UNIVERSITÀ DEGLI STUDI DI BARI
Facoltà di Scienze Matematiche, Fisiche e Naturali
Dipartimento di Informatica

Personalization: Learning User Profiles

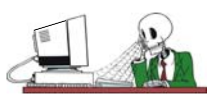
Prof. Giovanni Semeraro
Dott. Pasquale Lops
Dott. Marco Degemmis

Corso di Gestione della Conoscenza d'Impresa
Anno Accademico 2005-2006

The general problem…

- Huge number of Web sites and volume of on-line data *(information overloading)*
  - Users overloaded with a large amount of information
  - Difficulty in finding relevant documents

- Consequence: searching may be time consuming!
  - Demand for automated user support
  - Need for intelligent solutions able to support users in finding documents according to their interests



...and problems in e-commerce

- Critical aspect in e-commerce
  - Millions of products for sale
  - Customers overloaded with a large amount of product information
  - Searching may be time consuming!

- Need for personalized solutions able to support customers in retrieving relevant products

User interests useful to achieve personalization

User Profiling

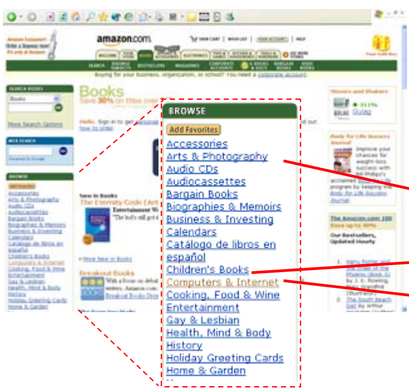
Personalization: User Profiles

USER PROFILE: RAPPRESENTAZIONE STRUTTURATA DEGLI INTERESSI E DELLE PREFERENZE DELL'UTENTE

- Questionnaire-based personalization
  - La definizione manuale di profili è un processo noioso per gli utenti
  - Gli utenti non aggiornano i loro profili
  - La personalizzazione potrebbe basarsi su dati non affidabili

Induzione di modelli di preferenze:
SUPERVISED MACHINE LEARNING

User Preferences: categories



Preferences

Arts & Photography
Children's books
Computers & Internet

User Preferences: items

Preferences

Arts & Photography
Children's books
Computers & Internet

content-based recommendations by learning from TEXT

Book description at Amazon.com

The system

ITem Recommender (ITR)

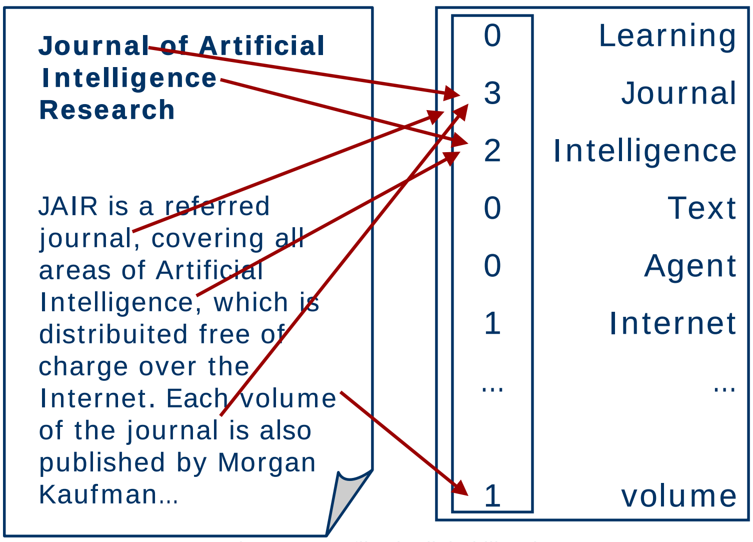
- Content-based item recommending on the basis of rates given by users
- Naïve Bayes text classification to assign a score (level of interest) to items according to the user preferences
  - ✓ Performance comparable to more complex algorithms
  - ✓ Increasing popularity in text classification
- Result: user profile containing the probabilistic model (words + frequencies) learned by the classifier

Document Representation

Molti metodi di text learning usano la reppresentazione *bag-of-words*

- Valore booleano che indica la presenza di una parola
- Frequenza della parola all'interno del documento
- Informazione aggiuntiva (word position, n-grams,...)

Document Representation: bag of words



Bag of words and slots

Journal of Artificial Intelligence Research

JAIR is a referred journal, covering all areas of Artificial Intelligence, which is distribuited free of charge over the Internet. Each volume of the journal is also published by Morgan Kaufman...

1	Journal
1	Intelligence
1	Artificial
1	Research

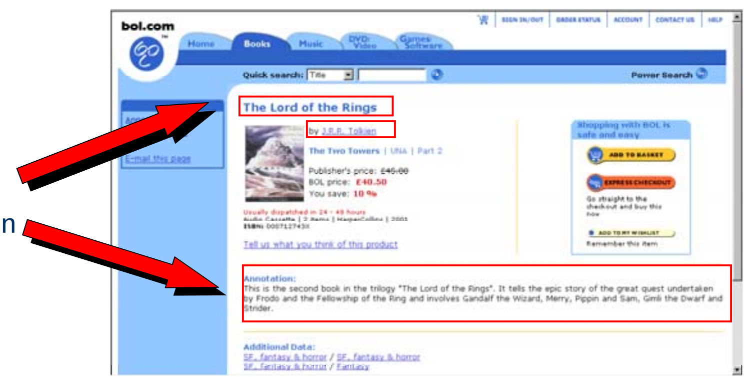
Slot "title"

2	Journal
1	Intelligence
1	Artificial
...	...

Slot "abstract"

Book Information at uk.bol.com

Book description



Utilizzo di tecniche basate sul contenuto: applicazione di metodi di text categorization alla descrizione testuale dei libri

Instance description

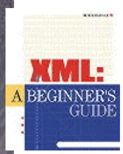
Book recommending on the Web

User Rates: 1-10

Instance (book)

- Title → {lord:1, ring:1}
- Authors → {j_tolkien:1}
- Annotation → {frodo:1, lord:1, ring:2,…}

BOW Extraction



{begin: 1, xml: 2, guid: 1, buzzword: 1, internet: 1, matur: 1, technolog: 1, power: 1, real: 1, world: 1, applic: 1, manag: 1, displa: 1, organ: 1 …}

BAG OF WORDS

TEXTUAL ANNOTATION

tokenization → Stopwords elimination → Stemming → Word Count

Bayes' Theorem

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

$$h \in H$$

D: training data

h: hypothesis from the space H

Bayesian Learning

The *learner* considers a set H of candidate hypotheses and tries to find the most probable hypothesis by taking into account the observed data D

Maximum A Posteriori (MAP) hypothesis

$$\begin{aligned} h_{MAP} &\equiv \operatorname{argmax} P(h \mid D) \\ &= \operatorname{argmax} \frac{P(D \mid h)P(h)}{P(D)} \\ &= \operatorname{argmax} P(D \mid h)P(h) \end{aligned}$$

$$h \in H$$

Bayesian Learning

TEXT CLASSIFICATION

H={“user-likes”, “user-dislikes”}

Each instance (document) is represented by n attributes (keywords) $\{a_1, a_2, \ldots, a_n\}$

$$h_{MAP} = \operatorname{argmax} P(h_j \mid a_1, a_2, \ldots, a_n)$$

$$h_{MAP} = \operatorname{argmax} \frac{P(a_1, a_2, \ldots, a_n \mid h_j) P(h_j)}{P(a_1, a_2, \ldots, a_n)}$$

$$h_{MAP} = \operatorname{argmax} P(a_1, a_2, \ldots, a_n \mid h_j) P(h_j)$$

$$h_j \in H$$

Bayesian Learning

Indipendence assumption

Naive Bayes Classifier :

$$h_{MAP} = \operatorname{argmax} \boxed{P(h_j)} \prod_{i=1}^{n} \boxed{P(a_i \mid h_j)}$$

$$h_j \in H$$

The probabilities are estimated from data

Classification Phase

- Ogni libro è un vettore di bag of words (BOW), una BOW per ogni slot
- Ogni slot è indipendente dagli altri

Title: {lord :1, ring :1}

Authors: {j_tolkien :1}

Annotation: {epic :4, novel :2, lord :2, ring :3, elf :2,…}

Le probabilità a posteriori per un libro d_i sono così calcolate:

$$P(c_j \mid d_i) = \frac{\boxed{P(c_j)}}{P(d_i)} \prod_{m=1}^{|S|} \prod_{k=1}^{|b_{im}|} \boxed{P(t_k \mid c_j, s_m)^{n_{kim}}}$$

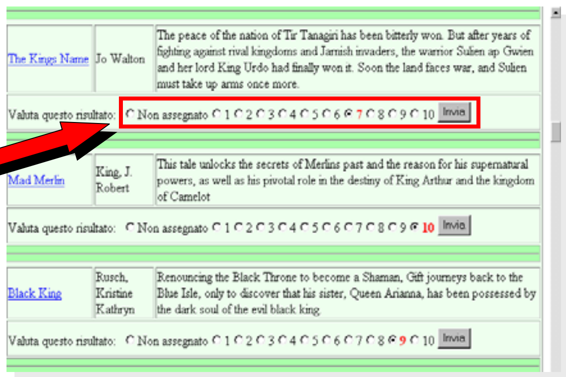
$S = \{s_1, s_2, \ldots, s_{|S|}\}$ è il set degli slots

b_{im} è la BOW nello slot s_m dell'istanza d_i

t_k è la k^{ma} word (che occorre n_{kim} volte nella BOW b_{im})

Acquisizione voti

Ogni utente giudica un set di libri di training in accordo con i suoi gusti, esprimendo un voto da 1 a 10



Sulla base dei giudizi dell'utente, il sistema apprende le sue preferenze su un particolare tipo di documenti in merito al topic di ricerca

Training

- $C = \{c_+, c_-\}$
  - c_+ likes rates 6-10
  - c_- dislikes rates 1-5

User rates r_i ➔ Weighted Instances

$$\boxed{\omega_+^i = \frac{r_i - 1}{9}} \quad \boxed{\omega_-^i = 1 - \omega_+^i}$$

$$\hat{P}(c_j) = \frac{\sum_{i=1}^{|TR|} \omega_j^i}{|TR|} \qquad \hat{P}(t_k \mid c_j, s_m) = \frac{O(t_k, c_j, s_m)}{L(c_j, s_m)}$$

training set cardinality

$$O(t_k, c_j, s_m) = \sum_{i=1}^{|TR|} \omega_j^i n_{kim} \qquad L(c_j, s_m) = \sum_{i=1}^{|TR|} \omega_j^i |b_{im}|$$

Example of ITR Profile

Profile for User: 117
Category: "Computer & Internet"

User ID: 117
Category: Computer & Internet
Class Priors: P(YES)= 0.6169947952025346 P(NO)= 0.3830052047974654

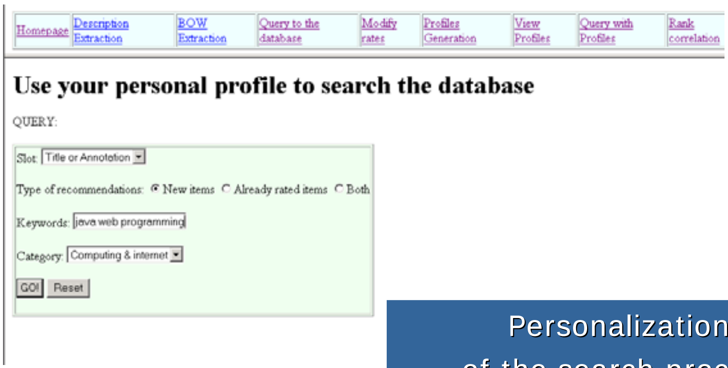
Slot: title

Feature	Strength
gam	2.4155710451915797
directx	2.2707400973131238
enterpris	1.6517008889069003
edit	1.5504909869753871
gem	1.4822827369488536
visu	1.4822827369488536
e-commerc	1.4822827369488534
...	
th	-2.303970881190259
sympos	-2.303970881190259
iee	-2.4741920310501264

The terms of the profile are ranked according to a *strength* measuring the discriminatory power of a word in classifying a book

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

Personalized Searching 1/2



Personalization
of the search process
using ITR PROFILES

Personalized Searching 2/2

Result set ranked by the classification value $P(c_+|d_i)$ computed by ITR

ITR & VIKEF

ITem Recommender (ITR) currently integrated in the *Personalization* component for *context-aware access*, *dissemination*, *visualization*, *knowledge sharing* and *collaboration* (WP6) within the 6FP Project IST-2003-507173 VIKEF

Virtual Information and Knowledge Environment Framework

Learning Preferences of Users Accessing Digital Libraries

O. Licchelli*, P. Lops*, G. Semeraro*, L. Bordoni** and F. Poggi**

(*) Dipartimento di Informatica, Università di Bari, Italy
(**) Enea/Uda-Adivisor, Roma, Italy

Overview

- ✓ Introduction
- ✓ User Profiling
- ✓ The COVAX project
- ✓ The Profile Extractor system in the COVAX architecture
  - ✓Extraction of User Profiles: the learning process
  - ✓Rules generation process
  - ✓Classification process
- ✓ Improve searching in COVAX

Introduction

- The World Wide Web has the potential to be the largest information source for Digital Libraries
- Limit of the Web: information overload problem
- Possible solution: personalization on the basis of the preferences of individual users

User Profiling

- *Personalization* within a Digital Library can occur along several directions:
  - *Explicit and implicit profiling*
  - *Static and dynamic profiling*
  - *Personal annotations*
  - *Person-dependent system behaviour*

The goal of the COVAX project

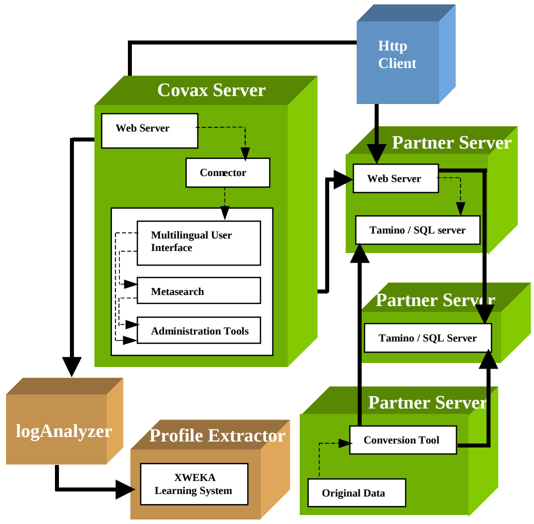
- **COVAX** (5FP IST-1999-11820) - **Contemporary Culture Virtual Archives in XML** - is a global system for search and retrieval, that increases accessibility via Internet to electronic resources regardless of their location.

HOW?

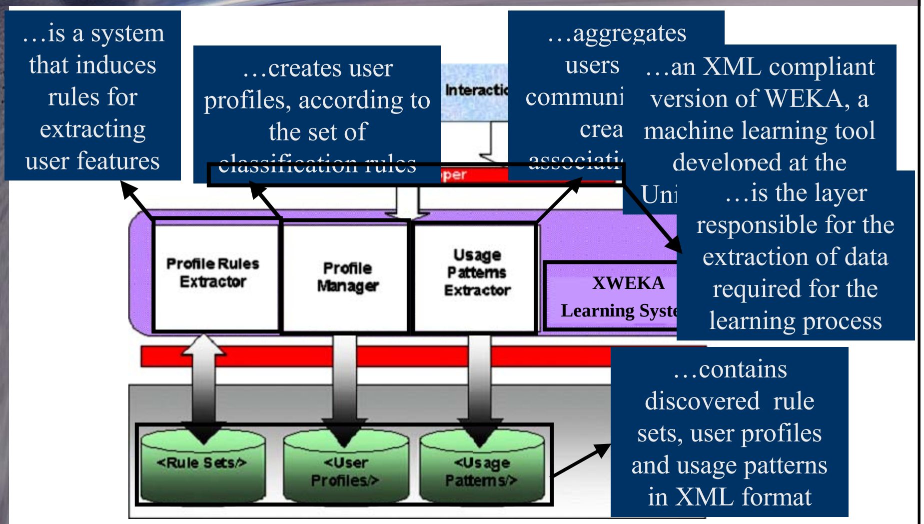
www.covax.org

- ✓ Using Web services for search and retrieval of documents
- ✓ Making accessible over the Internet existing documents in libraries, archives and museums
- ✓ Implementing standards to guarantee interoperability among systems

The Profile Extractor system in the COVAX architecture



The Profile Extractor architecture



Extraction of User Profiles: a ML process

Input: user instances classified in categories on the basis of preferences, abilities and goals.

Method: supervised learning techniques (decision trees, PART) provided by XWEKA.

Output: classification rules.

Profile Extractor: Input

Personal Data

- User_ID
- User_Type (citizen, researcher, scholar)

Interaction Data

- Number of Searches
  ∀ Collection and Database

Collection types:

Bibliographic (libraries)

Archive

Museum

Electronic_Text

Further personal data are not used for privacy problems

Databases:

Spanish libraries

ENEA scientific bibliographic

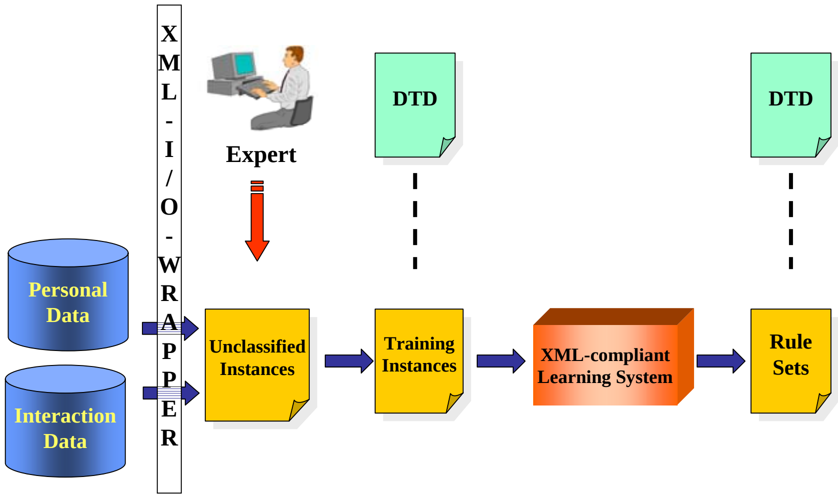
Naval Museum Computerized Library Museum Fronfeste Leather

Prix Multimedia Austria

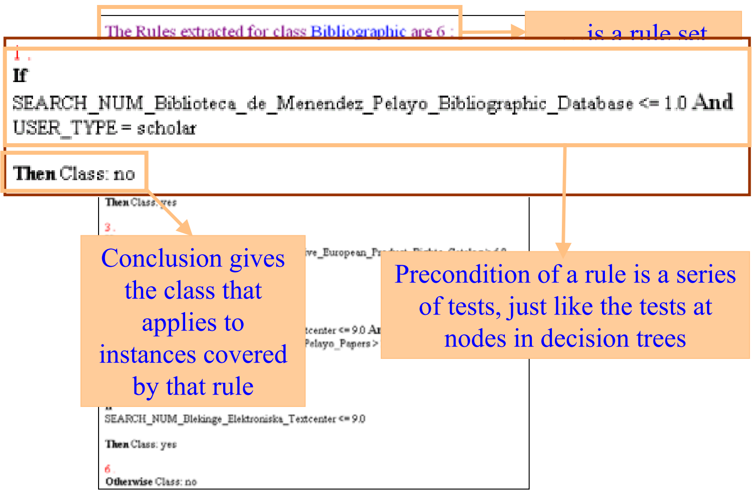
Blekinge Elektroniska Textcenter

etc…

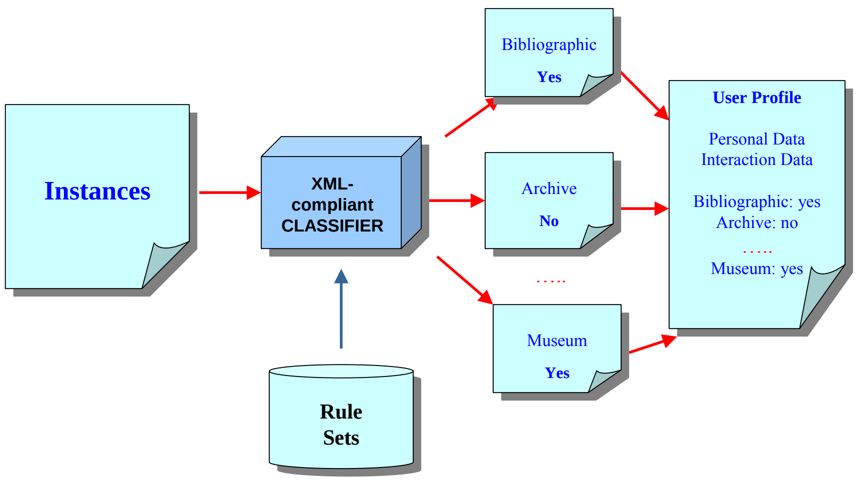
The classification rule generation process



Classification rules: an example

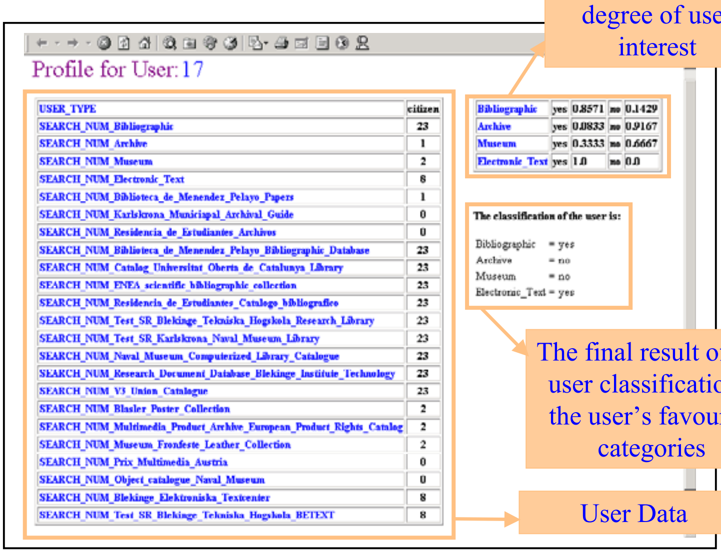


Classification process



User Profile: an example

The list of the 4 collections and the degree of user interest



Profile for User: 17

USER_TYPE	citizen
SEARCH_NUM_Bibliographic	23
SEARCH_NUM_Archive	1
SEARCH_NUM_Museum	2
SEARCH_NUM_Electronic_Text	8
SEARCH_NUM_Biblioteca_de_Menendez_Pelayo_Papers	1
SEARCH_NUM_Karlskrona_Municiapal_Archival_Guide	0
SEARCH_NUM_Residencia_de_Estudiantes_Archivos	0
SEARCH_NUM_Biblioteca_de_Menendez_Pelayo_Bibliographic_Database	23
SEARCH_NUM_Catalog_Universitat_Oberta_de_Catalunya_Library	23
SEARCH_NUM_ENEA_scientific_bibliographic_collection	23
SEARCH_NUM_Residencia_de_Estudiantes_Catalogo_bibliografico	23
SEARCH_NUM_Test_SR_Blekinge_Tekniska_Hogskola_Research_Library	23
SEARCH_NUM_Test_SR_Karlskrona_Naval_Museum_Library	23
SEARCH_NUM_Naval_Museum_Computerized_Library_Catalogue	23
SEARCH_NUM_Research_Document_Database_Blekinge_Institute_Technology	23
SEARCH_NUM_V3_Union_Catalogue	23
SEARCH_NUM_Blasler_Poster_Collection	2
SEARCH_NUM_Multimedia_Product_Archive_European_Product_Rights_Catalog	2
SEARCH_NUM_Museum_Fronfeste_Leather_Collection	2
SEARCH_NUM_Prix_Multimedia_Austria	0
SEARCH_NUM_Object_catalogue_Naval_Museum	0
SEARCH_NUM_Blekinge_Elektroniska_Textcenter	8
SEARCH_NUM_Test_SR_Blekinge_Tekniska_Hogskola_BETEXT	8

Bibliographic	yes	0.8571	no	0.1429
Archive	yes	0.0833	no	0.9167
Museum	yes	0.3333	no	0.6667
Electronic_Text	yes	1.0	no	0.0

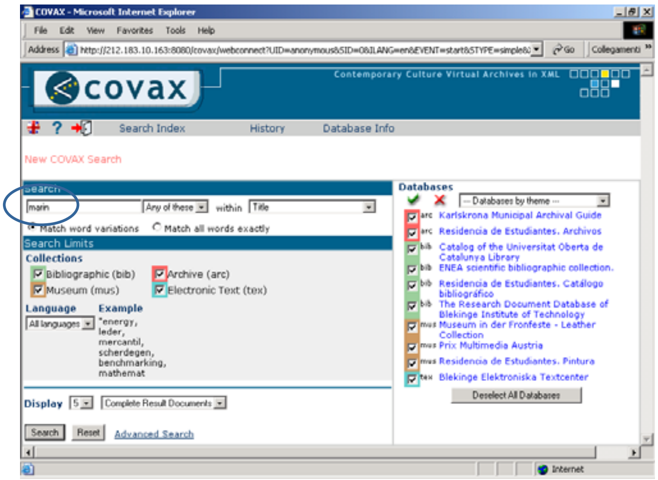
The classification of the user is:

Bibliographic = yes
Archive = no
Museum = no
Electronic_Text = yes

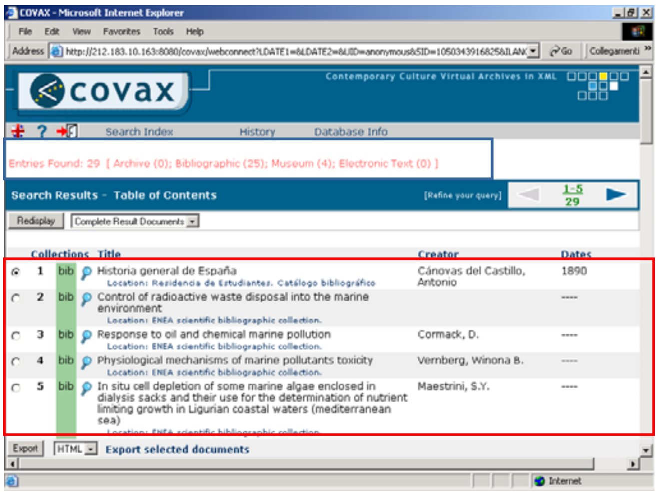
The final result of the user classification - the user's favourite categories

User Data

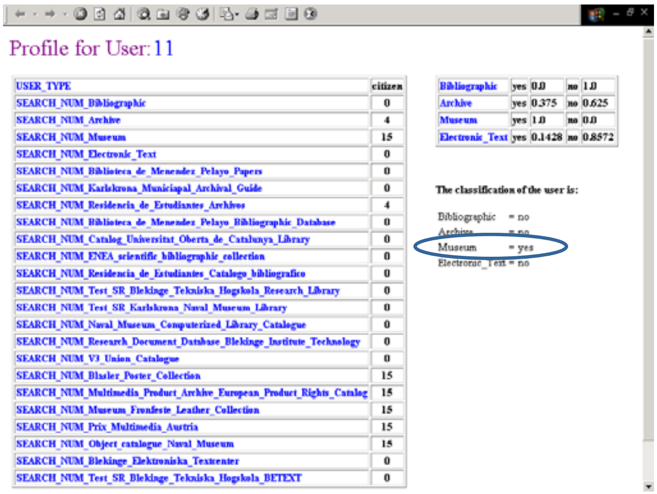
How to improve searching in COVAX



Scenario 1: user profile not available



Scenario 2 – user profile available



Profile for User: 11

USER_TYPE	citizen
SEARCH_NUM_Bibliographic	0
SEARCH_NUM_Archive	4
SEARCH_NUM_Museum	15
SEARCH_NUM_Electronic_Text	0
SEARCH_NUM_Biblioteca_de_Menendez_Pelayo_Papers	0
SEARCH_NUM_Karlskrona_Municiapal_Archival_Guide	0
SEARCH_NUM_Residencia_de_Estudiantes_Archivos	4
SEARCH_NUM_Biblioteca_de_Menendez_Pelayo_Bibliographic_Database	0
SEARCH_NUM_Catalog_Universitat_Oberta_de_Catalunya_Library	0
SEARCH_NUM_ENEA_scientific_bibliographic_collection	0
SEARCH_NUM_Residencia_de_Estudiantes_Catalogo_bibliografico	0
SEARCH_NUM_Test_SR_Blekinge_Tekniska_Hogskola_Research_Library	0
SEARCH_NUM_Test_SR_Karlskrona_Naval_Museum_Library	0
SEARCH_NUM_Naval_Museum_Computerized_Library_Catalogue	0
SEARCH_NUM_Research_Document_Database_Blekinge_Institute_Technology	0
SEARCH_NUM_V3_Union_Catalogue	0
SEARCH_NUM_Blasler_Poster_Collection	15
SEARCH_NUM_Multimedia_Product_Archive_European_Product_Rights_Catalog	15
SEARCH_NUM_Museum_Fronfeste_Leather_Collection	15
SEARCH_NUM_Prix_Multimedia_Austria	15
SEARCH_NUM_Object_catalogue_Naval_Museum	15
SEARCH_NUM_Blekinge_Elektroniska_Textcenter	0
SEARCH_NUM_Test_SR_Blekinge_Tekniska_Hogskola_BETEXT	0

Bibliographic	yes	0.0	no	1.0
Archive	yes	0.375	no	0.625
Museum	yes	1.0	no	0.0
Electronic_Text	yes	0.1428	no	0.8572

The classification of the user is:

Bibliographic = no
Archive = no
Museum = yes
Electronic_Text = no

References

F. Abbattista, L. Bordoni & G. Semeraro. Artificial Intelligence for Cultural Heritage and Digital Libraries. *Applied Artificial Intelligence: An International Journal*, 17(8-9):681-686, Taylor & Francis, September-October 2003.

Balabanovic, M. and Shoham, Y. (1997) *Fab: Content-Based, Collaborative recommendation*. Communications of ACM, 40 (3), 66-72.

L. Bordoni, F. Esposito, O. Licchelli, F. Poggi & G. Semeraro. Personalized Search for Digital Library Users. In L. Bordoni and F. Poggi (Eds.), *Proc. of the AI*IA 2003 Workshop on Artificial Intelligence for the Cultural Heritage*, 104-109, Pisa, Italy, September 23, 2003.

Breese, J.S., Heckerman, D., Kadie, C. (1998) *Empirical Analysis of Predictive Algorithms for Collaborative Filtering.* In Cooper, G. F., Moral, S. (Eds.): UAI '98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, July 24-26, 1998, University of Wisconsin Business School, Madison, Wisconsin, USA. 43-52. Morgan Kaufmann.

Buono, P., Costabile, M. F., Guida, S., Piccinno, A. (2002) *Integrating User Data and Collaborative Filtering in a Web Recommendation Systems*. In Reich, S., Tzagarakis, M.M. and Debra, P.M.E. (Eds.), Hypermedia: Openness, Structural Awareness and Adaptivity, Lecture Notes in Computer Science, vol. 2266. 315-321. Springer, Berlin.

Burke, R. (2002) *Hybrid Recommender Systems: Survey and Experiments*. User Modeling and User-Adapted Interaction, Kluwer Academic Publishers, the Netherlands, 12, 331-370.

M. Degemmis, P. Lops, G. Semeraro & F. Abbattista. Extraction of User Profiles by Discovering Preferences through Machine Learning. In M.A. Kłopotek, S.T. Wierzchoń and K. Trojanowski (Eds.), *Intelligent Information Processing and Web Mining: Proceedings of the International IIS IIPWM´03 Conference*, Advances in Soft Computing, 69-78, Springer:Berlin, 2003.

M. Degemmis, P. Lops, G. Semeraro, M.F. Costabile, S.P. Guida & O. Licchelli. Improving Collaborative Recommender Systems by means of User Profiles. In C.-M. Karat, J. Blom, and J. Karat (Eds.), *Designing Personalized User Experiences in eCommerce*, Chapter 14, 253-274, Kluwer Academic, 2004.

M. Degemmis, P. Lops, G. Semeraro, M.F. Costabile, O. Licchelli & S.P. Guida. A Hybrid Collaborative Recommender System based on User Profiles. *Proceedings of the 6th International Conference on Enterprise Information Systems ICEIS 2004*, Volume 4: Software Agents and Internet Computing, 162-169, Porto, Portugal, 2004.

Frank, E. and Witten, I. (1998) *Generating accurate rule sets without global optimization*. In Shavlik, J. W. (Ed.): Proceedings of the Fifteenth International Conference on Machine Learning, July 24-27, 1998, Madison, Wisconson, USA. 144-151. Morgan Kaufmann.

References (cont'd)

Konstan, J. A., Riedl, J., Borchers, A. and Herlocker, J. L. (1998) *Recommender Systems: A GroupLens Perspective*. In Recommender Systems. Papers from 1998 Workshop. Technical Report WS-98-08. 60-64. AAAI Press.

O. Licchelli, P. Lops, G. Semeraro, L. Bordoni & F. Poggi. Learning Preferences of Users Accessing Digital Libraries. In J. Cha, R. Jardim-Gonçalves, and A. Steiger-Garção (Eds.), *Concurrent Engineering – The Vision for the Future Generation in Research and Applications, Advanced Design, Production and Management Systems, Proceedings of the 10th ISPE International Conference on Concurrent Engineering: Research and Applications*, 457-465, A.A. Balkema Publishers, Madeira Island, Portugal, 2003.

O. Licchelli, F. Esposito, G. Semeraro & L. Bordoni. Personalization to Improve Searching in a Digital Library. In P. Isaías, F. Sedes, J.C. Augusto and U. Ultes-Nitsche (Eds.), *New Technologies for Information Systems, Proceedings of the 3rd International Workshop on New Developments in Digital Libraries*, in conj. with 5th International Conference on Enterprise Information Systems ICEIS 2003, 46-55, Angers, France, April 22, 2003.

Oard, D., and Kim, J. (1998) *Implicit Feedback for Recommender Systems*. In Kautz, H. (Ed.), Recommender Systems. In Proceedings of the AAAI Workshop on Recommender Systems, July 1998. AAAI Press.

Pazzani, M. (1999) *A Framework for Collaborative, Content-Based and Demographic Filtering*. Artificial Intelligence Review, 13 (5-6), 393-408.

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

Ricci, F. and Del Missier, F. (2004) *Supporting Travel Decision Making Through Personalized Recommendation* In Karat, C.M., Blom, J., and Karat, J. (Eds.), Designing Personalized User Experiences for eCommerce, Kluwer.

Sarwar, B. M., Karypis, G., Konstan, J. and Riedl, J. (2002) *Recommender Systems for Large-Scale E-Commerce: Scalable Neighborhood Formation Using Clustering*. In Proceedings of the Fifth International Conference on Computer and Information Technology, December 2002, East West University, Bangladesh.

Schafer, J. B., Konstan, J. and Riedl, J. (1999) *Recommender Systems in E-Commerce.* In EC '99: Proceedings of the First ACM Conference on Electronic Commerce, November 3-5, 1999, Denver, CO. 158-166.

Sebastiani, F. (2002) *Machine Learning in Automated Text Categorization*. ACM Computing Surveys, 34 (1), 1–47.

G. Semeraro, M. Degemmis & P. Lops. User Profiling to Support Internet Customers: what do you want to buy today?. *Informatica: An International Journal of Computing and Informatics*, 26(4):407-418, Informatika Society, 2002.

References (cont'd)

G. Semeraro, P. Lops & M. Degemmis. Learning Probabilistic User Profiles from Labelled and Unlabelled Data. *Informatica: An International Journal of Computing and Informatics*, 27(4):385-389, Informatika Society, 2003.

G. Semeraro, M. Degemmis, P. Lops, U. Thiel & M. L'Abbate. A Personalized Information Search Process based on Dialoguing Agents and User Profiling. In F. Sebastiani (Ed.), *ECIR-03: Proceedings of the 25th European Conference on Information Retrieval Research*, Lecture Notes in Computer Science 2633, 613-621, Springer:Berlin, 2003.

G. Semeraro, F. Abbattista, M. Degemmis, O. Licchelli, P. Lops, & F. Zambetta. Agents, Personalisation, and Intelligent Applications. In R. Corchuelo, A. Ruiz Cortés and R. Wrembel (Eds.), *Technologies Supporting Business Solutions*, Part IV: Data Analysis and Knowledge Discovery, Chapter 7, 141-160, Nova Sciences Publishers, 2003.

G. Semeraro, M. Degemmis & P. Lops. Exploiting supervised learning techniques to model user preferences: personalized e-salespersons for personalized e-stores. In C. Stephanidis (Ed.), *Universal Access in HCI: Inclusive Design in the Information Society, Proceedings of the 2nd International Conference on Universal Access in Human-Computer Interaction UAHCI 2003*, 617-621, Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey, 2003.

G. Semeraro, P. Lops & M. Degemmis. People I Know. In K. Tochtermann and H. Maurer (Eds.), *Journal of Universal Computer Science, Proc. of the 3rd Int. Conf. on Knowledge Management I-KNOW '03*, 463-471, Graz, Austria, July 2-4, Springer:Berlin, 2003.

Shardanand, U., Maes, P., 1995. *Social Information Filtering: Algorithms for automating "word of mouth"*. In Proceedings of the Conference on Human Factors in Computing Systems, May 7-11. 210-217, Denver, CO. ACM Press.

Terveen, L. and Hill, W. (2001) *Human-Computer Collaboration in Recommender Systems*. In J. Carroll (Ed.), Human Computer Interaction in the New Millenium. 223-242. Addison-Wesley, New York.

Tuzhilin, A. and Adomavicius, G. (1999) *Integrating user behaviour and collaborative methods in recommender systems*. In CHI' 99 Workshop on Interacting with Recommender Systems, May 1999, Pittsburgh, PA, USA.

Witten, I. H. and Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann Publishers, San Francisco.

Yao, Y. Y. (1995) *Measuring Retrieval Effectiveness Based on User Preference of Documents*. In Journal of the American Society for Information Science, 46 (2), 133–145.

Contact

Giovanni Semeraro
Dipartimento di Informatica
Università degli Studi di Bari
Via E. Orabona, 4
I-70126 Bari

Tel/Fax: (+39) 080 5442140
Email: semeraro@di.uniba.it
Web:
http://lacam.di.uniba.it:8000/people/semeraro.htm