PISA

A **P**ersonalized ***I***nformation **S**earch **A**ssistant

Maria Elena Renda

Scuola Superiore
Sant'Anna
di Studi Universitari e di Perfezionamento

Outline

- Introduction
  - Search Scenario
  - Personalization
  - Our Approach
- P *I* S A
  - System Functionality
  - Architecture
  - Prototype & Demo
- Conclusions and Future Work

Information Retrieval

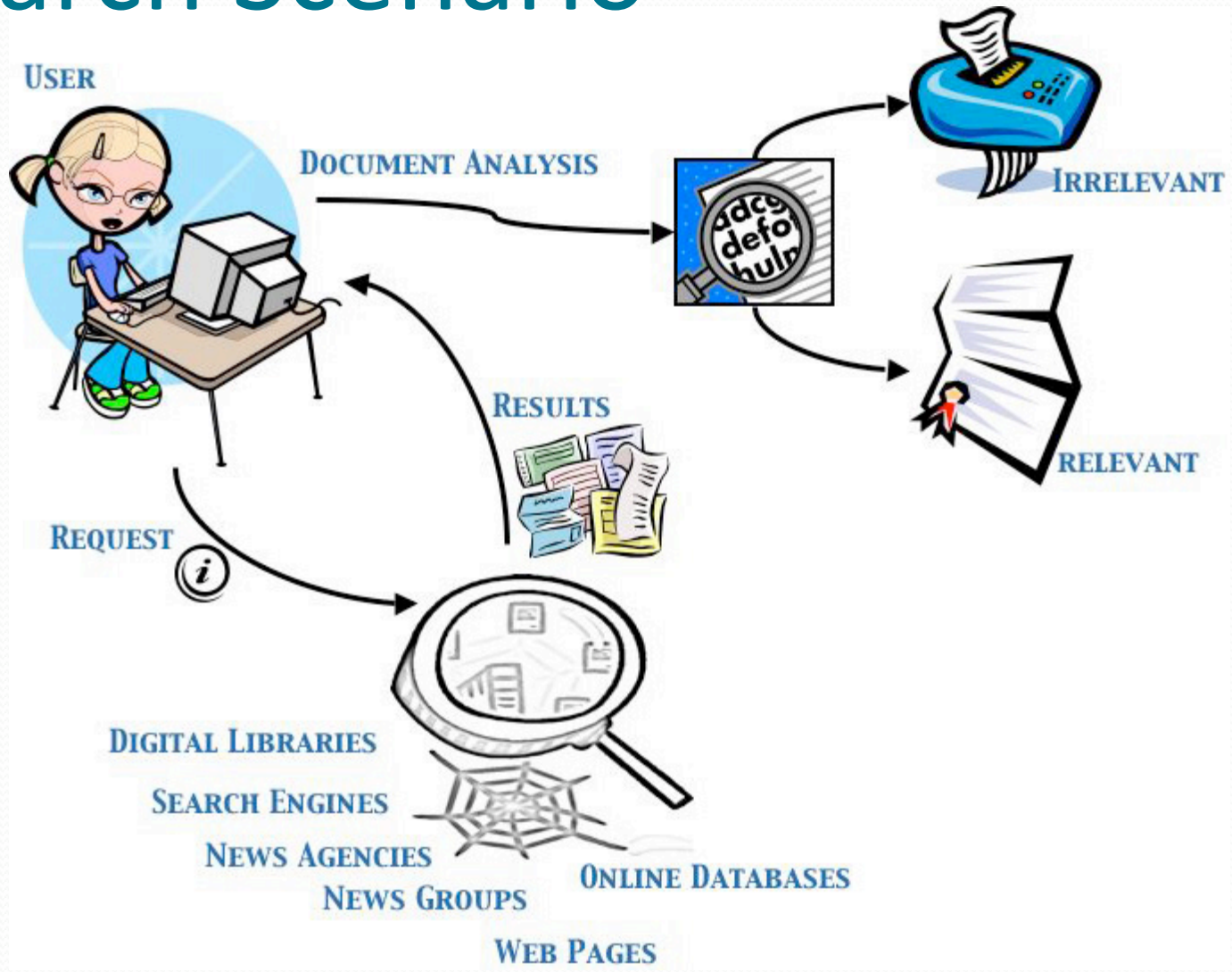
Three steps

1. *Collection Representation*
2. *Query Processing*
3. *Document Retrieval*

IR most known application

- Web documents search and retrieval

Search Scenario



Web Search

Two actors

1. Users (with short & long term information needs)
2. Web Information Resources

Three main tasks

1. Fetching
2. Browsing
3. Selecting

… relevant information

We are here

- Introduction
  - ✓ Search scenario
  - Personalization
  - Our Approach
- P *I* S A
  - System Functionality
  - Architecture
  - Prototype & Demo
- Conclusions and Future Work

Profiling

- Learn the user information needs

User Profile: a structured (machine) representation of the user's information needs and preferences

  - *What* to gather
  - *How* to deliver/present information

- Automatic (*Dynamic Profiling*)
- Manual (*Customization*)

Information Filtering

Filter relevant information

- On "explicit" user request (*Pull Modality*)
- On "implicit" user request (*Push Modality*)
- User Profile Matching (*Collaborative Filtering*)

Information Resources

Information Resources Heterogeneity

1. *Topic* of the information provided

   content (*what they provide*)

2. *Schema* of the information provided

   representation (*how they provide*)

Query example:

q = {author = Vardi, abstract = logic, abstract = computer science}

Metasearch

- *Where* to search: selection of relevant resources (*Automatic Resource Selection*)
- *How to query: reformulation of q into q'* *(Automatic Schema Matching)*
- *How to combine results: merging* (*Rank Fusion*)

Result Presentation

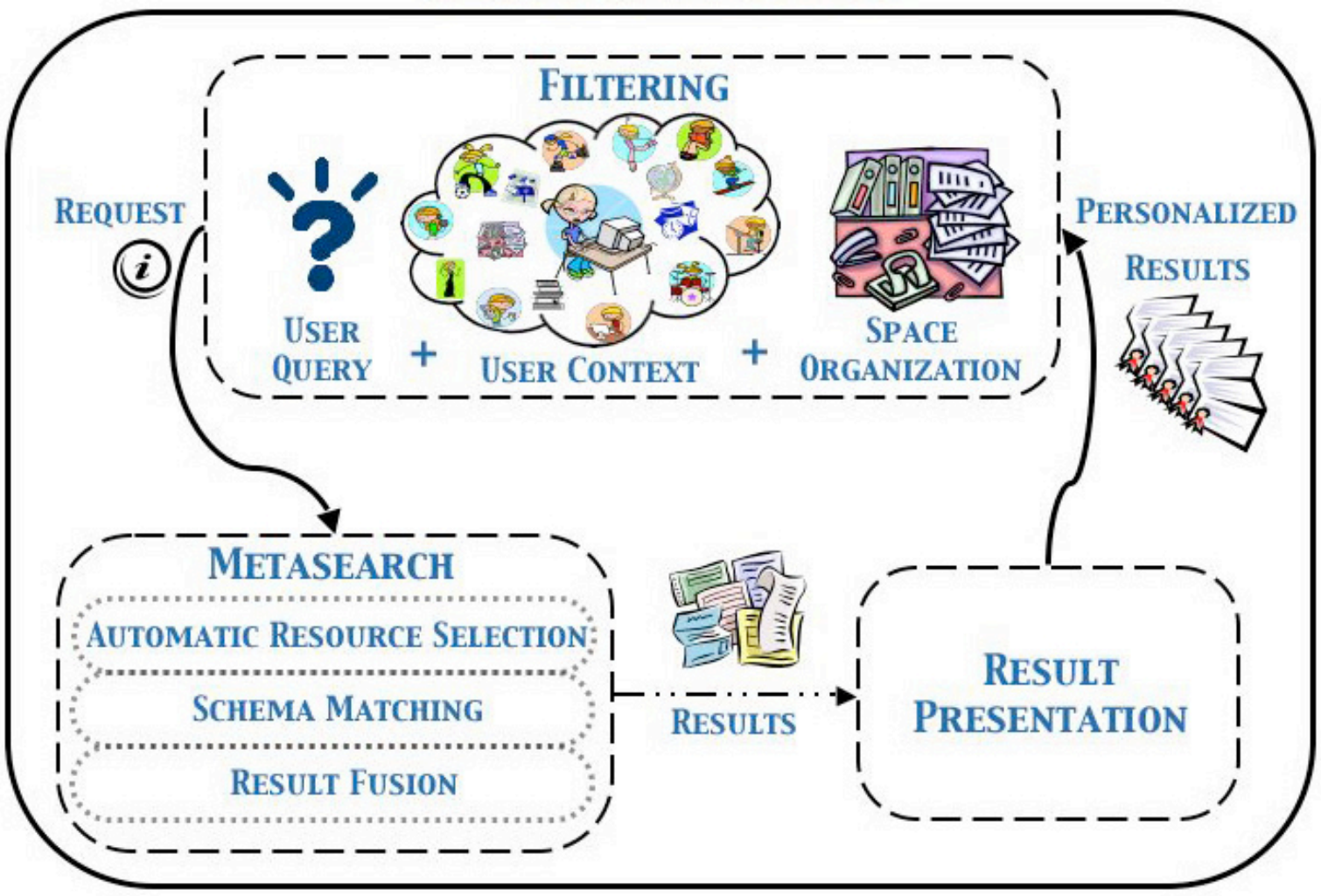
1. Personalized information summarization (*Document Summarization*)

2. Personalized information passage identification (*Passage Retrieval*)

We are here

- Introduction
  - ✓ Search scenario
  - ✓ Personalization
  - Our Approach
- P *I* S A
  - System Functionality
  - Architecture
  - Prototype & Demo
- Conclusions and Future Work

Personalized Search Assistant



State of the Art

Several environments using personalization:

- iGoogle: user-defined profiles
- Amazon
- MyYahoo
- Alerting DLs
- PENG - PErsonalised News content programminG
- PIA: agent based personal information system
- Cyclades: open collaborative virtual archive environment supporting users and communities
- … Databases, Newsgroups, Search engines, Web sites, etc.

No desktop application, most require user collaboration, all provide partial personalized features

We are here

- ✓ Introduction
  - ✓ Search scenario
  - ✓ Personalization
  - ✓ Our Approach
- P *I* S A
  - System Functionality
  - Architecture
  - Prototype & Demo
- Conclusions and Future Work

P I S A

For each personalization task:

- ✓ Best solution identification
- ✓ Algorithm Implementation
- ✓ Experimental evaluation

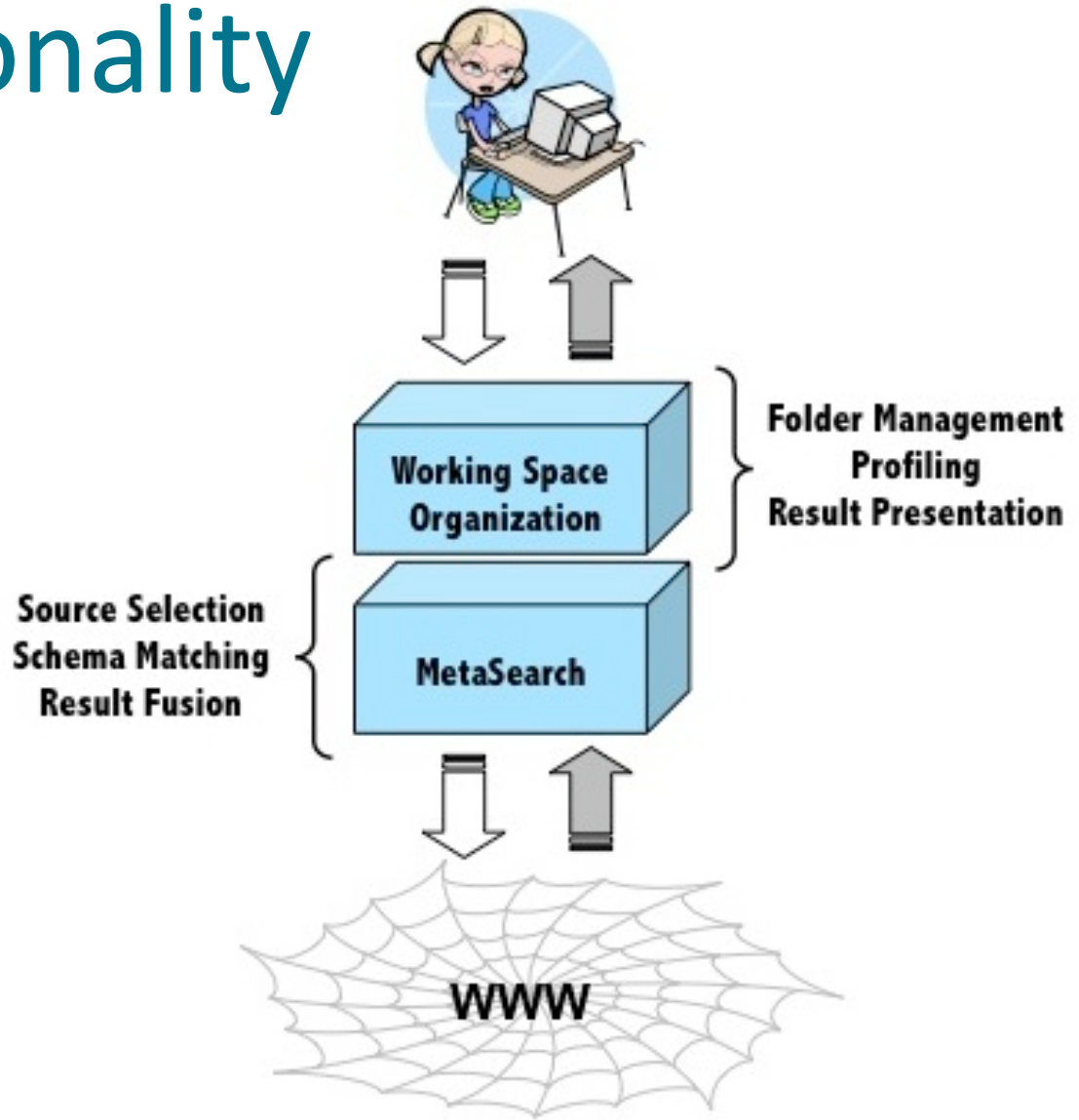
- Profiling and Filtering

  [Renda@IPM'05], [Renda@ICADL'02], [RendaTR'02], [RendaD6.1.1], [RendaD6.2.1], [RendaD5.1.1] [RendaD4.2.1], [RendaD3.0.1], [RendaD2.2.1]

- Metasearch

  [Renda@SAC'06], [Renda@CIKM'04], [Renda@SAC'03]

Functionality



Functionality w.r.t the System

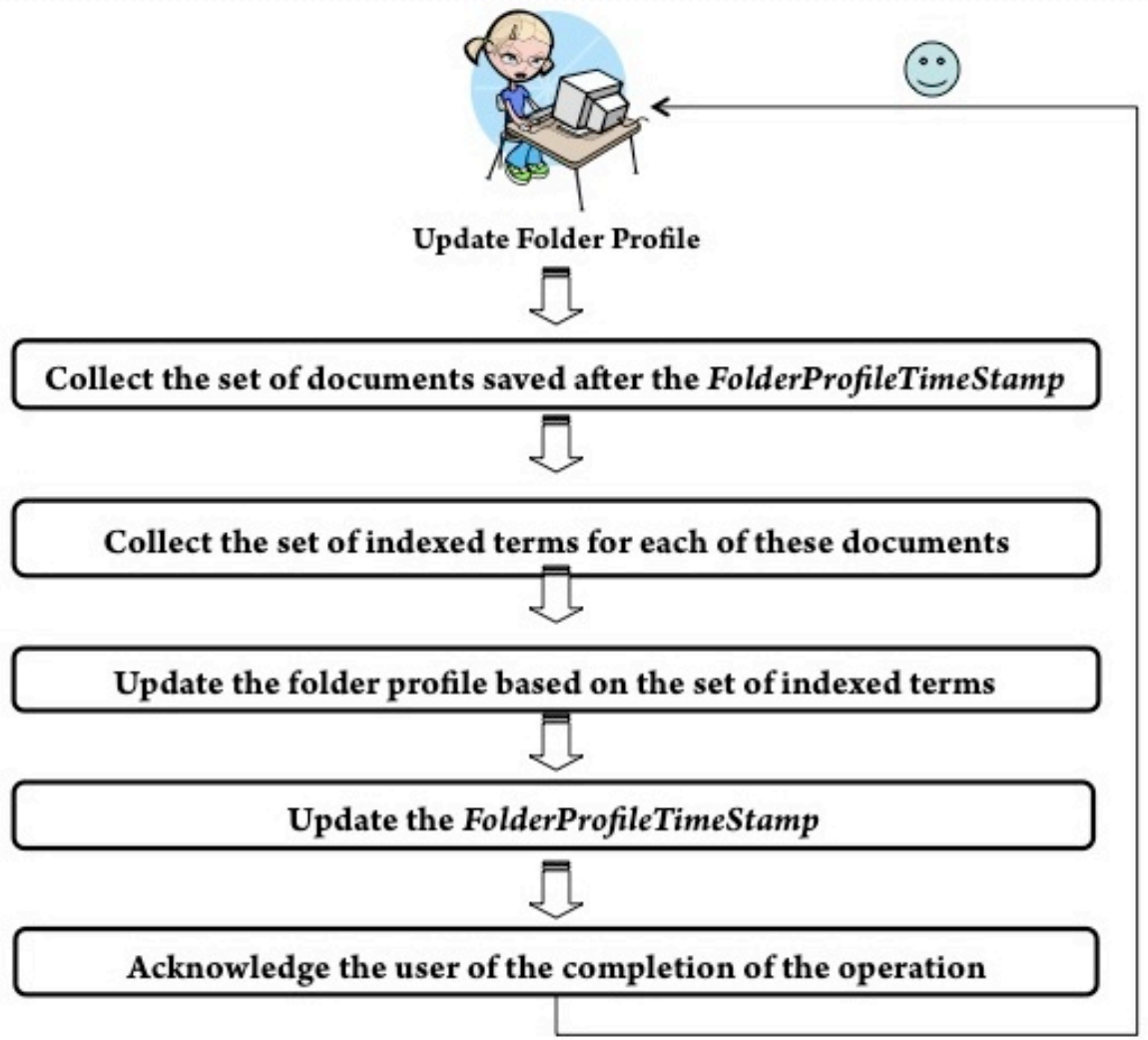
Actions the system performs in background:

1. Automatic resource selection
2. Schema matching
3. Profile management
4. Rank fusion
5. Document filtering

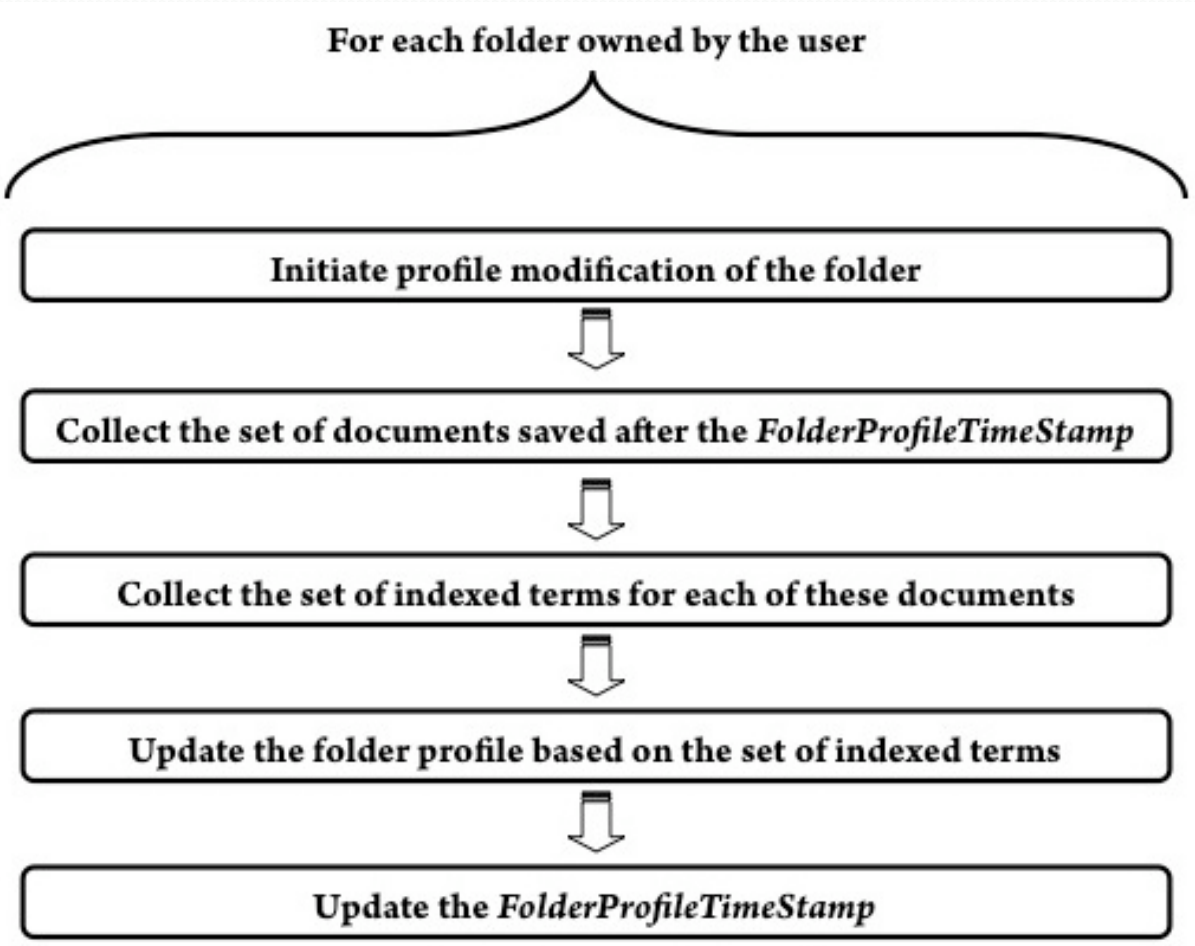
Functionality w.r.t the User

1. Login to the system
2. View/edit personal information
3. Set up system preferences
4. Manage folders
   - create, rename, move, empty, delete
5. Manage (and browse) documents
   - view, delete, cut, paste
6. Update the user and folder profiles
   - on-demand, at scheduled time
7. Search for new documents
   - *Filtered Search* (*Search New*/*Personalized*), *Simple Search*

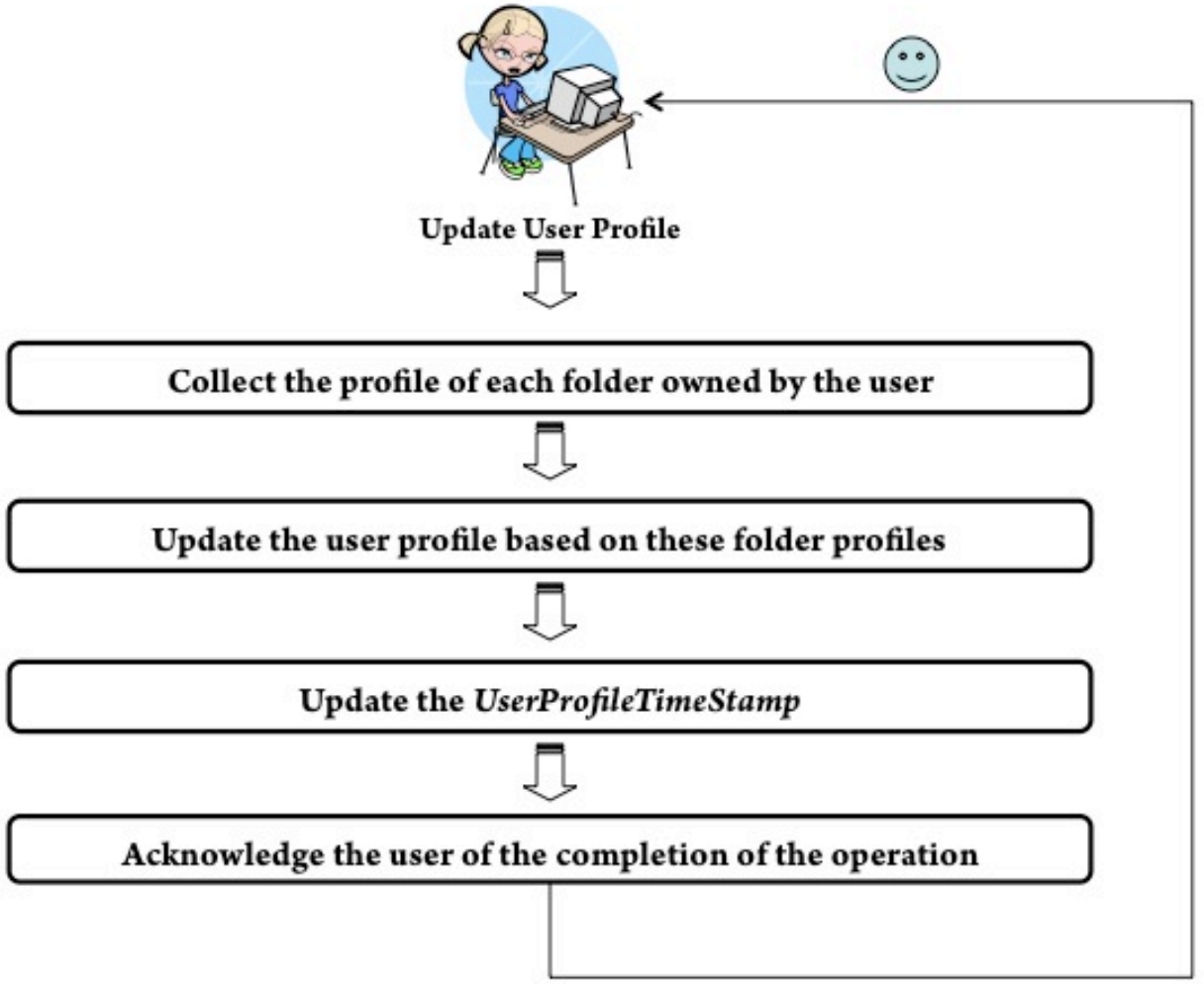
Update FP On-Demand



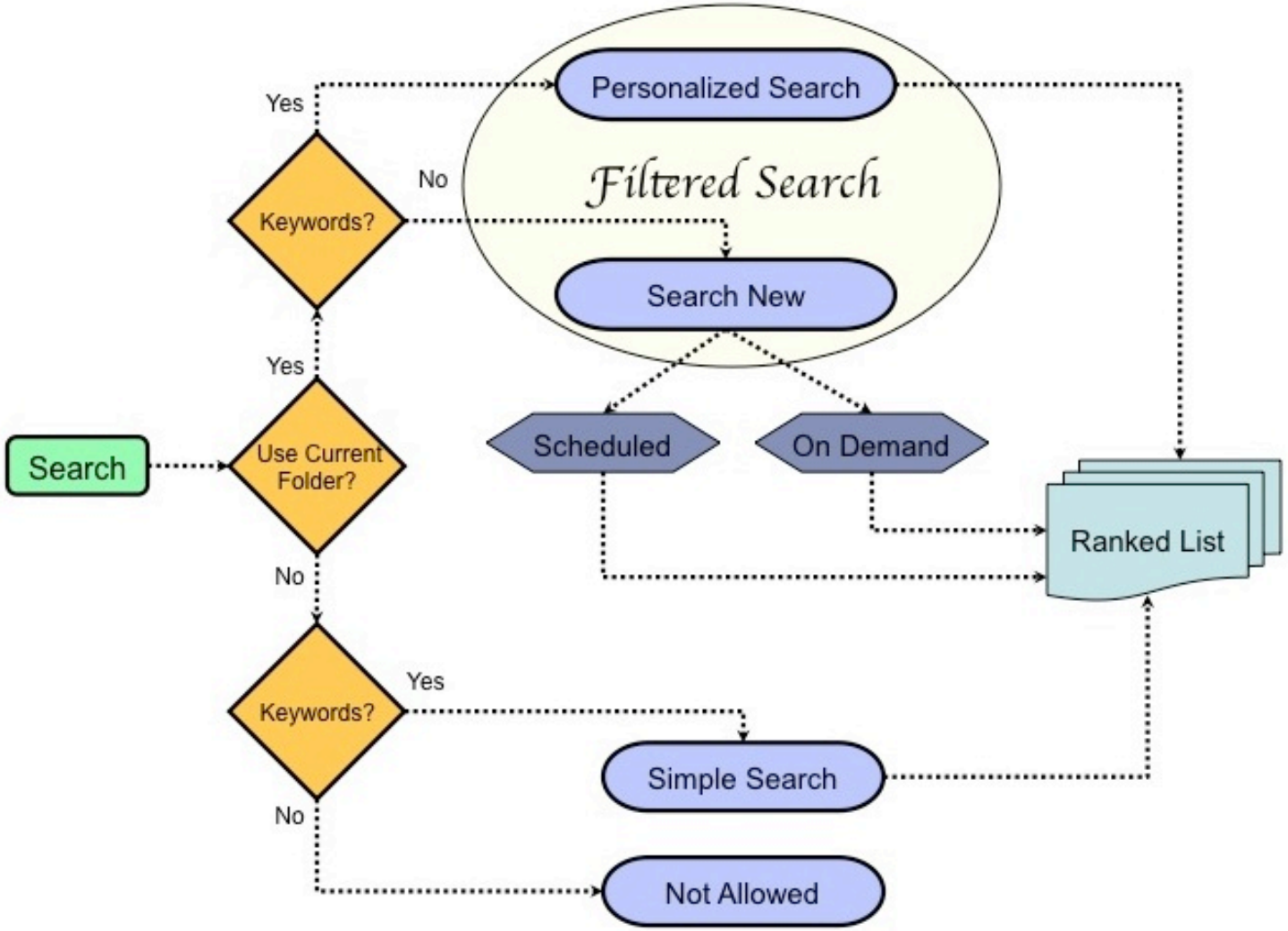
Scheduled FP Updating



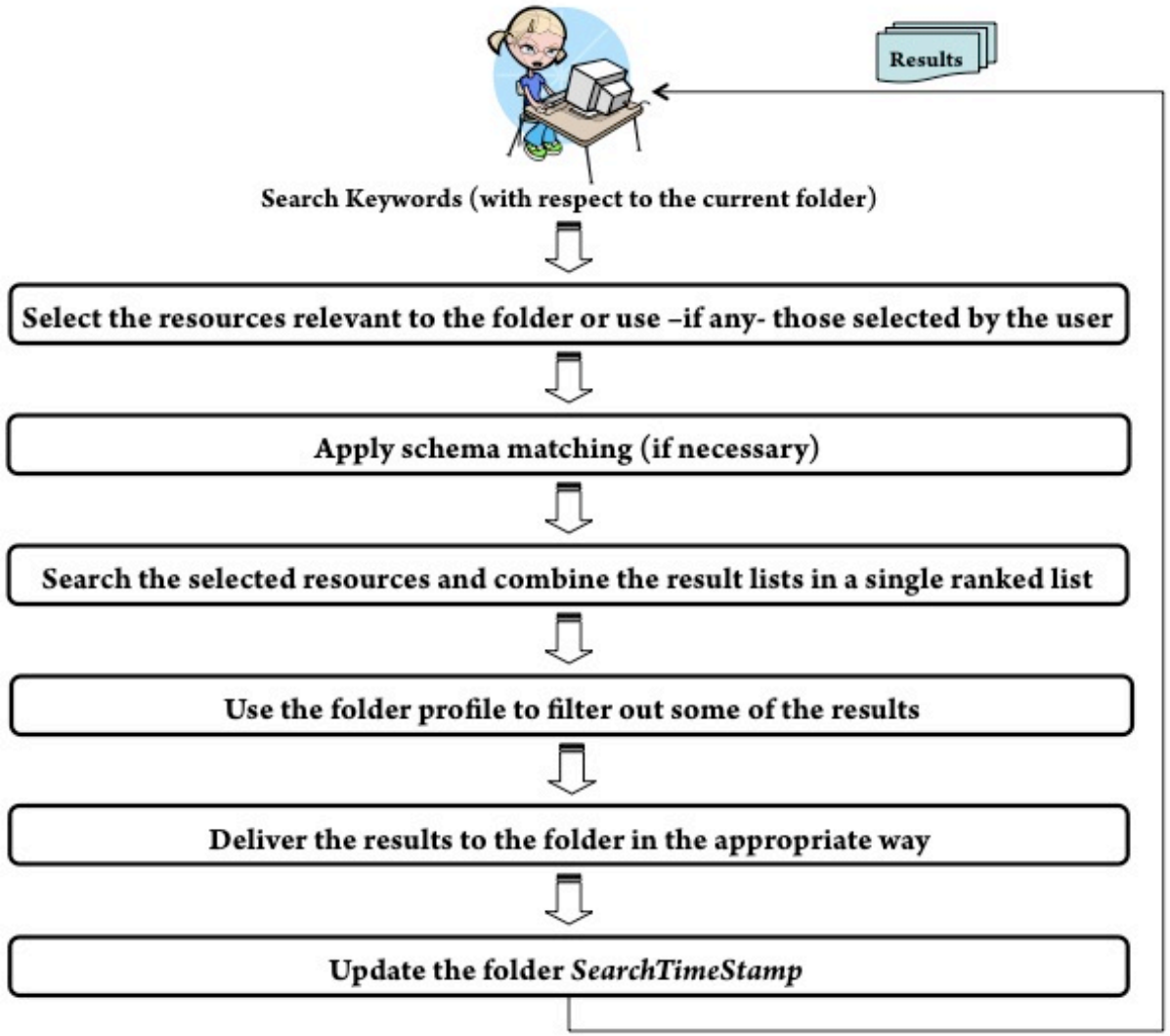
Update UP On-Demand



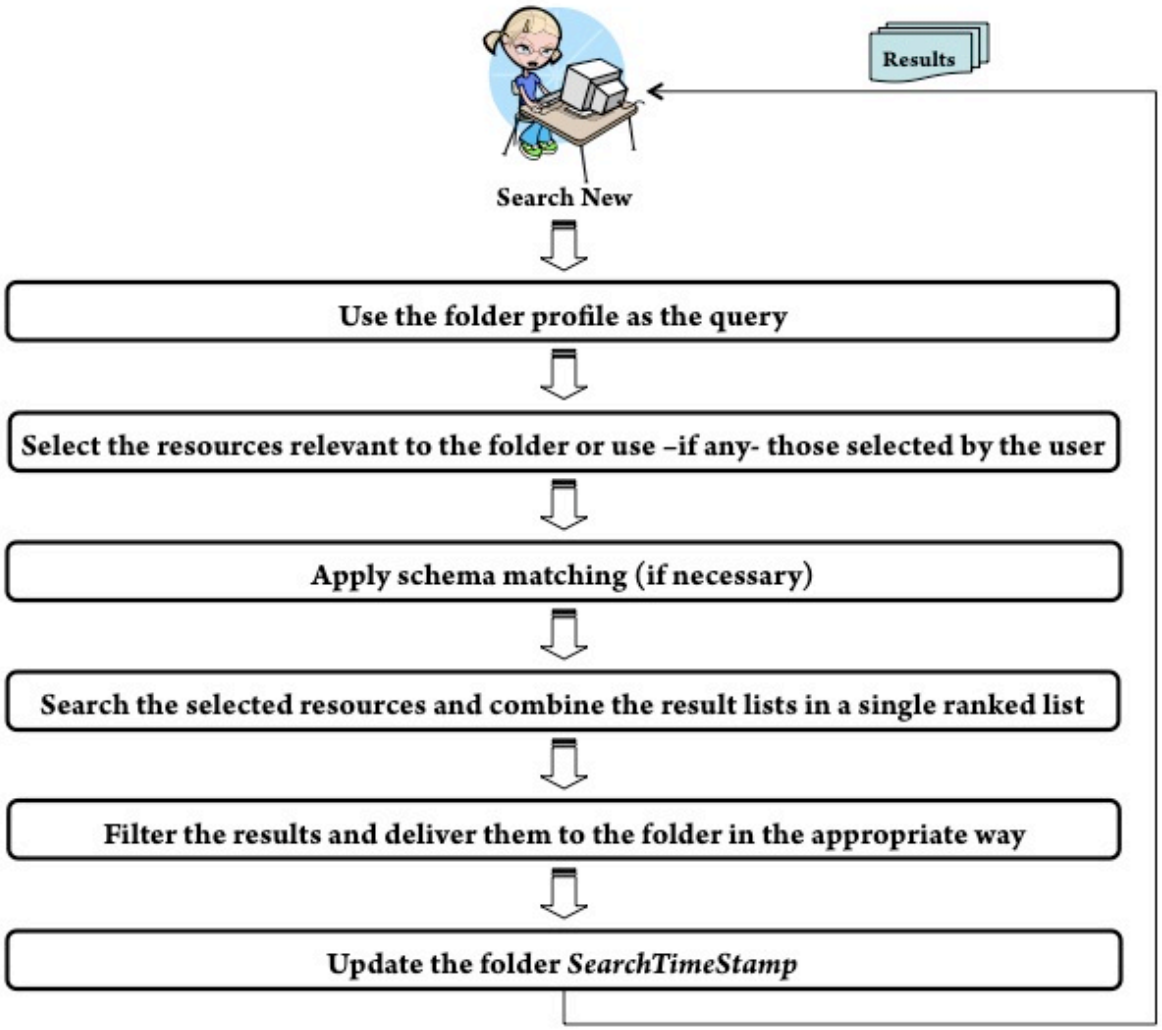
Search Mechanisms



Personalized Search



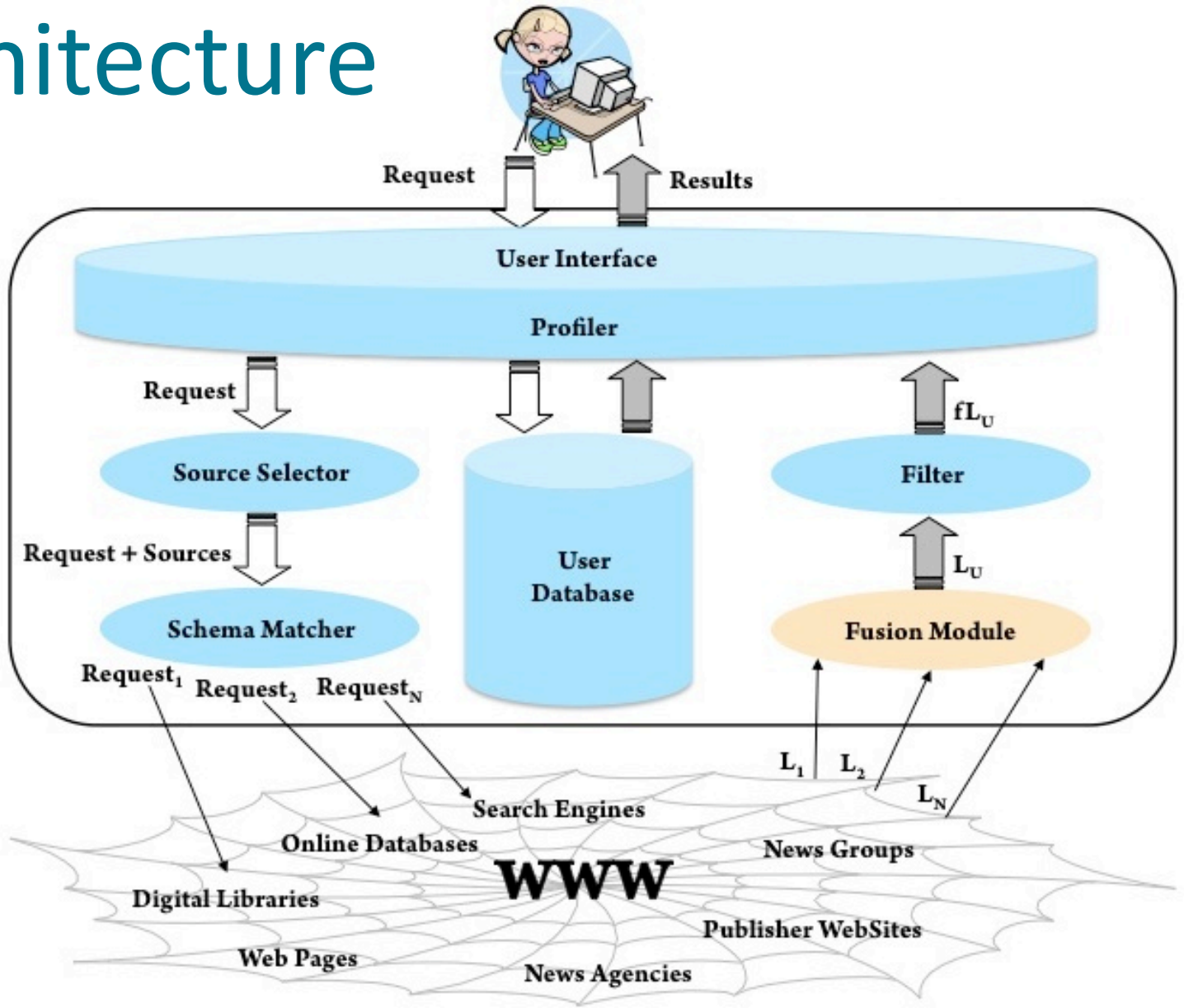
On-Demand Search New



We are here

- ✓ Introduction
  - ✓ Search scenario
  - ✓ Personalization
  - ✓ Our Approach
- P *I* S A
  - ✓ System Functionality
  - Architecture
  - Prototype & Demo
- Conclusions and Future Work

Architecture



User Database

- Preferences Table
- Settings Table
- Folders Table
- Documents Table
- Profiles Table

MySQL version 5.0.51

Profiler

Folder Profile $f_i = \frac{1}{|F_i|} \cdot \sum_{d_j \in F_i} d_j \quad \rightarrow \quad f_i = \langle w_{i1}, \ldots, w_{im} \rangle$

Where each $w_{ik} = \frac{1}{|F_i|} \cdot \sum_{d_j \in F_i} w_{jk}$

User Profile $p_u = \frac{1}{|\mathbb{F}_u|} \cdot \sum_{F_i \in \mathbb{F}_u} f_i \quad \rightarrow \quad p_u = \langle w_{u1}, \ldots, w_{um} \rangle$

[Renda@ICADL'02]
[Renda@IPM'05]

Filter

- *Max number of records*
- *Personalized Search*: document-profile content similarity

	t_1	...	t_k	...	t_m
d_1	w_{11}	...	w_{1k}	...	w_{1m}
d_2	w_{21}	...	w_{2k}	...	w_{2m}
...	...	...	...	...	...
d_j	w_{j1}	...	w_{jk}	...	w_{jm}
...	...	...	...	...	...
d_n	r_{n1}	...	w_{nk}	...	w_{nm}

	t_1	...	t_k	...	t_m
f_1	w_{11}	...	w_{1k}	...	w_{1m}
f_2	w_{21}	...	w_{2k}	...	w_{2m}
...	...	...	...	...	...
f_i	w_{i1}	...	w_{ik}	...	w_{im}
...	...	...	...	...	...
f_v	w_{v1}	...	w_{vk}	...	w_{vm}

$w_{ij} = tf_{ij} \cdot idf_i$, where

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \ (term\ frequency \text{ of } t_i \in d_j)$$

$$idf_i = \log \frac{|D|}{df_i} \ (inverse\ document\ frequency),\ df_i = \left|\{d \in D : t_i \in d\}\right|$$

[Renda@ICADL'02]
[Renda@IPM'05]

Source Selector

$$Sampling \; -> \; Approx(R_i)$$

$$G(q, R_i) = \frac{1}{|q|} \cdot \sum_{v_k \in q} p(v_k \mid R_i), \text{ where } p(v_k \mid R_i) = T_{i,k} \cdot I_k \cdot w_k \text{ (CORI)}$$

$$\mathrm{T_{i,k}} = df_{i,k} \Big/ \left(df_{i,k} + 50 + 150 \cdot \frac{cw_i}{\overline{cw}} \right) \quad (Collection\ Term\ Frequency)$$

$$I_k = \frac{1}{\log(|\Re| + 1.0)} \cdot \log\left(|\Re| + 0.5 / cf_k\right) \quad (Inverse\ Resource\ Frequency)$$

w_k: weight of the term in the query

$df_{i,k}$: number of documents in $Approx(R_i)$ with value v_k

cw_i: number of values in $Approx(R_i)$, $\overline{cw}$ the mean value

cf_k: number of approximated resources containing value v_k

$|\Re|$: number of resources

[Renda@CIKM'04]
[Renda@IPM'05]

Schema Matcher

$$q = \{A_1 = v_1, ..., A_q = v_q\},\ Mapping: A_T \rightarrow A_S$$

$$\forall A_k \in T, \forall R_i \text{ with schema } S_i, \text{ the most relevant } A_{ij} \in S_i : A_k \rightarrow A_{ij}$$

$$\forall R_k \in \Re, \text{ compute } Approx(R_k)$$

$$\text{Document } r_s = \{A_k = v_{k1}, ..., A_{ks} = v_{ks}\},\ r_s \in Approx(R_k)$$

$$\forall A_{kj} \in S_k,\ C_{k,j} = \bigcup_{r_s \in Approx(R_k)} \{r \mid r := \{A_{kj} = v_{kj}\}, A_{kj} = v_{kj} \in r_s\}$$

$$\forall q^* = \{A_i = v_i\} \in q, \text{ compute } G_{1 \le a \le q}(q^*, C_{k,k_a})\ (CORI)$$

$$G(q^*, C_{k,j}) = \max_a\ G(q^*, C_{k,k_a}) \text{ and set } A_i \rightarrow A_j \text{ for } R_k$$

[Renda@SAC'06]

Fusion Module

Ranking: linear ordering of a set of items

Rank fusion problem: compute a consensus ranking

Let $R = \{\tau_1, ..., \tau_n\}$ be a set of n rankings, T the *fused ranking*

Linear Combination Ranking Fusion Methods define the score as:

$$s^{\mathrm{T}}(i) = h(i,R)^{y} \cdot \sum_{\tau \in R} \alpha_{\tau} \cdot w^{\tau}(i)$$

where:

i) Each $\tau \in R$ has been normalized and w^{τ} is the *normalized weight*

ii) $y \in \{0,1\}$, 1 hits do count, 0 hits do not count

iii) $\sum_{\tau \in R} \alpha_{\tau} = 1,\ \alpha_{\tau} > 0$ (*source priority*)

[Renda@SAC'03]

Fusion Module (cont.)

Rank-based method: *CombMNZ*

$$s^{\mathrm{T}}(i) = h(i,R)^{0} \cdot \sum_{\tau \in R} \frac{1}{|R|} \cdot w^{\tau}(i)$$

where:

$$i)\ w^{\tau}(i) = 1 - \frac{\tau(i) - 1}{|\tau|}\ (rank\ normalization)$$

$$ii)\ y = 0\ (do\ not\ consider\ hits)$$

$$iii)\ \alpha_{\tau} = \frac{1}{|R|}\ (resources\ with\ same\ priority)$$

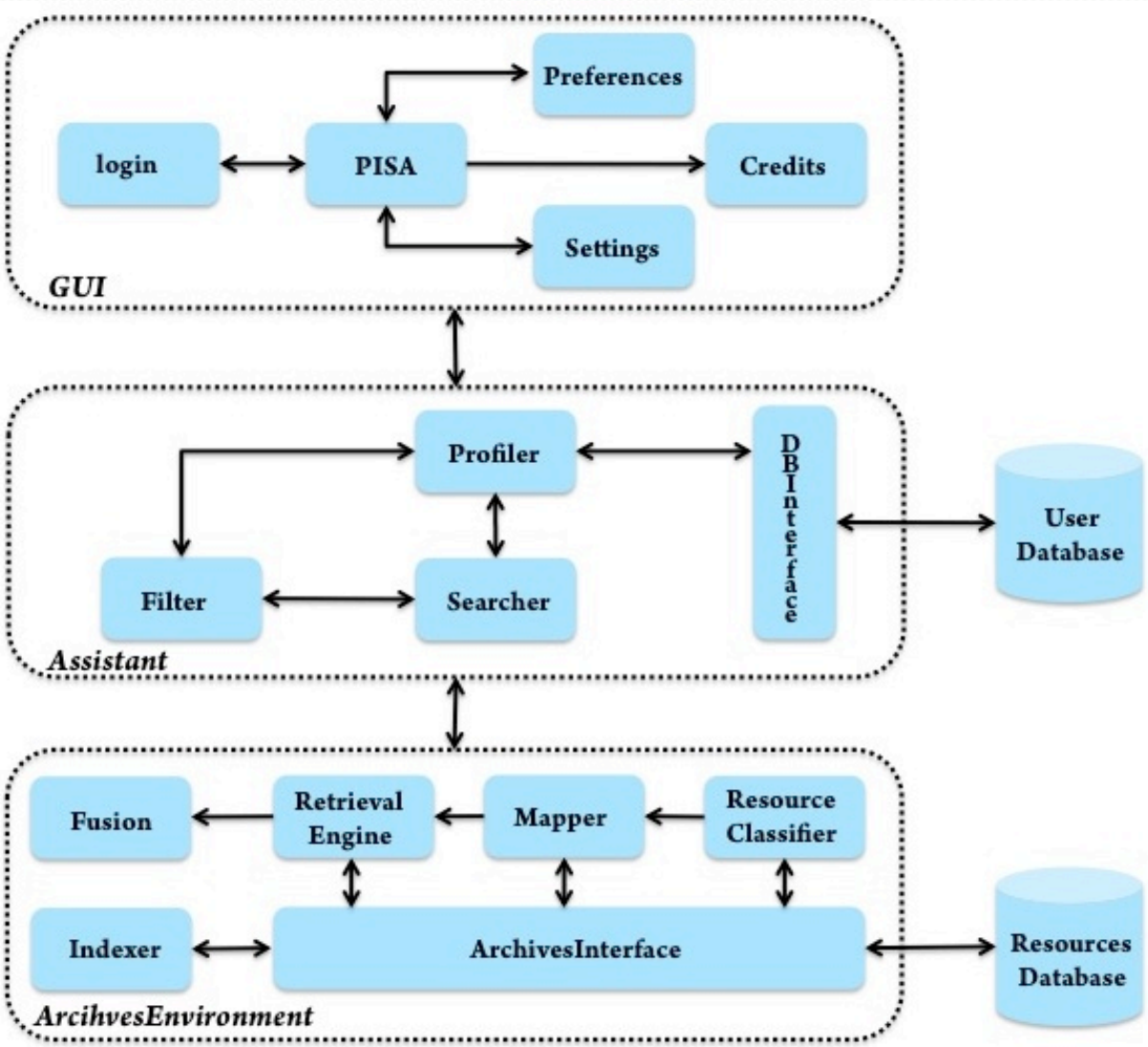
We are here

- ✓ Introduction
  - ✓ Search scenario
  - ✓ Personalization
  - ✓ Our Approach
- P *I* *S* A
  - ✓ System Functionality
  - ✓ Architecture
  - Prototype & Demo
- Conclusions and Future Work

P I S A Prototype

- Tools and Libraries
  - Java Platform, Standard Edition, and the Java Development Kit (JDK), version 6.0
  - MySQL version 5.0.51
  - MySQL Connector/JDBC version 3.1.8
  - Apache Lucene library version 2.2
    - 8 resources with a total of 45k indexed documents
- Code
  - three packages: GUI, Assistant, ArchivesEnvironment
  - 80 Java classes
  - more than 15,000 lines of source code

Internal Architecture



User Interface & Functionality

Demo…

We are here

- ✓ Introduction
  - ✓ Search scenario
  - ✓ Personalization
  - ✓ Our Approach
- ✓ P I S A
  - ✓ System Functionality
  - ✓ Architecture
  - ✓ Prototype & Demo
- Conclusions and Future Work

Conclusions

- ✓ Personalization
  - ✓ New services
    - ✓ Smart Searching
    - ✓ Information Organization
- ✓ P *I* S A
  - ✓ Desktop application offering single user personalization
  - ✓ Learns user's preferences to improve search and retrieval
  - ✓ Users gain in time, quality of documents and satisfied information needs

Future work

- Result Presentation
  - Passage Retrieval, Summarization
- Individual and community P *I* SA user

  Community networks sharing
  - Data and documents (*P2P*, *Grid*) **[Renda@CIKM'04]**
  - Wireless communication facilities (*Mesh Community Networks*) **[Renda@PerCom'08, Renda@ComNet'07]**

Centralized (or hierarchical) profile maintenance

Related Publications

1. S. Burresi, C. Canali, **M.E. Renda**, and P. Santi. *Meshchord: A location-aware, cross-layer specialization of chord for wireless mesh networks*. In Proc. of the 6h Annual IEEE International Conference on Pervasive Computing and Communications, pages 206-212, Hong Kong, 2008. **[Renda@PerCom'08]**
2. L. Galluccio, G. Morabito, S. Palazzo, M. Pellegrini, **M.E. Renda**, and P. Santi. *Georoy: A location-aware enhancement to viceroy peer-to-peer algorithm*. Computer Networks, 51(8):1998-2014, 2007. **[Renda@ComNet'07]**
3. **M.E. Renda** and U. Straccia. *Automatic structured query transformation over distributed digital libraries*. In Proc. of the 21st Annual ACM Symposium on Applied Computing, pages 1078-1083, Dijon, France, 2006. ACM Press. **[Renda@SAC'06]**
4. **M.E. Renda** and U. Straccia. *A personalized collaborative digital library environment: a model and an application*. Information Processing & Management, 41(1):5-21, 2005. **[Renda@IPM'05]**
5. **M.E. Renda** and J. Callan. The robustness of content-based search in hierarchical peer to peer networks. In Proc. of the 13th Annual ACM Conference on Information and Knowledge Management, pages 562-570, Washington D.C., U.S.A., 2004. ACM Press. **[Renda@CIKM'04]**
6. **M.E. Renda** and U. Straccia. *Web metasearch: Rank vs. score based rank aggregation methods*. In Proceedings 18th Annual ACM Symposium on Applied Computing, pages 841-846, Melbourne, Florida, USA, 2003. ACM. **[Renda@SAC'03]**
7. **M. E. Renda** and U. Straccia. *A personalized collaborative digital library environment*. In 5th International Conference on Asian Digital Libraries, number 2555 in LNCS, pages 262{274, Singapore, Republic of Singapore, 2002. Springer-Verlag. **[Renda@ICADL'02]**

Related Publications (cont.)

8. **M. E. Renda** and U. Straccia. *A recommendation system in a collaborative digital library environment*. Technical Report 2002-TR-06, IEI - CNR, Pisa, Italy, 2002. **[RendaTR02]**
9. U. Straccia, G. Fischer, N. Fuhr, H. Avancini, L. Candela, **M.E. Renda**, F. Sebastiani, N. Papadopoulos, D. Plexousakis, T. Gross, and T. Kreifelts. *It consortium and noe questionnaire report*. Technical report, CYCLADES project IST-2000-25456, 2003. Deliverables D6.1.1 and D6.2.1. **[RendaD6.1.1-D6.2.1]**
10. U. Straccia, G. Fischer, N. Fuhr, H. Avancini, L. Candela, **M.E. Renda**, F. Sebastiani, N. Papadopoulos, D. Plexousakis, T. Gross, and T. Kreifelts. *System validation report*. Technical report, CYCLADES project IST-2000-25456, 2003. Deliverable 5.1.1. **[RendaD5.1.1]**
11. U. Straccia, G. Fischer, N. Fuhr, D. Castelli, P. Pagano, **M.E. Renda**, F. Sebastiani, N. Papadopoulos, D. Plexousakis, T. Gross, and T. Kreifelts. *System testing report*. Technical report, CYCLADES project IST-2000-25456, 2003. Deliverable D4.2.1. **[RendaD4.2.1]**
12. U. Straccia, G. Fischer, N. Fuhr, D. Castelli, P. Pagano, **M.E. Renda**, F. Sebastiani, N. Papadopoulos, D. Plexousakis, T. Gross, and T. Kreifelts. *Detailed system specification report*. Technical report, CYCLADES project IST-2000-25456, 2002. Deliverable D3.0.1. **[RendaD3.0.1]**
13. U. Straccia, G. Fischer, N. Fuhr, D. Castelli, P. Pagano, **M.E. Renda**, F. Sebastiani, N. Papadopoulos, D. Plexousakis, T. Gross, and T. Kreifelts. *Global system architecture report*. Technical report, CYCLADES project IST-2000-25456, 2001. Deliverable D2.2.1. **[RendaD2.2.1]**

Thank you all!!

Dedicated to Paolo, Bianca & Marta…