WordNet-based User Profiles for Semantic Personalization

Giovanni Semeraro, Marco Degemmis, Pasquale Lops, Ignazio Palmisano

LACAM – Knowledge Acquisition and Machine Learning Lab
Department of Computer Science – University of Bari

July 24th, 2005, Edinburgh, Scotland, UK
Outline

✔ Introduction
✔ Personalized Access
✔ WordNet-based profiles
✔ Experiments
✔ Final Remarks
Today’s Information Society

Problems…

- Explosion of irrelevant information
- Users overloaded by this information

…and consequences

- Searching is time consuming
- Need for intelligent solutions to support users
Personalized Portals & Stores
Learning User Profiles as a Text Categorization Problem

Preferences
Arts & Photography
Children’s books
Computers & Internet

content-based recommendations by learning from TEXT and users' ratings on items
Research Goal

Intelligent Information Access =

3. Personalized Access by user profiles +
4. Semantic Access by concept identification in documents

USER PROFILE: A STRUCTURED REPRESENTATION OF USER INTERESTS AND PREFERENCES

1. Automated induction of user profiles by means of supervised machine learning techniques
2. Taking into account the meaning of the words
Intelligent Personalized Information Access

- **Introduction**
  - Training documents
  - Positive examples
  - Negative examples

- **Personalized Access**
  - Learning
  - Concept-based Profile

- **WordNet-based profiles**
  - Concept identification
  - Concept-based document representation

- **Experiments**
  - Documents
  - Matching
Movie Recommending on the Web

**Instance**
(movie)

- **Title**
- **Director**
- **Cast**
- **Summary**
- **Keywords**

**Bag of Words (BOW)**

Tokenization + Stopword elimination + Stemming

User Ratings: 0-5

**Movie**

*Young Frankenstein (1974)*

- Directed by Mel Brooks

**Keywords**

- User Ratings: 0-5
- **Keywords**

**Summary**

A young neurosurgeon (Gene Wilder) inherits the castle of his grandfather, the famous Dr. Victor von Frankenstein. In the castle he finds a funny hunchback called Igor, a pretty lab assistant named Inga and the old housekeeper, Frau Blucher -miklahh-. Young Frankenstein believes that the work of his grandfather is only crap, but when he discovers the book where the mad doctor described his reanimation experiment, he suddenly changes his mind and tries to create his own creature. The result of this experiment is a giant hairy monster who tastes good, and the mad doctor's grandson decides to repeat the experiments to the granddad's castle and repeats the experiments: (more) (view trailer)
Word Sense Disambiguation (WSD)

1. Many meanings for polysemous words, known as senses
2. One sense at a time is used in a specific context.
3. Deciding which sense to use is Word Sense Disambiguation

Approaches to WSD

- **Knowledge-based**: uses Machine Readable Dictionaries
- **Corpus-based**: uses sense-tagged corpus
WordNet

1. Lexical reference database whose design is inspired by current psycholinguistic theories of human lexical memory
   - The work started in 1985 by a group of psychologists and linguists at Princeton University

2. English *nouns*, *verbs*, *adverbs* and *adjectives* are organized into SYNonym SETs, each representing one underlying lexical concept

3. Relations among synsets can be used to engineer a change of representation in text data by transforming vectors of words into vectors of word meanings
   - The synonymy relation can be used to map words with similar meanings together
   - Hypernymy (corresponding to the IS–A relation) can be used to generalize noun and verb meanings to a higher level of abstraction

http://www.cogsci.princeton.edu/~wn/
**Synset Semantic Similarity**

24: `function SINSIM(a, b)` \[ \text{▷ The similarity of the synsets } a \text{ and } b \]

25: \[ N_p \leftarrow \text{the number of nodes in path } p \text{ from } a \text{ to } b \]

26: \[ D \leftarrow \text{maximum depth of the taxonomy} \]

27: \[ r \leftarrow -\log(N_p/2D) \]

28: `return r`

29: `end function`

\[ \text{SINSIM(cat,mouse) = } -\log(5/32)=0.806 \]
A document $d$ is mapped into a list of WordNet synsets following these steps:

1. Each monosemous word $w$ in a slot of a document $d$ is mapped into the corresponding WordNet synset;

2. For each couple of words $<\text{noun, noun}>$ or $<\text{adjective, noun}>$, a search in WordNet is made in order to verify if at least one synset exists for the bigram $<w_1, w_2>$. In the positive case, WSD algorithm is applied on the bigram, otherwise it is applied separately on $w_1$ and $w_2$, using all words in the slot as the context $C$ of $w$;

3. Each polysemous unigram $w$ is disambiguated, using all words in the slot as the context $C$ of $w$. 
Experimental Evaluation

1. Extended Eachmovie
   - Internet Movie Database

2. 10-fold stratified cross-validation
   - Precision, Recall, F-measure, NDPM

3. Movie relevant if rating >2
   - Rocchio: Cosine Similarity (positive/negative profile)

4. Experiments: BOW-generated profiles vs. BOS-generated profiles
   - Wilcoxon signed rank test
   - Low number of independent trials
   - Classification for each genre is a trial
   - Significance level $p < 0.05$
The EachMovie Dataset

1. Project conducted by Compaq Research Centre (1996-1997)
2. Dataset of user-movie ratings
   - About 2.8 millions ratings
   - Over 72,000 users
   - 1,628 items (movies) subdivided in 10 categories
   - Discrete rating between 0 and 5
   - Movies content crawled from the Internet Movie Database (IMDb)
3. 10 movie categories
   - 933 randomly selected users
   - 100 users for each category, only for Category 2 – Animation, 33 users selected
   - Each user rated between 30 and 100 movies
## Extended Eachmovie (ratings+content)

<table>
<thead>
<tr>
<th>Id Genre</th>
<th>Genre</th>
<th>#Rated Movies</th>
<th>%POS</th>
<th>%NEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Action</td>
<td>4,474</td>
<td>72.05</td>
<td>27.95</td>
</tr>
<tr>
<td>2</td>
<td>Animation</td>
<td>1,103</td>
<td>56.67</td>
<td>43.33</td>
</tr>
<tr>
<td>3</td>
<td>Art_Foreign</td>
<td>4,246</td>
<td>76.21</td>
<td>23.79</td>
</tr>
<tr>
<td>4</td>
<td>Classic</td>
<td>5,026</td>
<td>91.73</td>
<td>8.27</td>
</tr>
<tr>
<td>5</td>
<td>Comedy</td>
<td>4,714</td>
<td>63.46</td>
<td>36.54</td>
</tr>
<tr>
<td>6</td>
<td>Drama</td>
<td>4,880</td>
<td>76.24</td>
<td>23.76</td>
</tr>
<tr>
<td>7</td>
<td>Family</td>
<td>3,808</td>
<td>63.71</td>
<td>36.29</td>
</tr>
<tr>
<td>8</td>
<td>Horror</td>
<td>3,631</td>
<td>59.89</td>
<td>40.11</td>
</tr>
<tr>
<td>9</td>
<td>Romance</td>
<td>3,707</td>
<td>72.97</td>
<td>27.03</td>
</tr>
<tr>
<td>10</td>
<td>Thriller</td>
<td>3,709</td>
<td>71.94</td>
<td>28.06</td>
</tr>
</tbody>
</table>

- Green: 60-65% positive
- Yellow: 70-75% positive
- Red: 75-100% positive
# Bag of Synsets

## Bag of Words

<table>
<thead>
<tr>
<th>Id</th>
<th>Movie</th>
<th>Word Form</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td></td>
<td>aaron</td>
<td>1</td>
</tr>
<tr>
<td>67</td>
<td></td>
<td>murder</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1134</td>
<td></td>
<td>roll</td>
<td>3</td>
</tr>
<tr>
<td>1134</td>
<td></td>
<td>wheel</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1161</td>
<td></td>
<td>zoom</td>
<td>1</td>
</tr>
</tbody>
</table>

#Features=172.296

## Bag of Synsets

<table>
<thead>
<tr>
<th>Id</th>
<th>Movie</th>
<th>Word Form</th>
<th>Id Synset</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td></td>
<td>aaron</td>
<td>8844021</td>
<td>1</td>
</tr>
<tr>
<td>67</td>
<td></td>
<td>murder</td>
<td>6712568</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1134</td>
<td></td>
<td>roll</td>
<td>2051720</td>
<td>5</td>
</tr>
<tr>
<td>1161</td>
<td></td>
<td>zoom</td>
<td>1618551</td>
<td>1</td>
</tr>
</tbody>
</table>

#Features=107.990

1. 38% Reduction of features representing movies in the EachMovie dataset
   - Mainly on slots containing proper names
2. Recognition of bigrams
3. Synonyms represented by the same synsets
## Semantic Profiles Evaluation

<table>
<thead>
<tr>
<th>Id</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>NDPM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BOW</td>
<td>BOS</td>
<td>BOW</td>
<td>BOS</td>
</tr>
<tr>
<td>1</td>
<td>0.72</td>
<td>0.75</td>
<td>0.82</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.79</td>
<td>0.75</td>
<td>0.79</td>
</tr>
<tr>
<td>2</td>
<td>0.65</td>
<td>0.64</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>0.64</td>
<td>0.63</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>3</td>
<td>0.77</td>
<td>0.85</td>
<td>0.79</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>0.77</td>
<td>0.84</td>
<td>0.77</td>
<td>0.84</td>
</tr>
<tr>
<td>4</td>
<td>0.92</td>
<td>0.94</td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>0.93</td>
<td>0.94</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>5</td>
<td>0.66</td>
<td>0.69</td>
<td>0.72</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>0.67</td>
<td>0.70</td>
<td>0.67</td>
<td>0.70</td>
</tr>
<tr>
<td>6</td>
<td>0.78</td>
<td>0.79</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>0.80</td>
<td>0.81</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>7</td>
<td>0.68</td>
<td>0.74</td>
<td>0.75</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>0.69</td>
<td>0.77</td>
<td>0.69</td>
<td>0.77</td>
</tr>
<tr>
<td>8</td>
<td>0.64</td>
<td>0.69</td>
<td>0.74</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>0.67</td>
<td>0.73</td>
<td>0.67</td>
<td>0.73</td>
</tr>
<tr>
<td>9</td>
<td>0.73</td>
<td>0.76</td>
<td>0.79</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>0.74</td>
<td>0.77</td>
<td>0.74</td>
<td>0.77</td>
</tr>
<tr>
<td>10</td>
<td>0.74</td>
<td>0.75</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>0.77</td>
<td>0.78</td>
<td>0.77</td>
<td>0.78</td>
</tr>
<tr>
<td>Mean</td>
<td>0.73</td>
<td>0.76</td>
<td>0.78</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>0.74</td>
<td>0.84</td>
<td>0.74</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Mean F1: 0.74
Mean Recall: 0.83
Mean Precision: 0.78
Results

1. Improvement in precision (+3%) and recall (+5%)

2. The BOS model outperforms the BOW model specifically on datasets:
   - 3 (+8% of precision, +7% of recall)
   - 7 (+6% of precision, +9% of recall)
   - 8 (+5% of precision, +8% of recall)

3. No improvement on dataset 2 (Animation)
   - Low number of rated movies
   - WSD errors (difficulty in disambiguating stories)
Conclusions & Future Works

1. Extending BOW to BOS improves classification accuracy when WSD is performed on short documents.
2. Improved results are independent from the distribution of positive and negative examples in the dataset.

1. Integration of user profiles into UUCM [Metha et al. 2005]
2. Ontologies and user profiles
   - Domain-specific ontologies