Intelligent Information Access
Methods, Perspectives and Applications

Prof. Giovanni Semeraro
semeraro@di.uniba.it
Marco Degemmis, Ph.D
degemmis@di.uniba.it
Pasquale Lops, Ph.D
lops@di.uniba.it

Dipartimento di Informatica – Università di Bari

Outline

Today's Information Society

People across the world…

- Chat
- Exchange e-mail, sms, pictures
- Buy products and services online
- Use search engines to find information their work and day-to-day life
- Exploit the Web for obtaining informati conventional sources like books, maga libraries

4/89

Today's Information Society

People across the world…

- Chat
- Exchange e-mail, sms, pictures
- Buy products and services online
- Use search engines to find information useful in their work and day-to-day life
- Exploit the Web for obtaining information instead of conventional sources like books, magazines and libraries

5/89

Today's Information Society

People across the world…

- Chat
- Exchange e-mail, sms, pictures
- Buy products and services online
- Use search engines to find information useful in their work and day-to-day life
- Exploit the Web for obtaining information instead of conventional sources like books, magazines and libraries

6/89

Today's Information Society

People across the world…

- Chat
- Exchange e-mail, sms, pictures
- Buy products and services online
- Use search engines to find information useful in their work and day-to-day life
- Exploit the Web for obtaining information instead of conventional sources like books, magazines and libraries

7/89

Today's Information Society

People across the world…

- Chat
- Exchange e-mail, sms, pictures
- Buy products and services online
- Use search engines to find informatio their work and day-to-day life
- Exploit the Web for obtaining information instead of conventional sources like books, magazines and libraries

8/89

Today's Information Society

Se si vuole trovare una metafora del rapporto fra l'uomo e i mezzi di comunicazione, Umberto Eco suggerisce quella dell'automobilista: la tecnologia ha messo a disposizione vetture sempre più sofisticate, potenti e veloci; che vengano usate per portare una persona all'ospedale o per fare le gare di velocità sulle strade, dipende da chi è seduto al posto di guida. Lo stesso si può dire di quella che ormai è diventata una delle relazioni fondamentali della nostra vita quotidiana, cioè il nostro modo di interagire con i mass media, dalla televisione al telefonino, da Internet alla radio, dai libri ai cd e dvd (ebbene sì, anch'essi sono media), alla posta elettronica: dipende dalla cultura e dalla volontà di ciascuno di noi, educato soprattutto dalla scuola e dalla famiglia, mettere a punto una "dieta mediatica" – suggerisce Gianfranco Bettetini – che non provochi né obesità né anoressia.

da: "Mettete a dieta i mass-media"

INTERVISTA A DUE VOCI CON UMBERTO ECO E GIANFRANCO BETTETINI

di Paolo Perazzolo, Famiglia Cristiana n.20 del 20-05-2007
(http://www.sanpaolo.org/fc/0720fc/0720fc54.htm)

9/89

Today's Information Society

Problems…

- Explosion of irrelevant, unclear, inaccurate information
- Users overloaded with a large amount of information impossible to absorb

…and consequences

- Searching is time consuming
- Need for intelligent solutions able to support users in finding documents according to their interests

10/89

How Much Information?

1. Print, film, magnetic and optical storage media produced between 3.4 and 5.4 exabytes of unique information in 2002
   - ✓ 92% stored on magnetic media, mostly hard disks
   - ✓ 500-800 MB per person each year
2. Information flows through electronic channels - telephone, radio, TV, and the Internet – contained almost 18 exabytes of new information in 2002, three and a half times more than is recorded in storage media

Medium	Terabytes
Radio	3,488
Television	68,955
Telephone	17,300,000
Internet	532,897
TOTAL	17,905,340

Medium	Terabytes
Surface Web	167
Hidden Web	91,850
E-mail	440,606
Inst. Messaging	274
TOTAL	532,897

Source: Lyman, Peter and Hal R. Varian, "How Much Information", 2003. School of Information Management and Systems, University of California at Berkeley. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003. Last access: May 23rd, 2007

11/89

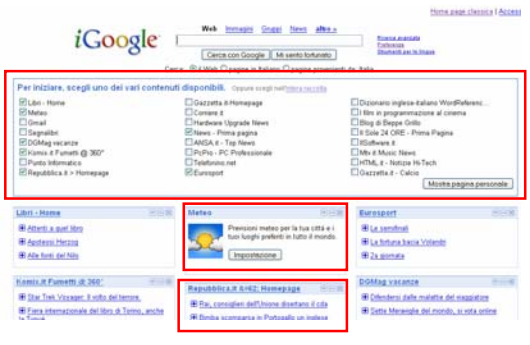
How big is an Exabyte?

Kilobyte (KB)	1,000 bytes or 10^3 bytes. 2 Kilobytes: A Typewritten page. 100 Kilobytes: A low-resolution photograph.
Megabyte (MB)	1,000,000 bytes or 10^6 bytes. 1 Megabyte: A small novel or a 3.5 inch floppy disk. 2 Megabytes: A high-resolution photograph. 5 Megabytes: The complete works of Shakespeare. 10 Megabytes: A minute of high-fidelity sound. 100 Megabytes: 1 meter of shelved books.
Gigabyte (GB)	1,000,000,000 bytes or 10^9 bytes. 1 Gigabyte: a pickup truck filled with books. 20 Gigabytes: A good collection of the works of Beethoven. 100 Gigabytes: A library floor of academic journals.
Terabyte (TB)	1,000,000,000,000 bytes or 10^{12} bytes 1 Terabyte: 50,000 trees made into paper and printed. 2 Terabytes: An academic research library.
Petabyte (PB)	1,000,000,000,000,000 bytes or 10^{15} bytes. 2 Petabytes: All U.S. academic research libraries. 20 Petabytes: Production of hard-disk drives in 1995. 200 Petabytes: All printed material.
Exabyte (EB)	1,000,000,000,000,000,000 bytes or 10^{18} bytes. 2 Exabytes: Total volume of information generated in 1999. 5 Exabytes: All words ever spoken by human beings.

TABLE 1.1: How Big is an Exabyte? Source: Many of these examples were taken from Roy Williams "Data Powers of Ten" web page at Caltech.

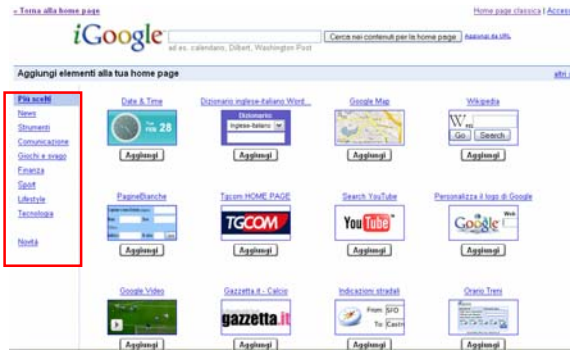
12/89

My...Web: *i*Google 1/2



13/89

My…Web: *i*Google 2/2

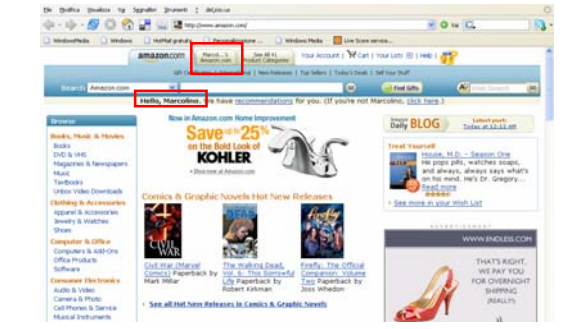


14/89

My… Web: Google News

15/89

My…Web: Personalized Stores 1/2



16/89

My…Web: Personalized Stores 2/2

17/89

Information Access Strategies

Process	Information Need	Information Sources
Information Retrieval	Dynamic-Specific	Stable-Unstructured
Information Filtering	Stable-Specific	Dynamic-Unstructured
Text Mining	Stable-Specific	Stable
Database Access	Dynamic-Specific	Stable-Structured
Exploration	Broad	Varied

Information Retrieval [Baeza-Yates and Ribeiro-Neto 1999]

- "Information Retrieval (IR) deals with the representation, storage, organization of, and access to information items"
- "…the user must first translate this information need into a *query* which can be processed by a search engine (or IR system)".
- "Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the user. The emphasis is on the retrieval of *information* as opposed to the retrieval of *data*".

18/89

Information Access Strategies

Process	Information Need	Information Sources
Information Retrieval	Dynamic-Specific	Stable-Unstructured
Information Filtering	Stable-Specific	Dynamic-Unstructured
Text Mining	Stable-Specific	Stable
Database Access	Dynamic-Specific	Stable-Structured
Exploration	Broad	Varied

Information Filtering [Hanani et al. 2001]

- "The aim of Information Filtering is to expose users to only the information that is relevant to them. Some examples of filtering applications are: filters for search results on the internet,… e-mail filters based on *personal profiles*, … filters for e-commerce applications that address products and promotions to potential customers only…"
- "There are many systems of widely varying philosophies, but all share the goal of automatically directing the most valuable information to users in accordance with their user model…"

19/89

Comparing IR and IF

Common Mechanisms

- **Representation:** Both the user's information need – query or profile – and the document set must be represented for *comparison*
- **Comparison:** String matching? Concept matching? User-User Correlation? Item-Item correlation?
- **Feedback:** To improve the performance of the IR/IF system, a feedback mechanism is usually incorporated.

20/89

Comparing IR and IF

1. Where IR is concerned with the *collection and organization* of texts, IF is concerned with the *distribution* of texts to groups or individuals.
2. Where IR is typically concerned with the selection of texts from a relatively *static database*, IF is mainly concerned with the selection or elimination of texts from a *dynamic datastream*.
3. Where IR is concerned with responding to the user's interaction with texts within a *single information-seeking episode*, IF is concerned with long-term changes over *a series of information-seeking episodes*.

[Belkin and Croft 1992]

21/89

Comparing IR and IF

Parameters	Information Retrieval	Information Filtering
Representation of Information Needs	queries	profiles
Goal	selection of relevant items for query	filtering out irrelevant items or collecting items
Frequency of use	ad hoc use one time user	repetitive use long term users
Type of Users	Not known to the system	"Profiled"
Database	(relatively) static	very large dynamic

Table adapted from [Hanani et al. 2001]

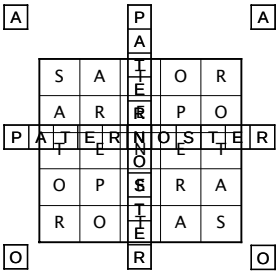
22/89

Some Problems in IF systems…

1. IF systems perform the filtering task on the basis of user profiles
   - ✓ Structured model of the user interests
   - ✓ User profiles compared against item descriptions to provide recommendations
2. Problems: keywords not appropriate for representing content, due to polysemy, synonymy, multi-word concepts (*homography*, *homophony*) – "Sator arepo eccetera" (Eco, 2007)

23/89

Some Problems in IF systems… (cont'd)



24/89

Keyword-based profiles

doc1
AI is a branch of computer science

doc2
the 2007 International Joint Conference on **Artificial Intelligence** will be held in India

doc3
apple launches a new product…

USER PROFILE

artificial	0.02
intelligence	0.01
apple	0.13
AI	0.15
…	

MULTI-WORD CONCEPTS

25/89

Keyword-based profiles

doc1
AI is a branch of computer science

doc2
the 2007 International Joint Conference on **Artificial Intelligence** will be held in India

doc3
apple launches a new product…

USER PROFILE

artificial	0.02
intelligence	0.01
apple	0.13
AI	0.15
…	

SYNONYMY

26/89

Keyword-based profiles

doc1
AI is a branch of computer science

doc2
the 2007 International Joint Conference on **Artificial Intelligence** will be held in India

doc3
apple launches a new product…

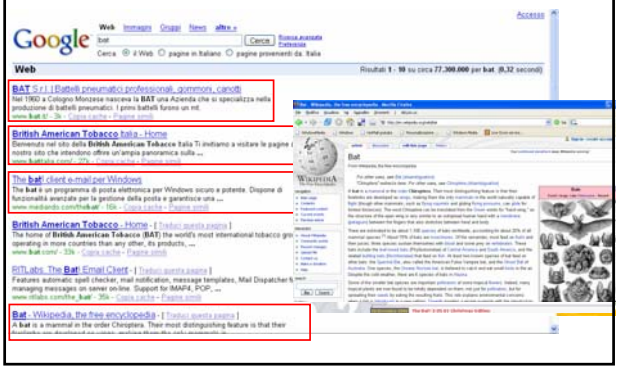
USER PROFILE

artificial	0.02
intelligence	0.01
apple	0.13
AI	0.15
…	

POLYSEMY

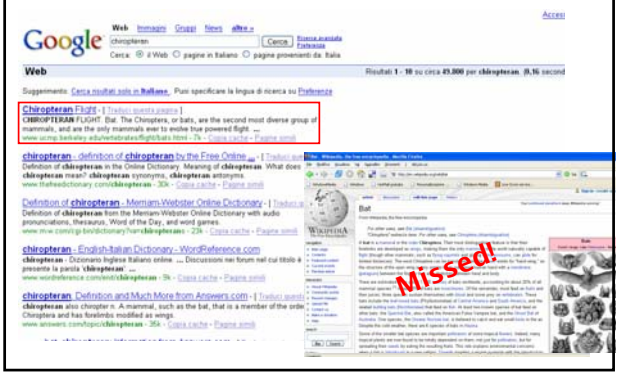
27/89

Some Problems in IR systems… Polysemy



28/89

Some Problems in IR systems… Synonymy



29/89

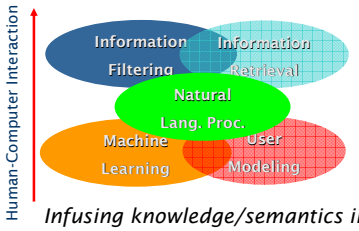
Research Directions…

Intelligent Information Access =

1. Personalized Access by user profiles +
2. Semantic Access by concept identification in documents

30/89

Research Areas

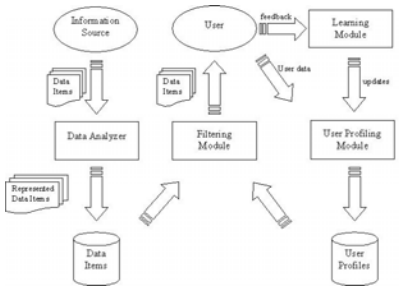


Infusing knowledge/semantics into (traditional) software and services

"technology should adapt to people rather than vice versa"

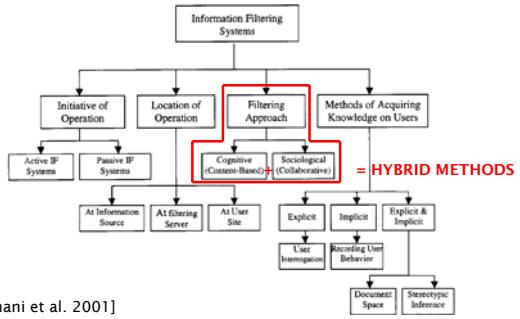
31/89

IF systems architecture



32/89

A classification of IF systems

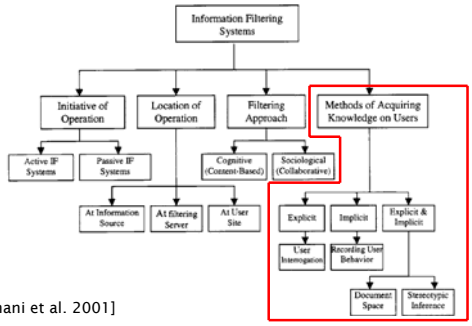


= HYBRID METHODS

[Hanani et al. 2001]

33/89

A classification of IF systems



[Hanani et al. 2001]

34/89

Collaborative / Social Filtering

1. Makes use of a database of user preferences in order to:
   - ✓ find users with *similar* interests
   - ✓ predict whether an *unseen* information item is likely to be of interest for a user based on how *other* users rated that item
2. Widely adopted in recommender systems [Resnick and Varian 1997, Linden et al. 2003]
3. Different implementations
   - ✓ user-to-user
   - ✓ item-to-item

Recommender Systems have the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options [Burke, 2002]

35/89

Collaborative / Social Filtering

1. Makes use of a database of user preferences in order to:
   - ✓ find users with *similar* interests
   - ✓ predict whether an *unseen* information item is likely to be of interest for a user based on how *other* users rated that item
2. Widely adopted in recommender systems [Resnick and Varian 1997, Linden et al. 2003]
3. Different implementations
   - ✓ user-to-user
   - ✓ item-to-item

Recommender Systems provide personalized suggestions about items that the user might find interesting, by matching items to user profiles or groups [Kangas, 2001]

36/89

Collaborative / Social Filtering

1. Makes use of a database of user preferences in order to:
   - ✓ find users with *similar* interests
   - ✓ predict whether an *unseen* information item is likely to be of interest for a user based on how *other* users rated that item
2. Widely adopted in recommender systems [Resnick and Varian 1997, Linden et al. 2003]
3. Different implementations
   - ✓ user-to-user
   - ✓ item-to-item

Recommendation process

Given a large set of items and a description of the user's needs, the goal is to present a small set of the items that are suited to the user needs

37/89

User-to-User CF

Each user represented as an *N*-dimensional vector of *items*

active user

	StarWars	Batman	Harry Potter	Matrix	I Robot	Star Trek
Joe	2	5	4	2	?	?
Howard	5	1	2		1	
Dan	3	2		4		
Rob		3	3			3
Jenny	5		5			
Mary	1	5				1
Mark	2		5		5	

Recommendations based on a few users – *neighbors* – most similar to the active user

38/89

User-to-User CF

active user

	StarWars	Batman	Harry Potter	Matrix	I Robot	Star Trek
Joe	2	5	4	2	5	1
Howard	5	1	2		1	
Dan	3	2		4		
Rob		3	3			3
Jenny	5		5			
Mary	1	5				1
Mark	2		5		5	

Different strategies for computing similarity between users
Cosine similarity and Pearson's correlation coefficient widely used [Herlocker et al. 1999]

Recommendations *selected* from the neighbors using various methods: Items ranked according to how many users liked them

Joe is recommended to see "I robot" and to avoid "Star Trek" based on the suggestions by his *neighbors* Mary and Mark

39/89

Item-to-Item CF

1. Adopted by Amazon.com [Linden et al. 2003]
   - ✓ Amazon.com has more that 30 million customers and several million catalog items
   - ✓ scales to massive datasets and produces high-quality real-time recommendations
2. Similar-items table containing similar items that customers tend to purchase together
3. The algorithm:
   - ✓ finds items similar to each of the active user's purchases and ratings
   - ✓ aggregates those items and recommends the *most popular* or *correlated* items

Item-to-Item CF

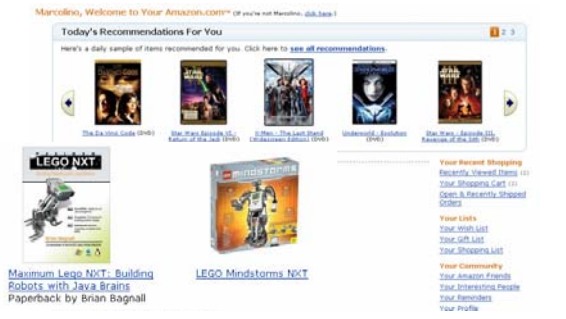
	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9
U_1				X		X			X
U_2	X			X		X			
U_3					X			X	
U_4	X			X			X		
A		X		X					

Shopping cart

Set of recommendation for A: I_4, I_1, I_6 , I_2

Amazon Recommendation System@work 1/3

Amazon Recommendation System@work 2/3



43/89

Amazon Recommendation System@work 3/3

Recomm. based on content similarity

44/89

Amazon Recommendation System@work 3/3

45/89

Traps and Pitfalls

-----Messaggio originale-----
Da: Silvano Petruzzelli
Inviato: giovedì 15 marzo 2007 14.36
A: Gianni Semeraro - Uniba; Gigi Intonti - Software Design; Leo Iaquinta - Uniba; Anna Lisa Gentile - Uniba
Oggetto: Amazon Recommendation System... Che abbaglio!!!! - FW: Save 32% at Amazon.com on "Hidden in Plain Sight: The Secret of More" by Mark Buchanan

Ciao a tutti.
Ho ricevuto il solito messaggio da parte di Amazon che mi suggerisce un libro. Questa volta però sono rimasto un po' spiazzato perché a me, dei libri del pastore Buchanan, sinceramente mi importa poco.
Poi ho capito che questa raccomandazione mi è arrivata perché tra i libri che Amazon sa che ho comprato c'è Nexus: Small Worlds and the Groundbreaking Theory of Networks di Mark Buchanan. Il problema è che questo Mark Buchanan è decisamente un'altra persona, non il pastore.
Un bell'abbaglio per uno dei sistemi di raccomandazione più efficaci e noti sul web!
Chissà, se scrivessi che posseggo una Vespa, forse mi suggerirebbe i libri di Bruno...

S.

Possible solution:

Entity/Identity recognition [Dichev et al. 2007, Bouquet et al. 2007]

Identity on the Web: *WWW 2007 Workshop I3: Identity, Identifiers, Identifications -- Entity-Centric Approaches to Information and Knowledge Management on the Web*, online http://ceur-ws.org/Vol-249.

46/89

Content-based Filtering

1. Each user is assumed to operate independently
2. Items are represented by some features
   - ✓ Movies: actors, directors, plot, …
   - ✓ Music: players, titles, …
3. Filtering based on the comparison btw the *content* of the items and the user preferences as defined in the user profile
4. How to put "intelligence" into CB filtering?
   - ✓ novel strategies to represent items (mostly based on models and NLP operations inherited from IR)
   - ✓ novel strategies to build and represent profiles (mostly based on AI and Machine Learning)

47/89

Word Sense Disambiguation (WSD) (1/2)

1. The different meanings of polysemous words are known as *senses*
2. Only one sense of a polysemous word is used in a specific linguistic context. The context determines the correct sense
3. The process of deciding which sense is used in a specific context is called WSD [Miller, 90]

Approaches to WSD

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

48/89

Word Sense Disambiguation (WSD) (2/2)

1. WordNet: a lexical reference database, inspired by current psycholinguistic theories of human lexical memory
2. English *nouns*, *verbs*, *adverbs* and *adjectives* organized into SYNonym SETs (synset), each one representing an underlying lexical concept
3. Change of text representation from vectors (bag) of words (BOW) into vectors (bag) of synsets (BOS) to avoid polysemy, synonymy, etc.

49/89

Bag of Synsets

Bag of Words

Id doc	Word Form	Occurrence
31	artificial	1
31	intelligence	1
…	…	…
1134	WWW	3
1134	web	2
…	…	…
1135	Java	2
…	…	…

Bag of Synsets

Id doc	Word Form	Id Synset	Occurrence
31	artificial intelligence	05766061	1
…	…		…
1134	WWW,web	04425517	5
…	…	…	…
1135	Java	08357098	1
1135	Java	07452170	1
…	…		…

- ✓ Recognition of bigrams
- ✓ Synonyms represented by the same synsets
- ✓ Polysemous words disambiguated (hopefully)

50/89

WordNet as a sense repository: The Lexical Matrix

WordNet a lexical database for the English language

Synonym word forms: SYNSET

Word Meanings	Word Forms					
	F_1	F_2	F_3	…	…	F_n
M_1	E(1,1)	E(2,1)				
M_2		E(2,2)	E(3,2)			
M_3						
$M_{…}$						
M_m						E(m,n)

Mapping between word forms and word meanings

51/89

WordNet as a sense repository: The Lexical Matrix

WordNet a lexical database for the English language

the word form is polysemous: WSD needed

Word Meanings	Word Forms					
	F_1	F_2	F_3	…	…	F_n
M_1	E(1,1)	E(2,1)				
M_2		E(2,2)	E(3,2)			
M_3						
$M_{…}$						
M_m						E(m,n)

Mapping between word forms and word meanings

52/89

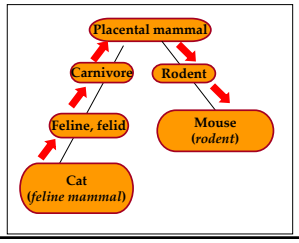
WSD algorithm

1. Input: $D = \langle w_1, w_2, \ldots, w_h \rangle$ document
2. Output: $X = \langle s_1, s_2, \ldots, s_k \rangle$ $(k \leq h)$
   - ✓ Each s_i obtained by disambiguating w_i based on the *context* of each word
   - ✓ Some words not recognized by WordNet
   - ✓ Groups of words recognized as a single concept
   - ✓ UniBA JIGSAW WSD algorithm [Semeraro et al. 2007]

53/89

Example: Noun Disambiguation

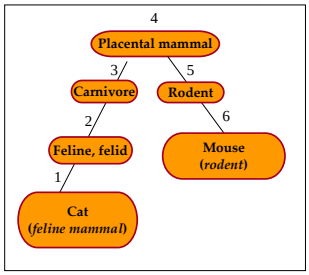
1. Semantic similarity between synsets inversely proportional to their distance in the WordNet IS-A hierarchy
2. Path length similarity between synsets used to assign scores to the candidate synsets of a polysemous word



54/89

Synset Semantic Similarity

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



SINSIM(cat,mouse) =
-log(6/32)=0.727

Leacock-Chodorow similarity [Leacock and Chodorow, 1998]

55/89

Cat-Mouse Disambiguation

"The white cat is hunting the mouse"

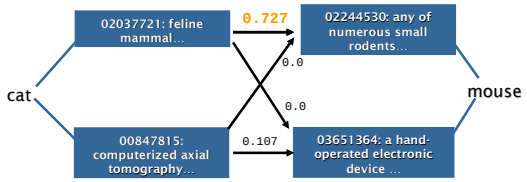
w = cat
C = {mouse}

56/89

Cat-Mouse Disambiguation

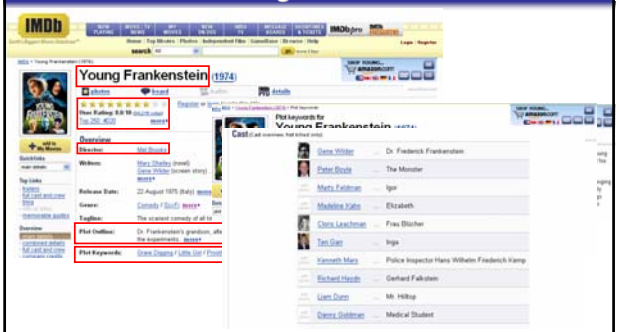
"The white cat is hunting the mouse"

w = cat
C = {mouse}

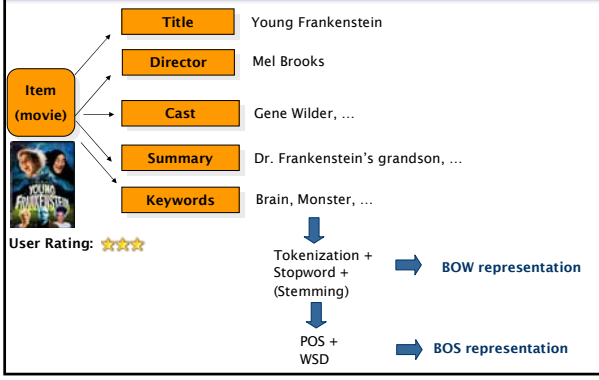


57/89

Movie Recommending (1/2)



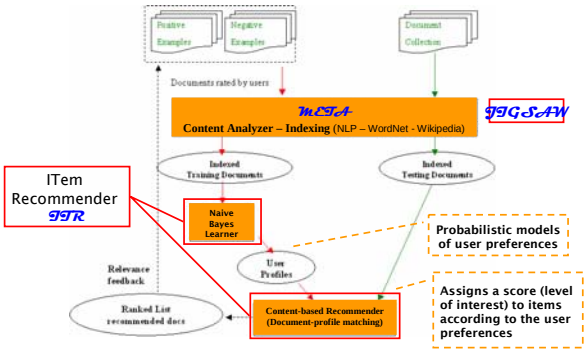
Movie Recommending (2/2)



ITem Recommender (ITR)

1. Advanced NLP techniques used to represent documents
2. Naïve Bayes text classification to assign a score (level of interest) to items according to the user preferences
3. Result: semantic user profile - as a binary text classifier (user-likes and user-dislikes) - containing the probabilistic model of user preferences

ITR: the complete architecture



61/89

Example of Keyword-based User Profile

User ID: 6 Category: dummy Class Priors: P(YES)= 0.5333333 P(NO)= 0.4666666

Slot: abstractContent

Feature	Strength
enhh	3.6138782
reposition	2.8947555
easi	2.8947555
reusabl	2.8947555
upper-level	2.8947555
way	2.8947555
avail	2.8947555
area	2.5893739
kim	2.4626222
seamless	2.1475411
subtasks:	2.1475411

$$strength(t_k, s_m) = \log \frac{P(t_k \mid c_+, s_m)}{P(t_k \mid c_-, s_m)}$$

Features are keywords

62/89

Example of Sense-based User Profile

User ID: 6 Category: dummy Class Priors: P(YES)= 0.5333333 P(NO)= 0.4666666

Slot: abstractContent

Feature	Strength
1742	3.6145387
2268652	2.8954161
2766412	2.8954161
2223910	2.8954161
4415376	2.8954161
5655492	2.5376664
5552847	2.3181007
2311478	2.1482017
5309075	2.1482017
1636312	2.1482017

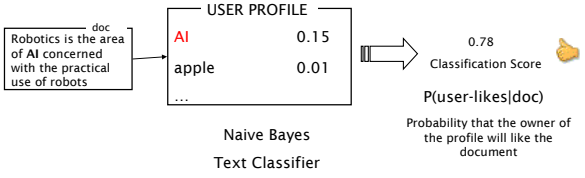
$$strength(t_k, s_m) = \log \frac{P(t_k \mid c_+, s_m)}{P(t_k \mid c_-, s_m)}$$

is computed on synsets instead of keywords

Features are WordNet synsets

63/89

Recommendations based on User Profiles

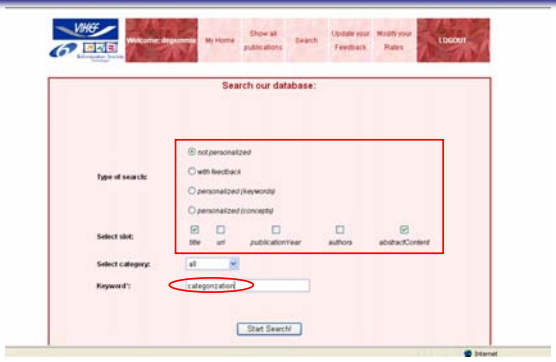


64/89

Conference Participant Advisor: Login

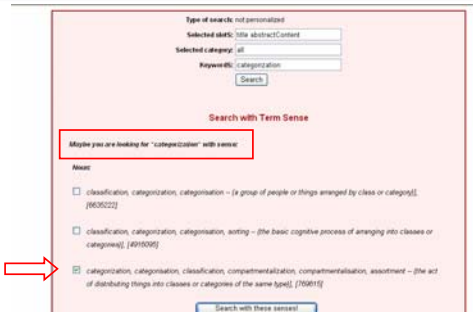
Conference Participant Advisor: Selecting Papers to train the system

65/89



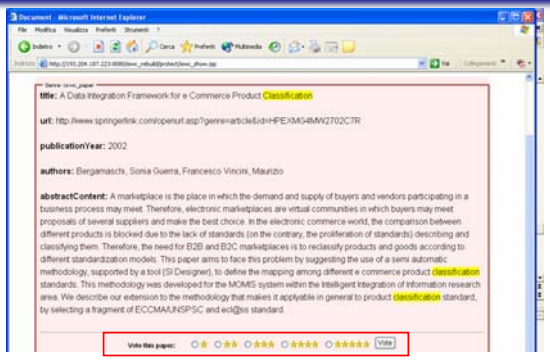
Conference Participant Advisor: Query disambiguation

66/89



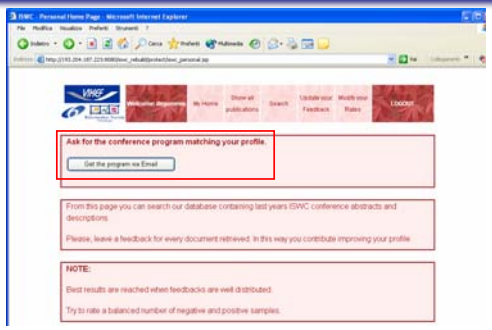
Conference Participant Advisor: Rating Retrieved Papers

67/89



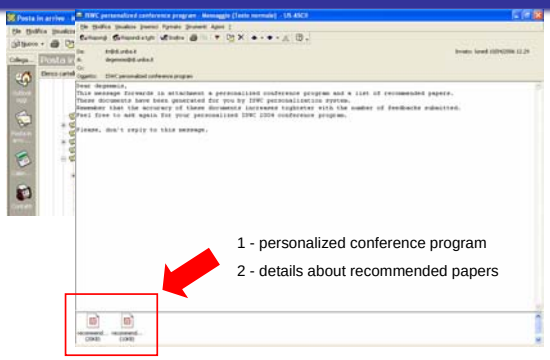
Conference Participant Advisor: Getting the Personalized Program

68/89



Personalized Program delivered by mail

69/89



Conference Participant Advisor: Personalized Program + Paper details

70/89

71/89

Content-based Filtering Systems (1/2)

System	Type	Document Representation	Profiling Strategy
Letizia [Lieberman95]	Web pages recommender	Keywords	Heuristics + implicit feedback
Syskill & Webert [Pazzani et al.96]	Web pages recommender	Keywords	Bayesian Learning + explicit feedback
WebMate[Chen & Sycara98]	Web pages recommender	Keywords	Profiles as TFIDF vectors
NewsDude [Billsus et al.99]	News Recommender	Keywords	Hybrid profile for short-term (TF-IDF) and long-term interests (Bayesian Learning)
PSUN [Sorensen et al. 94]	News Recommender	N-grams	Genetic algorithms + explicit feedback

72/89

Content-based Filtering Systems (2/2)

System	Type	Document Representation	Profiling Strategy
ifWeb [Asnicar & Tasso 96]	Document Searching	Keywords	Weighted semantic network + explicit feedback + separate profiles for interests and disinterests + aging strategy
LIBRA [Mooney & Roy 00]	Book Recommender	Keywords, documents structured into slots	Bayesian learning
siteIF [Magnini et al. 01]	News Recommender	Based on WordNet synsets	Weighted semantic network + implicit feedback + aging strategy

73/89

Hybrid Strategies

1. Try to combine different types of filters in order to overcome drawbacks of single techniques
2. Burke's classification [Burke, 2002]
   - ✓ Weighted – the scores of different recommendation techniques combined together to produce a single recommendation
   - ✓ Switching - use some criteria to switch between recommendation techniques
   - ✓ Mixed - recommendations from different systems presented at the same time
   - ✓ Feature Combination - features from different recommendation sources thrown together into a single recommendation algorithm (e.g., content-based techniques used over a set of augmented data containing collaborative information as simply additional features)
3. UniBA approach (Content-Collaborative) described in [Degemmis et al. 2007]

Intelligent IR: beyond keywords and the "one fits all" approach

74/89

1. Semantic Indexing and Retrieval: Use of lexicons, ontologies and on-line resources
   - ✓ Adaptation of VSM for Ontology-based Retrieval [Castells et al. 2007]
   - ✓ Indexing by WordNet Synsets [Gonzalo et. al 98], [Mihalcea & Moldovan 2000]
   - ✓ Indexing by using shared world knowledge, like Wikipedia [Gabrilovich and Markovitch 2007]
2. New models for personalized retrieval (ranking)
   - ✓ Contextual user profiles for query refinement [Liu et al. 2004]
   - ✓ Re-Ranking (modification of the original ranking) based on user profiles [Semeraro 2007, to appear]

75/89

What is a Semantic Digital Library?

Semantic digital libraries

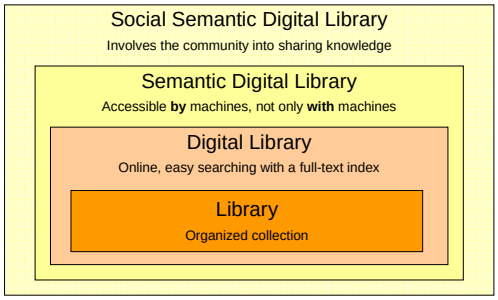
- ✓ integrate information based on different metadata, e.g.: resources, user profiles, bookmarks, taxonomies – high quality semantics = highly and meaningfully connected information

- ✓ provide interoperability with other systems (not only digital libraries) on either metadata or communication level or both – RDF as common denominator between digital libraries and other services
- ✓ delivering more robust, user friendly and adaptable search and browsing interfaces empowered by semantics

from: Sebastian R. Kruk, Stefan Decker, Bernhard Haslhofer, Predrag Kneževic, Sandy Payette, Dean Krafft. *WWW 2007 Tutorial on Semantic Digital Libraries*.

76/89

Evolution of Libraries



from: Sebastian R. Kruk, Stefan Decker, Bernhard Haslhofer, Predrag Kneževic, Sandy Payette, Dean Krafft. *WWW 2007 Tutorial on Semantic Digital Libraries.*

77/89

Benefits of Semantic Digital Libraries

Problems of today's libraries

1. rapidly growing islands of highly organized information
   - How to find things in a growing information space?
     - ✓ is it enough to have a full-text index (à la Google)?
     - ✓ typical "end-users" versus "expert users"
2. converging digital library systems
   - e.g. uniform access to Europe's digital libraries and cultural heritage
   - The European Library
     http://www.theeuropeanlibrary.org

from: Sebastian R. Kruk, Stefan Decker, Bernhard Haslhofer, Predrag Kneževic, Sandy Payette, Dean Krafft. *WWW 2007 Tutorial on Semantic Digital Libraries.*

78/89

Benefits of Semantic Digital Libraries

The two main benefits of Semantic Digital Libraries

1. new search paradigms for the information space
   - Ontology-based search / facet search
   - Community-enabled browsing
2. providing interoperability on the data level
   - integrating metadata from various heterogeneous sources
   - Interconnecting different digital library systems

from: Sebastian R. Kruk, Stefan Decker, Bernhard Haslhofer, Predrag Kneževic, Sandy Payette, Dean Krafft. *WWW 2007 Tutorial on Semantic Digital Libraries.*

79/89

Semantic DL as Evolving Knowledge Space

1. In state-of-the-art digital libraries users are consumers
   - ✓ Retrieve contents based on available bibliographic records
2. Recent trends: user communities
   - ✓ Connetea
   - ✓ Flickr
3. In Semantic digital libraries users are contributers as well
   - ✓ Tagging (Web 2.0)
   - ✓ Social Semantic Collaborative Filtering
   - ✓ Annotations
4. Semantic Digital libraries enforce the transition from a static information to a dynamic (collaborative) knowledge space

from: Sebastian R. Kruk, Stefan Decker, Bernhard Haslhofer, Predrag Kneževic, Sandy Payette, Dean Krafft. *WWW 2007 Tutorial on Semantic Digital Libraries.*

80/89

Existing Semantic Digital Library Systems

1. JeromeDL
   - ✓ a social semantic digital library makes use of Semantic Web and Social Networking technologies to enhance both interoperability and usability
2. BRICKS
   - ✓ aims at establishing the organizational and technological foundations for a digital library network in order to share knowledge and resources in the cultural heritage domain.
3. FEDORA
   - ✓ delivers flexible service-oriented architecture to managing and delivering content in the form of digital objects
4. SIMILE
   - ✓ extends and leverages DSpace, seeking to enhance interoperability among digital assets, schemata, metadata, and services

from: Sebastian R. Kruk, Stefan Decker, Bernhard Haslhofer, Predrag Kneževic, Sandy Payette, Dean Krafft. WWW 2007 Tutorial on Semantic Digital Libraries.

81/89

Concluding remarks

1. Need for *intelligent* solutions and tools for Information Access in the information overload era
2. New strategies for Information Filtering & Retrieval
   - ✓ Personalization & User Profiling for Recommender Systems
   - ✓ Semantics: to capture the meaning of content and user needs
3. The present and the future of Digital Libraries

To get introduced to the interesting world of IR, IF and NLP

- BOOKS
  - R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
  - I. Witten, M. Gori and T. Numerico. *Web Dragons. Inside the myths of Search Engine Technology*, 2007.
  - C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
  - M. Stevenson. *Word Sense Disambiguation*. Center for the Study of Language and Information, 2002.
  - C. D. Manning & H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
  - Pierre Baldi, Paolo Frasconi, Padhraic Smyth. *Modeling the Internet and the Web, Probabilistic Methods and Algorithms*. Wiley, 2003.
  - C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.

To get introduced to the interesting world of IR, IF and NLP

- Schools
  - 6th European Summer School in Information Retrieval (ESSIR 2007)
  - 19th European Summer School in Logic, Language and Information (ESSLLI 2007)
  - AI*IA Winter School of AI and Cultural Heritage (Milano, Feb 2007)

References

[Asnicar & Tasso 96] Asnicar, F. and C. Tasso. ifWeb: a Prototype of User Model-based Intelligent Agent for Documentation Filtering and Navigation in the Word Wide Web. In C. Tasso, A. Jameson, and C. L. Paris (eds.), Proceedings of the First International Workshop on Adaptive Systems and User Modeling on the World Wide Web, Sixth International Conference on User Modeling. Chia Laguna, Sardinia, Italy, pp. 3-12, 1997.

[Baeza-Yates and Ribeiro-Neto 1999] Baeza-Yates, R. and Ribeiro-Neto, B., Modern Information Retrieval. New York: Addison-Wesley, 1999.

[Belkin and Croft 1992] Belkin, N.J. and W.B. Croft. Information Filtering and Information Retrieval: Two Sides of the Same Coin? Communications of the ACM, 35(12): 29-38, 1992.

[Billsus et al.99] Billsus, D. and M.J. Pazzani. A Personal News Agent That Talks, Learns and Explains. Agents 1999, 268-275, 1999.

[Bouquet et al. 2007] Bouquet, P., H. Stoermer, and D. Giacomuzzi. Identity: OKKAM: Enabling a Web of Entities. In Proceedings of the WWW 2007 Workshop i3: Identity, Identifiers, Identification. Banff, Canada, May 8, 2007, CEUR Workshop Proceedings, ISSN 1613-0073, online http://ceur-ws.org/Vol-249/submission_150.pdf.

[Burke 2002] R. Burke. Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction. 12(4):331-370, 2002.

[Castells et al 2007] P. Castells, M. Fernandez and D. Vallet. An Adaptation of the Vector-Space Model for Ontology-based Information Retrieval. IEEE Trans. On Knowledge and Data Engeneering 19 (2), pp. 261-272, 2007.

References

[Chen and Sycara 1998] Chen, L. and K.P. Sycara. WebMate: A Personal Agent for Browsing and Searching. Agents 1998, 132-139, 199

[Degemmis et al. 2007] M. Degemmis, P. Lops & G. Semeraro. A Content-Collaborative Recommender that Exploits WordNet-based User Profiles for Neighborhood Formation. User Modeling and User-Adapted Interaction: The Journal of Personalization Research, Springer Netherlands, 2007, ISSN: 0924-1868 (to appear), ISSN online: 1573-1391 (March 22, 2007).

[Dichev et al. 2007] Dichev, C., D. Dicheva, and J. Fischer. Identity: How to name it, How to find it. In Proceedings of the WWW 2007 Workshop on Identity on the Web: I3: Identity, Identifiers, Identifications -- Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Alberta (Canada), May 8, 2007.

[Hanani et al. 2001] Hanani, U., B. Shapira, and P. Shoval. Information Filtering: Overview of Issues, Research and Systems. User Modeling and User-Adapted Interaction, 11(3): 203-259, 2001.

[Herlocker et al. 1999] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl: 1999, An Algorithmic Framework for Performing Collaborative Filtering. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 230-237.

[Kangas 2001] Sonja Kangas. Collaborative Filtering and Recommendations systems. Technical Report TTE4-2001-35. VTT Information Technology, Espoo, Finland.

[Kruk et al 2007] Sebastian R. Kruk, Stefan Decker, Bernhard Haslhofer, Predrag Knezevic, Sandy Payette, Dean Krafft. WWW 2007 Tutorial on Semantic Digital Libraries.

References

[Leacock and Chodorow 1998] C. Leacock and M. Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification. In C. Fellbaum, editor, WordNet: An Electronic Lexical Database, pages 266–283. MIT Press, 1998.

[Lieberman95] H. Lieberman. Letizia: An Agent That Assists Web Browsing. International Joint Conference on Artificial Intelligence, 924-929, Montreal, August 1995. (http://lcs.www.media.mit.edu/people/lieber/Lieberary/Letizia/Letizia.html)

[Linden et al. 2003] Linden, G., B. Smith, and J. York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Computing, 7(1), 76-80, 2003.

[Linden et. al 2001] G.D. Linden, J.A. Jacobi, and E.A. Benson, Collaborative Recommendations Using Item-to-Item Similarity Mappings, US Patent 6,266,649 (to Amazon.com), Patent and Trademark Office, Washington, D.C., 2001.

[Liu et al. 2004] Liu, F., Yu, C., and Meng, W. 2004. Personalized Web Search For Improving Retrieval Effectiveness. IEEE Transactions on Knowledge and Data Engineering 16, 1 (Jan. 2004), 28-40.

[Magnini et al. 01] Magnini, B. and C. Strapparava. Improving User Modelling with Content-based Techniques. In Proceedings of the Eighth International Conference on User Modeling. Sonthofen, Germany, pp. 74-83, Springer, 2001.

[Mihalcea and Moldovan 2000] Rada Mihalcea and Dan Moldovan. Semantic indexing using WordNet senses. In Proceedings of ACL 2000.

[Miller 1990] G. Miller. WordNet: An On-Line Lexical Database. International Journal of lexicography, 3(4), 1990.

References

[Mooney & Roy 00] Mooney, R. J. and L. Roy. Content-Based Book Recommending Using Learning for Text Categorization. In Proceedings of the Fifth ACM Conference on Digital Libraries. San Antonio, US, pp. 195-204, ACM Press, New York, US, 2000.

[Pazzani et al. 1997] Pazzani M. and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. Machine Learning, 27(3):313–331, 1997.

[Resnick and Varian 1997] Resnick, P. and H. Varian. Recommender Systems. Communications of the ACM, 40(3), 56-58, 1997.

[Sarwar et.] B.M. Sarwar, G. Karypis, J.A. Konstan, J. Riedl. Analysis of Recommendation Algorithms for E-Commerce, ACM Conf. Electronic Commerce, ACM Press, 2000, pp.158-167.

[Semeraro 2007] Personalized Searching by Learning WordNet-based User Profiles. Journal of Digital Information Management. Special Issue on Web Information Retrieval. 2007 (to appear)

[Semeraro et al. 2007] G. Semeraro, M. Degemmis, P. Lops, and P. Basile. Combining Learning and Word Sense Disambiguation for Intelligent User Profiling. In Proc. of the 20th Int. Joint Conf. on Artificial Intelligence, 2007, Hyderabad, India, pages 2856-2861, 2007.

[Sorensen et al. 95] Sorensen H. and M. McElligot. PSUN: A Profiling System for Usenet News. CKIM '95 Workshop on Intelligent Information Agents, Baltimore, Maryland, December 1995. (http://odyssey.ucc.ie/filtering/psun.ps)