10th International Conference on User Modeling

UM'05

DELOS Network of Excellence on Digital Libraries

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

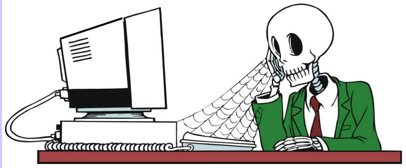
Today's Information Society

Problems…

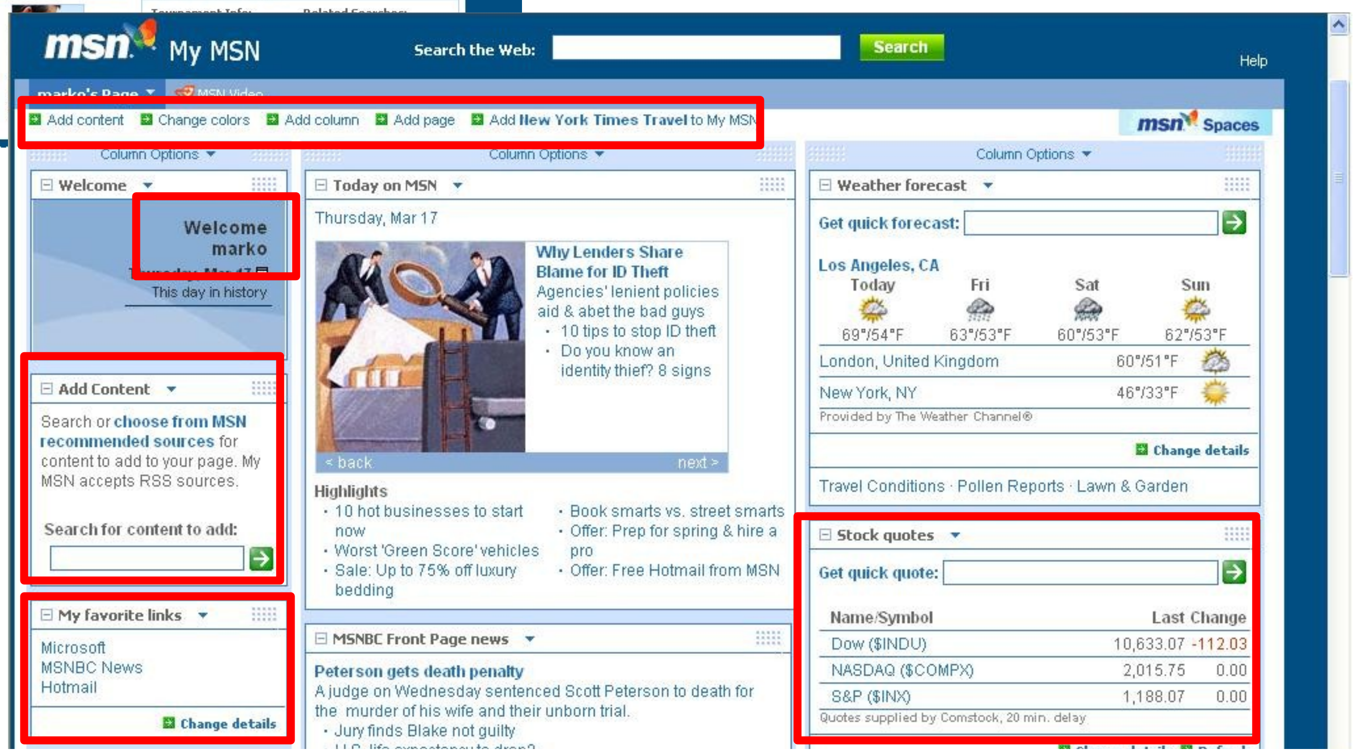
- Explosion of irrelevant information
- Users overloaded by this information

…and consequences

- Searching is time consuming
- Need for intelligent solutions to support users



My…Web?



Personalized Portals & Stores

Learning User Profiles as a Text Categorization Problem

Preferences

Arts & Photography

Children's books

Computers & Internet

Book description at Amazon.com

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

Research Goal

Intelligent Information Access =

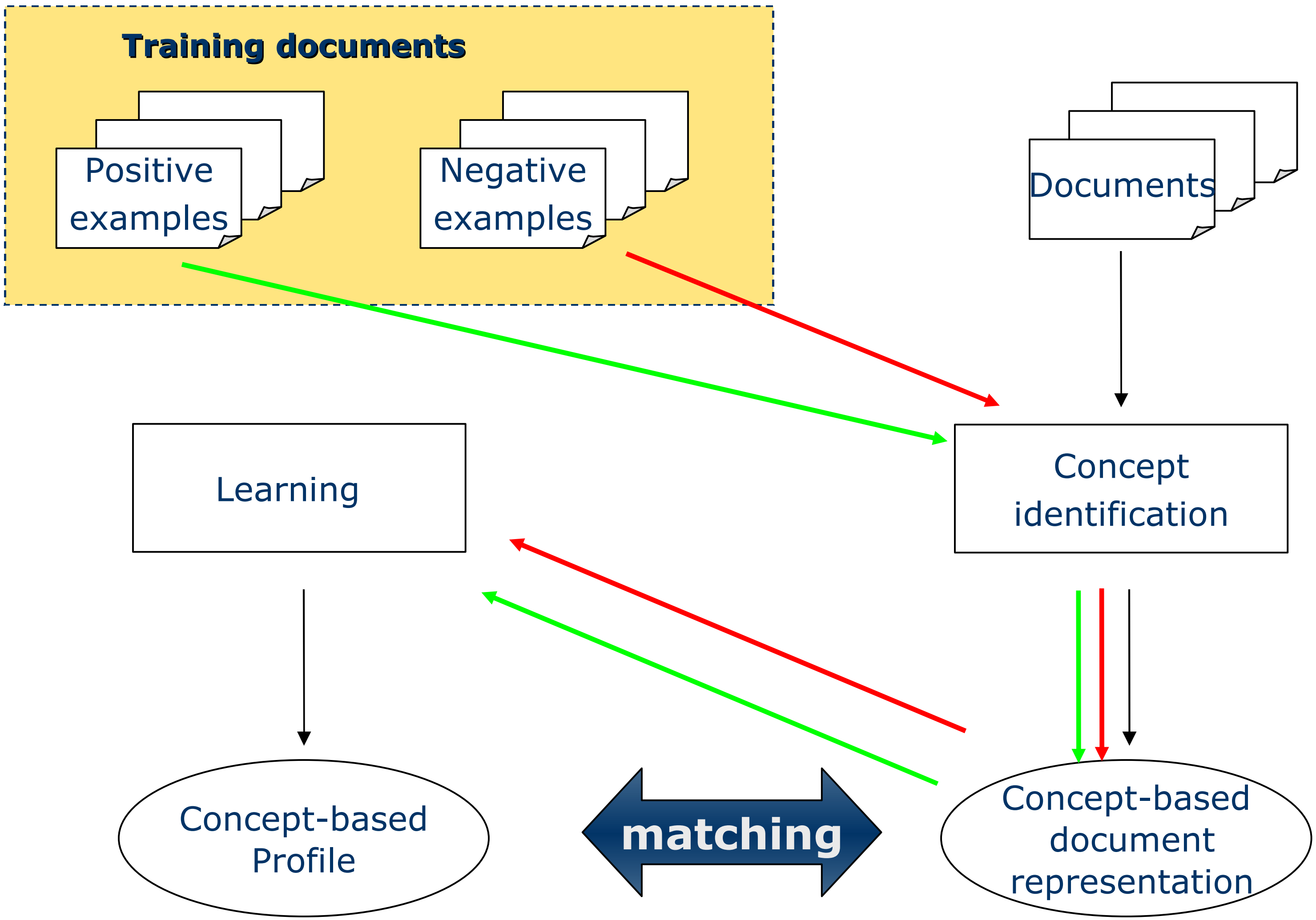
3. Personalized Access by user profiles +
4. Semantic Access by concept identification in documents

USER PROFILE: A STRUCTURED REPRESENTATION OF USER INTERESTS AND PREFERENCES

❶ Automated induction of user profiles by means of supervised machine learning techniques

❷ Taking into account the meaning of the words

Intelligent Personalized Information Access



Movie Recommending on the Web

Young Frankenstein (1974)

Directed by
Mel Brooks

Writing credits
Mary Shelley (novel)
Gene Wilder (screen story) ...
(more)

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 -iiiihhh!-. 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)

Tokenization +
Stopword elimination +
Stemming

Bag of Words (BOW)

Instance (movie) → Title, Director, Cast, Summary, Keywords

User Ratings:0-5

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 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

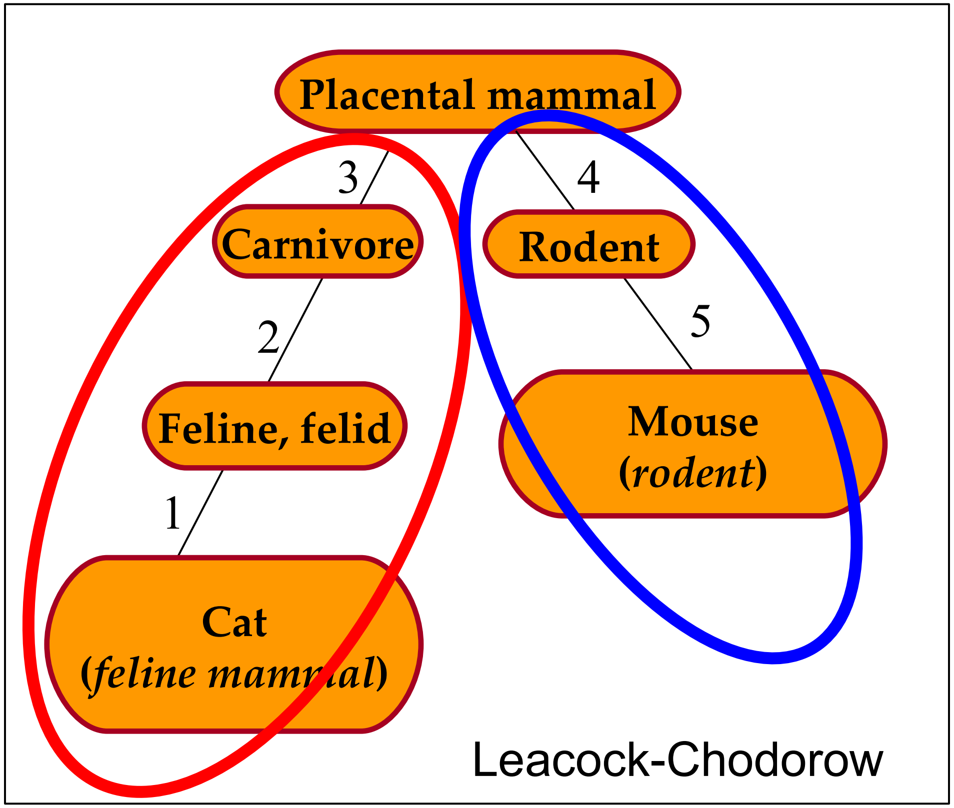
http://www.cogsci.princeton.edu/~wn/

Synset Semantic Similarity

function SINSIM(a, b) ▷ The similarity of the synsets a and b
  $N_p \leftarrow$ the number of nodes in path p from a to b
  $D \leftarrow$ maximum depth of the taxonomy ▷ In WordNet 1.7.1 $D = 16$
  $r \leftarrow -log(N_p/2D)$
  return r
end function

```
SINSIM(cat,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 <*noun,noun*> or <*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)

Id Genre	Genre	#Rated Movies	%POS	%NEG
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

Id Movie	Word Form	Occurrence
31	aaron	1
67	murder	1
…	…	…
1134	roll	3
1134	wheel	2
…	…	…
1161	zoom	1

#Features=172.296

Bag of Synsets

Id Movie	Word Form	Id Synset	Occurrence
31	aaron	8844021	1
67	murder	6712568	1
…	…		…
1134	roll	2051720	5
…	…		…
1161	zoom	1618551	1

#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

Id	Precision		Recall		F1		NDPM	
	BOW	BOS	BOW	BOS	BOW	BOS	BOW	BOS
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
Mean	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 in 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

VIKEF

Information Society Technologies

DELOS Network of Excellence on Digital Libraries