



10th International Conference on User Modeling



**PIA 2005 – Workshop on New Technologies for
Personalized Information Access**

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



My...Web?

1/2

Change colors: Simple White | Classic Blue

Select a Category: **Web** | News | Images | Music | Desktop^{BETA} | Encarta

msn Search the Web: Search

Thursday, Mar 17

HOTMAIL | MESSENGER | MY MSN | SIGN IN

St. Paddy's Day Where to Party

Parades, pubs & pints of green beer in 23 cities

- See 27 'green' e-cards
- Sample Irish tunes: U2, The Pogues & more

Microsoft Office has evolved. Have you? **Office**

msn Search

Personalized Portals & Stores

msn My MSN Search the Web: Search Help

Column Options

Welcome **marko**

This day in history

Add Content

Search or choose from MSN recommended sources for content to add to your page. My MSN accepts RSS sources.

Search for content to add:

My favorite links

- Microsoft
- MSNBC News
- Hotmail

Change details

Today on MSN

Thursday, Mar 17

Why Lenders Share Blame for ID Theft

Agencies' lenient policies aid & abet the bad guys

- 10 tips to stop ID theft
- Do you know an identity thief? 8 signs

Highlights

- 10 hot businesses to start now
- Worst 'Green Score' vehicles
- Sale: Up to 75% off luxury bedding
- Book smarts vs. street smarts
- Offer: Prep for spring & hire a pro
- Offer: Free Hotmail from MSN

MSNBC Front Page news

Peterson gets death penalty

A judge on Wednesday sentenced Scott Peterson to death for the murder of his wife and their unborn trial.

- Jury finds Blake not guilty

Weather forecast

Get quick forecast:

Los Angeles, CA

Today	Fri	Sat	Sun
69°/54°F	63°/53°F	60°/53°F	62°/53°F

London, United Kingdom 60°/51°F

New York, NY 46°/33°F

Provided by The Weather Channel®

Change details

Travel Conditions · Pollen Reports · Lawn & Garden

Stock quotes

Get quick quote:

Name/Symbol	Last Change
Dow (\$INDU)	10,833.07 -112.03
NASDAQ (\$COMPX)	2,015.75 0.00
S&P (\$INX)	1,188.07 0.00

Quotes supplied by Comstock, 20 min. delay.

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

Book description at Amazon.com

amazon.com. VIEW CART | WISH LIST | YOUR ACCOUNT | HELP

WELCOME YOUR STORE BOOKS APPAREL & ACCESSORIES ELECTRONICS TOYS & GAMES KITCHEN & HOUSEWARES COMPUTER & VIDEO GAMES SEE MORE STORES

SEARCH BROWSE SUBJECTS BESTSELLERS MAGAZINES CORPORATE ACCOUNTS E-BOOKS & DOCS BARGAIN BOOKS USED BOOKS

Smaller, faster, smarter...cheaper
Shop new releases in Digital Cameras [Shop now!](#)

Swing, Second Edition
by [Matthew Robinson](#), [Pavel Vorobiev](#), [Pavel A. Vorobiev](#), [David Karr](#)

List Price: \$49.95
Price: **\$34.97** & This item ships for FREE with Super Saver Shipping. [See details.](#)
You Save: \$14.98 (30%)
Availability: Usually ships within 24 hours

9 used & new from \$30.14

Edition: Paperback
[see larger photo](#)
[See more product details](#)

READY TO BUY?
[Add to Shopping Cart](#)
or
[Sign in](#) to turn on 1-Click ordering.

MORE BUYING CHOICE
9 used & new from \$30.14
Have one to sell? [Sell yours](#)
[Add to Wish List](#)
[Add to Wedding Registry](#)

Editorial Reviews

Amazon.com

Written for the experienced Java developer, *Swing* provides an in-depth guide to getting the most out of Sun's Swing/JFC user interface classes. Mixing real-world code examples and expert advice on advanced features, this book shows how to make use of this powerful library effectively within your own projects.

The best thing about this text has to be its sample programs, many of which incorporate other Java APIs to do "real" work. For example, a demo of the scroll pane Swing component uses other JFC classes to display JPG images. For working with lists, the authors show how to process .ZIP files in Java. For demonstrating table programming, there's coverage of JDBC to connect to databases. Other standout code samples include a working FTP client and a fully functional RTF word processor. (Many of these examples are enhanced in separate steps, showing off new Swing classes and features along the way.) The authors do a particularly good job of annotating code with clear explanations referenced with numbered bullets that point out important lines of code.

The other noteworthy feature here is the material on extending basic Swing functionality through custom code. (To use Swing effectively, you definitely need to be able to customize its classes. The authors show you how.) There are examples for enhancing Swing with custom layout managers and numerous samples that extend trees and tables, and even a section on the basics of creating new pluggable look and feel (PLAF) modules for Swing.

Research Goal

Intelligent Information Access =

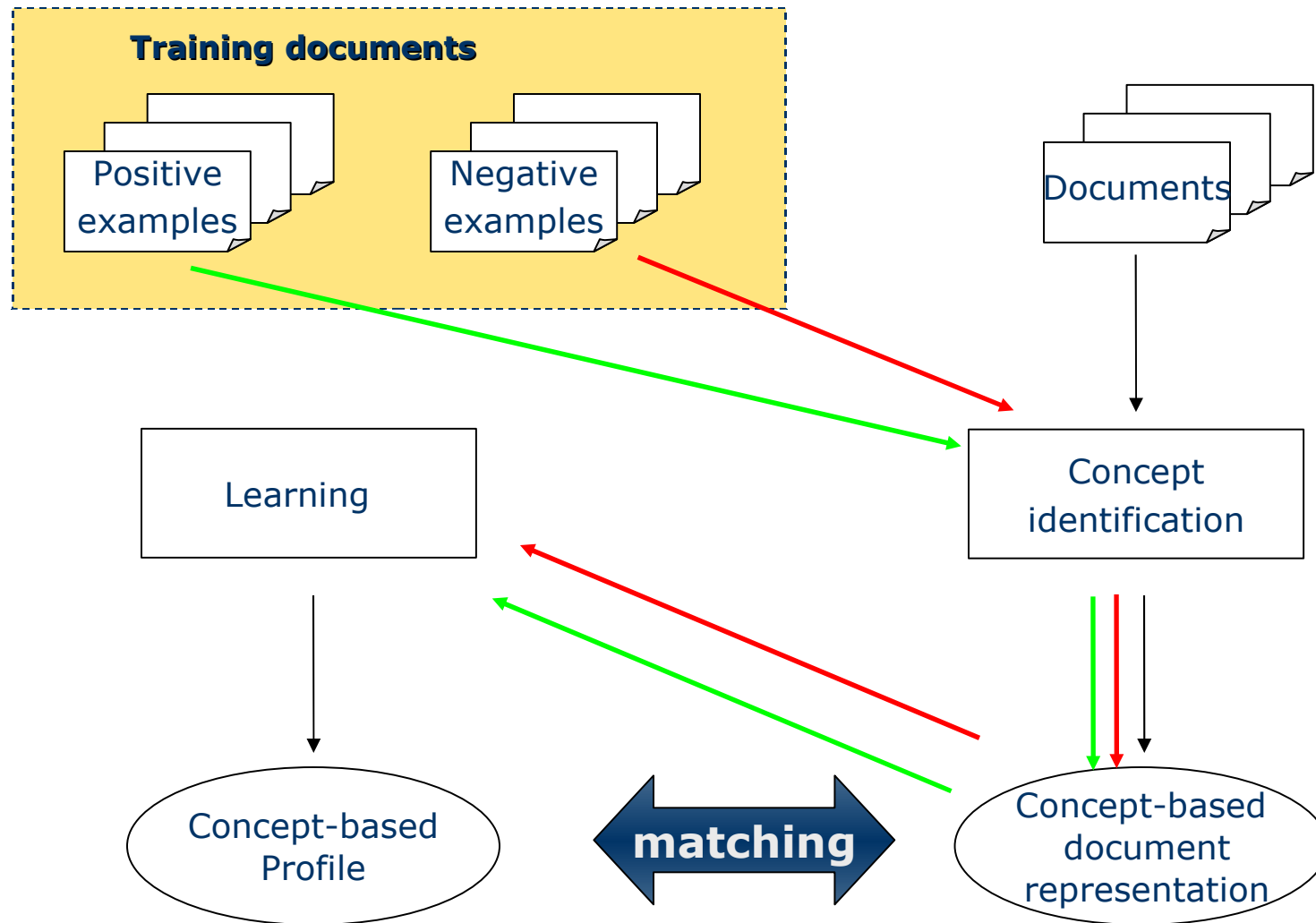
3. **Personalized** Access by **user profiles** +
4. **Semantic** Access by **concept identification** in documents

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



Movie Recommending on the Web

Young Frankenstein (1974)



Tokenization +
Stopword elimination +
Stemming

Bag of Words (BOW)

[Add to MyMovies](#) [IMDbPro Professional Details](#)

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 -mühhhl-. 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... granddad's castle and repeats the experiments. [\(more\)](#) [\(view trailer\)](#)



User Ratings: 0-5

Instance
(movie)

Title

Director

Cast

Summary

Keywords

Word Sense Disambiguation (WSD)

- 1 Many meanings for polisemous 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 **SYN**onym **SET**s, 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

Synset Semantic Similarity

```

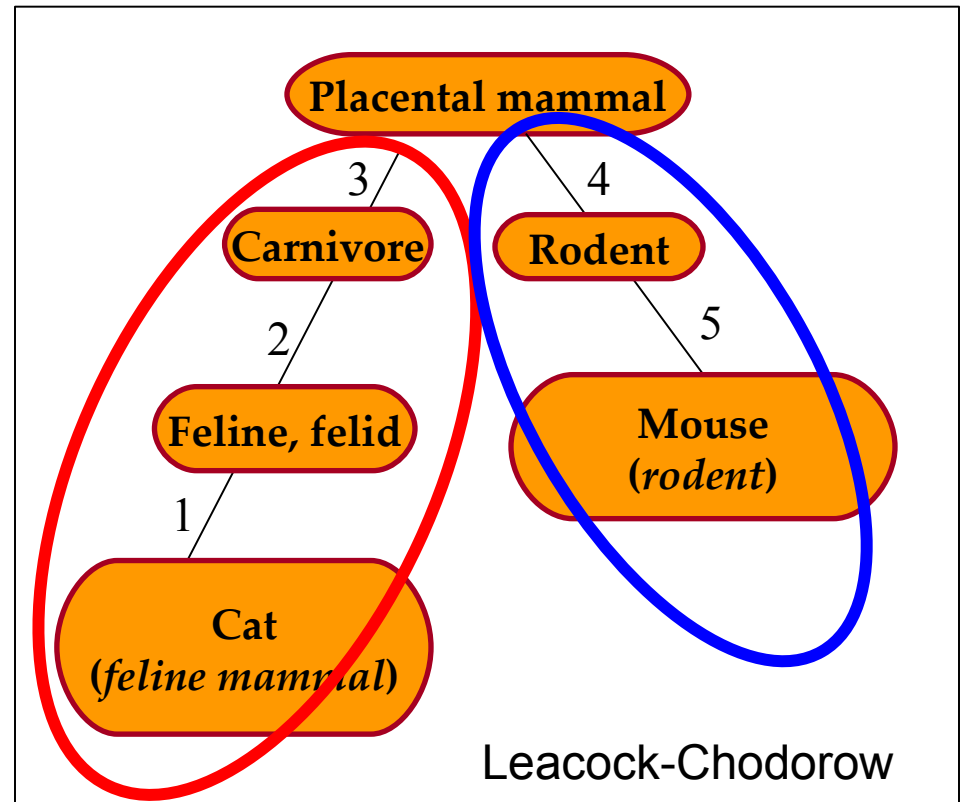
24: function SINSIM(a, b)
25:    $N_p \leftarrow$  the number of nodes in path p from a to b
26:    $D \leftarrow$  maximum depth of the taxonomy
27:    $r \leftarrow -\log(N_p/2D)$ 
28:   return r
29: end function

```

▷ The similarity of the synsets *a* and *b*

▷ In WordNet 1.7.1 $D = 16$

$$\text{SINSIM}(\text{cat}, \text{mouse}) = -\log(5/32) = 0.806$$



[Fellbaum 1998]

Semantic Indexing

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 $\langle noun, noun \rangle$ or $\langle adjective, noun \rangle$, a search in WordNet is made in order to verify if at least one synset exists for the bigram $\langle w_1, w_2 \rangle$. 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)

<i>Id Genre</i>	<i>Genre</i>	<i>#Rated Movies</i>	<i>%POS</i>	<i>%NEG</i>
1	 Action	4,474	72.05	27.95
2	 Animation	1,103	56.67	43.33
3	 Art_Foreign	4,246	76.21	23.79
4	 Classic	5,026	91.73	8.27
5	 Comedy	4,714	63.46	36.54
6	 Drama	4,880	76.24	23.76
7	 Family	3,808	63.71	36.29
8	 Horror	3,631	59.89	40.11
9	 Romance	3,707	72.97	27.03
10	 Thriller	3,709	71.94	28.06
		39,298	71.84	28.16

 60-65% positive

 70-75% positive

 75-100% positive

Bag of Synsets

Bag of Words

<i>Id Movie</i>	<i>Word Form</i>	<i>Occurrence</i>
31	aaron	1
67	murder	1
...
1134	roll	3
1134	wheel	2
...
1161	zoom	1

#Features=172.296

Bag of Synsets

<i>Id Movie</i>	<i>Word Form</i>	<i>Id Synset</i>	<i>Occurrence</i>
31	aaron	8844021	1
67	murder	6712568	1
...
1134	roll	2051720	5
...
1161	zoom	1618551	1

#Features=107.990

- 38% Reduction of features representing movies in the EachMovie dataset
 - ✓ Mainly on slots containing proper names
- Recognition of bigrams
- Synonyms represented by the same synsets

Semantic Profiles Evaluation

<i>Id</i>	<i>Precision</i>		<i>Recall</i>		<i>F1</i>		<i>NDPM</i>	
	<i>BOW</i>	<i>BOS</i>	<i>BOW</i>	<i>BOS</i>	<i>BOW</i>	<i>BOS</i>	<i>BOW</i>	<i>BOS</i>
1	0.72	0.75	0.82	0.86	0.75	0.79	0.46	0.44
2	0.65	0.64	0.66	0.66	0.64	0.63	0.34	0.38
3	0.77	0.85	0.79	0.86	0.77	0.84	0.46	0.48
4	0.92	0.94	0.94	0.96	0.93	0.94	0.45	0.43
5	0.66	0.69	0.72	0.75	0.67	0.70	0.44	0.46
6	0.78	0.79	0.84	0.87	0.80	0.81	0.45	0.45
7	0.68	0.74	0.75	0.84	0.69	0.77	0.41	0.40
8	0.64	0.69	0.74	0.82	0.67	0.73	0.42	0.44
9	0.73	0.76	0.79	0.81	0.74	0.77	0.48	0.48
10	0.74	0.75	0.85	0.84	0.77	0.78	0.45	0.44
Me an	0.73	0.76	0.78	0.83	0.74	0.84	0.44	0.44

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

