaged over the 30 study subjects
## Analysis of Precision

<table>
<thead>
<tr>
<th>Type of Content</th>
<th>Precision*</th>
<th>Recall*</th>
<th>F1*</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP#1: Static Content</td>
<td>75.86</td>
<td>94.27</td>
<td>84.07</td>
</tr>
<tr>
<td>EXP#2: Personal Tags</td>
<td>75.96</td>
<td>92.65</td>
<td>83.48</td>
</tr>
<tr>
<td>EXP#3: Social Tags</td>
<td>75.59</td>
<td>90.50</td>
<td>82.37</td>
</tr>
<tr>
<td>EXP#4: Static Content + Personal Tags</td>
<td>78.04</td>
<td>93.60</td>
<td>85.11</td>
</tr>
<tr>
<td>EXP#5: Static Content + Social Tags</td>
<td>78.01</td>
<td>93.19</td>
<td>84.93</td>
</tr>
</tbody>
</table>

* Results averaged over the 30 study subjects

Tag vs CB
Precision not improved
## Analysis of Precision

<table>
<thead>
<tr>
<th>Type of Content</th>
<th>Precision*</th>
<th>Recall*</th>
<th>F1*</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP#1: Static Content</td>
<td>75.86</td>
<td>94.27</td>
<td>84.07</td>
</tr>
<tr>
<td>EXP#2: Personal Tags</td>
<td>75.96</td>
<td>92.65</td>
<td>83.48</td>
</tr>
<tr>
<td>EXP#3: Social Tags</td>
<td>75.59</td>
<td>90.50</td>
<td>82.37</td>
</tr>
<tr>
<td>EXP#4: Static Content + Personal Tags</td>
<td>78.04</td>
<td>93.60</td>
<td>85.11</td>
</tr>
<tr>
<td>EXP#5: Static Content + Social Tags</td>
<td>78.01</td>
<td>93.19</td>
<td>84.93</td>
</tr>
</tbody>
</table>

* Results averaged over the 30 study subjects

Augmented vs CB Precision Improvement ≈ 2%
## Analysis of Recall

<table>
<thead>
<tr>
<th>Type of Content</th>
<th>Precision*</th>
<th>Recall*</th>
<th>F1*</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP#1: Static Content</td>
<td>75.86</td>
<td>94.27</td>
<td>84.07</td>
</tr>
<tr>
<td>EXP#2: Personal Tags</td>
<td>75.96</td>
<td>92.65</td>
<td>83.48</td>
</tr>
<tr>
<td>EXP#3: Social Tags</td>
<td>75.59</td>
<td>90.50</td>
<td>82.37</td>
</tr>
<tr>
<td>EXP#4: Static Content + Personal Tags</td>
<td>78.04</td>
<td>93.60</td>
<td>85.11</td>
</tr>
<tr>
<td>EXP#5: Static Content + Social Tags</td>
<td>78.01</td>
<td>93.19</td>
<td>84.93</td>
</tr>
</tbody>
</table>

Tag vs CB: Recall decrease 1.62% – 3.77%

* Results averaged over the 30 study subjects
## Analysis of Recall

<table>
<thead>
<tr>
<th>Type of Content</th>
<th>Precision*</th>
<th>Recall*</th>
<th>F1*</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP#1: Static Content</td>
<td>75.86</td>
<td>94.27</td>
<td>84.07</td>
</tr>
<tr>
<td>EXP#2: Personal Tags</td>
<td>75.96</td>
<td>92.65</td>
<td>83.48</td>
</tr>
<tr>
<td>EXP#3: Social Tags</td>
<td>75.59</td>
<td>90.50</td>
<td>82.37</td>
</tr>
<tr>
<td>EXP#4: Static Content + Personal Tags</td>
<td>78.04</td>
<td>93.60</td>
<td>85.11</td>
</tr>
<tr>
<td>EXP#5: Static Content + Social Tags</td>
<td>78.01</td>
<td>93.19</td>
<td>84.93</td>
</tr>
</tbody>
</table>

* Results averaged over the 30 study subjects

Augmented vs CB Recall decrease: 0.67% – 1.08%
### Analysis of F1

<table>
<thead>
<tr>
<th>Type of Content</th>
<th>Precision*</th>
<th>Recall*</th>
<th>F1*</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP#1: Static Content</td>
<td>75.86</td>
<td>94.27</td>
<td>84.07</td>
</tr>
<tr>
<td>EXP#2: Personal Tags</td>
<td>75.96</td>
<td>92.65</td>
<td>83.48</td>
</tr>
<tr>
<td>EXP#3: Social Tags</td>
<td>75.59</td>
<td>90.50</td>
<td>82.37</td>
</tr>
<tr>
<td>EXP#4: Static Content + Personal Tags</td>
<td>78.04</td>
<td>93.60</td>
<td>85.11</td>
</tr>
<tr>
<td>EXP#5: Static Content + Social Tags</td>
<td>78.01</td>
<td>93.19</td>
<td>84.93</td>
</tr>
</tbody>
</table>

*Results averaged over the 30 study subjects

**Overall accuracy F1 ≈ 85%**
## Putting Intelligence into CBRS: Challenges & Research Directions

<table>
<thead>
<tr>
<th>PROBLEMS</th>
<th>CHALLENGES</th>
<th>RESEARCH DIRECTIONS</th>
</tr>
</thead>
</table>
| Limited Content Analysis | Beyond keywords: novel strategies for the representation of items and profiles | ▪ Semantic analysis of content by means of external knowledge sources  
▪ Language-independent CBRS |
|                        | Taking advantage of Web 2.0 for collecting User Generated Content | Folksonomy-based CBRS                                        |
| Overspecialization      | Defeating homophily: recommendation diversification          | ▪ “computational” serendipity  
▪ Knowledge Infusion          |
Serendipity: Definitions

1. Serendipity
   - Making discoveries, by accidents and sagacity, of things which one were not in quest of (Horace Walpole, 1754)
   - The art of making an unsought finding (Pek van Andel, 1994) [vanAndel94]

2. Serendipitous ideas and findings
   - Gelignite by Alfred Nobel, when he accidentally mixed collodium (gun cotton) with nitroglycerin
   - Penicillin by Alexander Fleming
   - The psychedelic effects of LSD by Albert Hofmann
   - Cellophane by Jacques Brandenberger
   - The structure of benzene by Friedric August Kekulé

The challenge

1. Serendipity in RSs is the experience of receiving an *unexpected* and fortuitous, but *useful* advice
   - ✓ it is a way to *diversify* recommendations
2. The challenge is *programming* for serendipity
   - ✓ to find a manner to introduce serendipity into the recommendation process in an *operational way*
Strategies for computational serendipity [Toms00]

1. “Blind Luck”: random recommendations
2. “Prepared Mind”: Pasteur principle ("chance favors the prepared mind") - deep user modeling
3. “Anomalies and Exceptions”: searching for dissimilarity [Iaquinta10]
4. “Reasoning by Analogy”


Programming for Serendipity into CBRS: “Anomalies and Exceptions”

1. **Basic recommendation list** defined by the **best N** items ranked according to the user profile

2. **Idea for inducing serendipity**
   - extending the basic list with items programmatically supposed to be serendipitous for the active user
**Item Recommender (ITR)**

- Content-based recommender developed at Univ. of Bari [Semeraro07]
  - learns a probabilistic model of the interests of the user from textual descriptions of items
  - **user profile** = binary text classifier able to categorize items as interesting (LIKES) or not (DISLIKES)
  - a-posteriori probabilities as classification scores for LIKES and DISLIKES

Recommendation process: Ranked list approach

Profile Learner

USER PROFILE

<table>
<thead>
<tr>
<th>LIKES</th>
<th>DISLIKES</th>
</tr>
</thead>
<tbody>
<tr>
<td>future</td>
<td>alien</td>
</tr>
<tr>
<td>...</td>
<td>violence</td>
</tr>
<tr>
<td>...</td>
<td>blood</td>
</tr>
</tbody>
</table>

P(LIKES | ALF)

0.89

0.74

0.22
Potentially serendipitous items selected on the ground of categorization scores for LIKES and DISLIKES

- difference of classification scores tends to zero → uncertain classification
  
  \[ | P(\text{LIKES} \mid \text{ITEM}) - P(\text{DISLIKES} \mid \text{ITEM}) | \approx 0 \]

- assumption:
  
  uncertain classification ≡ items not known by the user
Programming for Serendipity into ITR: example

1. Basic recommendation list = $N$ most interesting items
2. Ranked list of “unpredictable” items obtained from ITR
3. Basic recommendation list augmented with some serendipitous items

USER PROFILE

<table>
<thead>
<tr>
<th>LIKES</th>
<th>DISLIKES</th>
</tr>
</thead>
<tbody>
<tr>
<td>future</td>
<td>violence</td>
</tr>
<tr>
<td>alien</td>
<td>blood</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

$P(\text{LIKES} | \text{ITEM})$

$| P(\text{LIKES} | \text{ITEM}) - P(\text{DISLIKES} | \text{ITEM}) |$
What about evaluation?

1. Classic evaluation metrics (Precision, Recall, F, MAE,…) don’t take into account obviousness, novelty and serendipity
   - Accurate recommendation ≠ Useful recommendation
   - emotional response associated with serendipity difficult to capture by conventional accuracy metrics
   - serendipity degree impossible to evaluate without considering user feedback

2. Novel metrics required
   - planned as a future work
Programming for Serendipity: cross-domain recommendations

Bram Stoker's Dracula [DVD] [1992]

DVD ~ Gary Oldman

RRP: £6.99
Price: £4.77 & eligible for Free UK delivery on orders over £5 with Super Saver Delivery. See details and conditions
You Save: £1.22 (20%)

In stock.
Items for dispatch to UK will be sold by Amazon's Preferred Merchant. (Why?) Gift-wrap available.
Only 2 left in stock—order soon.

16 new from £2.65
17 used from £1.00

Rent DVDs from LOVEFiLM.com
Amazon's choice for DVD rental.
With a 14 day FREE trial. Learn more

See larger image
Share your own customer images

s summertime, and the purchasing and watching of DVDs and Blu-rays is easy... Check out the Hottest Summer Offers in DVD.

Frequently Bought Together
Customers buy this item with The Shawshank Redemption [DVD] [1995]

Price For Both: £7.75

Surprise for you
Holiday in Transilvania

Add to Basket
“Reasoning by Analogy”: a serendipity strategy for cross-domain recommendations

User profile for Movies

<table>
<thead>
<tr>
<th>Cowboy</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amelie</td>
<td>0.5</td>
</tr>
<tr>
<td>Vampire</td>
<td>0.9</td>
</tr>
<tr>
<td>Pirate</td>
<td></td>
</tr>
</tbody>
</table>

“Parallel” user profile for Travels

<table>
<thead>
<tr>
<th>Texas</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paris</td>
<td></td>
</tr>
<tr>
<td>Transylvania</td>
<td>0.5</td>
</tr>
<tr>
<td>Caribbean</td>
<td>0.9</td>
</tr>
<tr>
<td>Zion</td>
<td></td>
</tr>
</tbody>
</table>

DBpedia

Ontology

Matrix
Ongoing work: DEVIUS

1. Analogy engine for computing “parallel” user profiles
   ✓ Spreading activation on DBpedia for mapping between domains
2. Open source code of DEVIUS available in September
3. Experimental evaluation
   ✓ books / movies
## Putting Intelligence into CBRS: Challenges & Research Directions

<table>
<thead>
<tr>
<th>PROBLEMS</th>
<th>CHALLENGES</th>
<th>RESEARCH DIRECTIONS</th>
</tr>
</thead>
</table>
| Limited Content Analysis | Beyond keywords: novel strategies for the representation of items and profiles | - Semantic analysis of content by means of external knowledge sources  
                       |                                                      |   - Language-independent CBRS |
|                     | Taking advantage of Web 2.0 for collecting User Generated Content | Folksonomy-based CBRS                                    |
| Overspecialization  | Defeating homophily: recommendation diversification  | - “computational” serendipity → programming for serendipity  
                       |                                                      |   - Knowledge Infusion       |
Knowledge Infusion (KI)

1. Humans typically have the *linguistic* and *cultural* experience to comprehend the meaning of a text
   - How to realize this *capability* into machines?

2. In NLP tasks, computers require access to vast amounts of common-sense and domain-specific world knowledge
   - Infusing lexical knowledge → Dictionaries (e.g. WordNet)
   - Infusing cultural knowledge → Wikipedia
   - ...

Enhancing CBRS by KI

1. Modeling the unstructured information stored in several (open) knowledge sources
2. Exploiting the acquired knowledge in order to better understand the item descriptions and extract more meaningful features
3. Inspired by a language game: The Guillotine [Semeraro09b]

The Guillotine: the game

Let’s try to play the game

APPLE

“An apple a day takes the doctor away”

JUDGMENT

Day of Judgment

SUNRISE

Beginning of the day

INDEPENDENCE

Independence day

SLEEPER

Daysleeper, a famous song by R.E.M.
What does OTTHO know about ‘stars’?

**Definitions**

**Star**: any one of the distant bodies appearing as a point of light in the sky at night | *Fixed star*, i.e. one which is not a planet

**Compound Forms**

```
<table>
<thead>
<tr>
<th>Lemma: Definitions</th>
<th>Compound Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star</td>
<td>any one of the distant bodies appearing as a point of light in the sky at night</td>
</tr>
</tbody>
</table>
```

**Dictionary Matrix**

```
<table>
<thead>
<tr>
<th>STAR</th>
<th>LIGHT</th>
<th>SKY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STAR</td>
<td>0.55</td>
<td>1.45</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

**Tag Matrix**

```
<table>
<thead>
<tr>
<th>ALIEN</th>
<th>SPACE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>STAR</td>
<td>0.27</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
```

**Tags in items’ tag cloud**

- SPACE 1.41
- ALIEN 0.27
- "STAR, SPACE, ALIEN"
KI@work for recommendation diversification

Plot Keywords
- STAR
- ROBOT
- ALIEN
- WAR
- BATTLE

KI-LIST
- SPACE 0.36
- FUTURE 0.10
- EXTRATERRESTRIAL 0.08
- CYBORG 0.07
- FIGHT 0.02
- JUSTICE 0.01
- ...

Search Results
Concluding Remarks

1. Research directions for overcoming some CBRS drawbacks
   - Main strategies adopted to introduce some semantics in the recommendation process
   - Main strategies for diversifying recommendations

2. Research agenda: glean meaning and user thought from the precious boxes (brain, Web, social networks, ...) they are hidden into:
   - fMRI & Eye/Head-tracking technologies for a new generation of evaluation metrics
   - Linked Open Data: interlinking user profiles with Semantic Web data and LOD
   - Semantic Cross-system Personalization: semantic matching of user profiles coming from heterogeneous systems
Thanks...  

...for your attention...  

...Questions?  

Semantic Web Access and Personalization research group  
http://www.di.uniba.it/~swap  

Pierpaolo Basile  
Marco de Gemmis  
Leo Iaquinta  
Piero Molino  
Fedelucio Narducci  
Eufemia Tinelli  

Annalina Caputo  
Michele Filannino  
Pasquale Lops  
Cataldo Musto  
Giovanni Semeraro
References 1/4


