Information Processing Competence Center

Implicit feedback learning in semantic and collaborative information retrieval systems

by Gérard Dupont[1, 2]

written under the direction of

Sébastien Adam[1], Yves Lecourtier[1], Bruno Grilheres[1, 2], Stephan Brunessaux[2]

[1] **Laboratoire d'Informatique de Traitement de l'Information et des Systèmes** (LITIS) - Saint-Étienne-du-Rouvray, France

[2] **EADS Defense and Security**, Information Processing and Competence Center - Val de Reuil, France

Summary

- Introduction
- Enhanced IRS with feedback learning
- Feedback learning in VITALAS
- Focus on learning using behavior measure as feedback
- Conclusion and future work

EADS DEFENCE & SECURITY

VITALAS

Introduction

Information retrieval ?

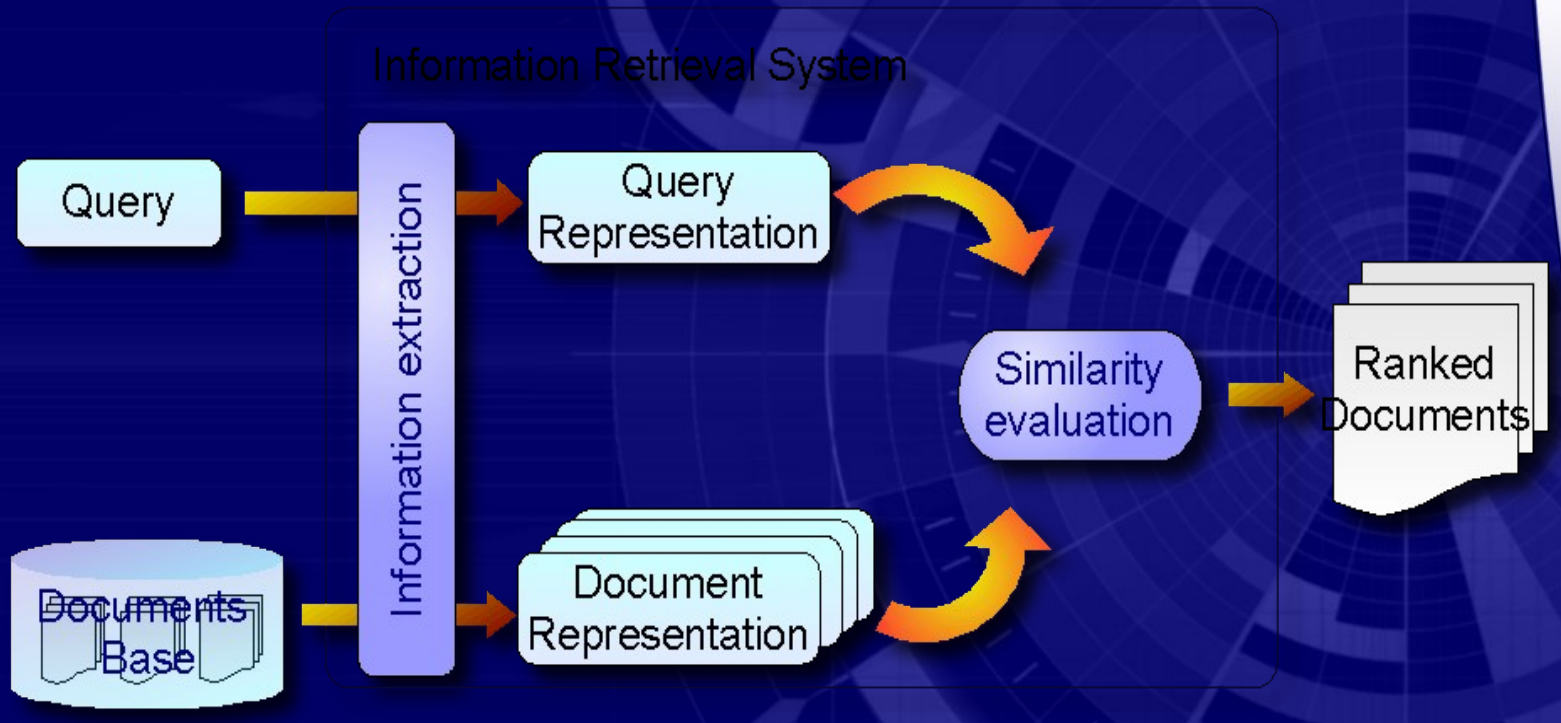
"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfy an information need from within large collections (usually stored on computers)"

An introduction to information retrieval - Manning , 2007.

Variables :

- Document and collection of documents : library, database, Intranet, Internet...
- Unstructured information without precise meaning
- Information needs expressed by users

Simple view of IRS



Aim : Matching query with documents (or part of documents)

– Information model to represent document and needs

– Similarity evaluation theory to produce ranked list of documents

Information model

– Vector model dedicated to text document retrieval

Document term vectors

$$\vec{D}_i = \begin{pmatrix} word_1 - 0{,}24 \\ word_2 - 0{,}4 \\ word_3 - 0{,}1 \\ \dots \\ word_N - 0{,}1 \end{pmatrix}$$

Query term vectors

$$\vec{Q} = \begin{pmatrix} word_1 - 0 \\ word_2 - 1 \\ word_3 - 0 \\ \dots \\ word_N - 1 \end{pmatrix}$$

Example of similarity formula (normalized cosinus)

$$score(\vec{D}_i, \vec{Q}) = \frac{\langle \vec{D}_i, \vec{Q} \rangle}{\|\vec{D}_i\| . \|\vec{Q}\|} = \frac{\sum_{k=1}^{N} d_{i,k} \times q_k}{\sqrt{\sum_{k=1}^{N} d_{i,k}^2} \times \sqrt{\sum_{k=1}^{N} q_k^2}}$$

– Generalized probabilistic model

$$P(D|L) = \prod_i P(A_i = a_i | L)$$

- Term vector extended to description through attributes/values
- Relevance as probability
- Possibility to handle multimedia features as attributes

EADS DEFENCE & SECURITY

VITALAS

Limits of current IRS

Strong assumptions :

- Dimensions of the vector model or attributes in the probabilistic model are independent
- User information needs is fully described by its query

Not verified in most of the cases :

- Linguistic study tells us that terms are not independent in texts (synonymous, antonymous,...) neither are features extracted in CBIR
- User can not define precisely their needs since they are trying to complete their knowledge

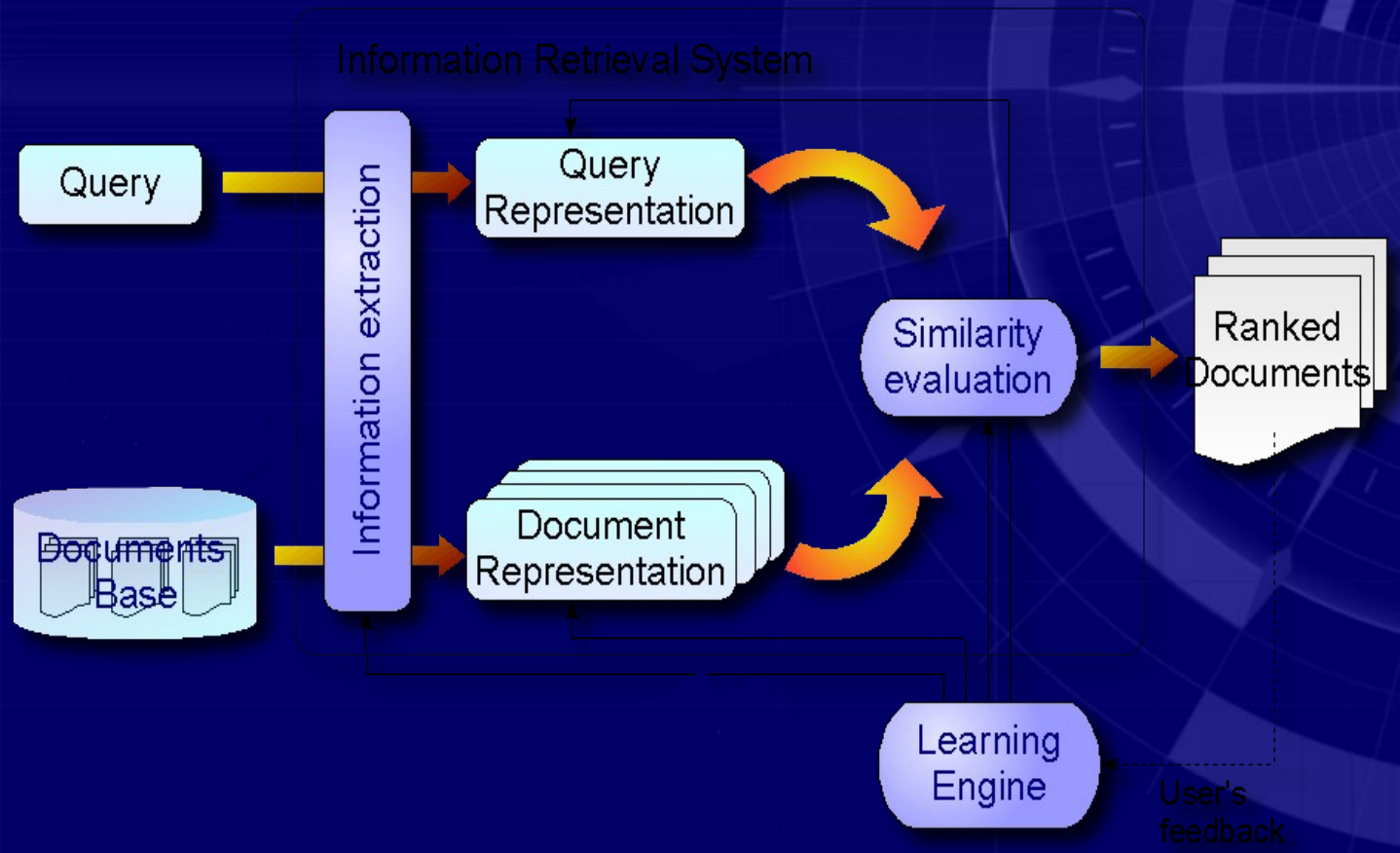
EADS DEFENCE & SECURITY

VITALAS

Enhanced IRS with feedback learning

Feedback learning

By giving feedback about the presented documents, users tell more about their needs to the system



Search becomes (again) an iterative process.

The IRS can enhance itself at multiple levels :

- information representation
- similarity evaluation

EADS DEFENCE & SECURITY

VITALAS

Feedback learning strategies

Many possibilities explored :

- Query rewriting (short term)
- Search context modelling (mid term)
- User model learning (long term)

Proven efficiency of explicit relevance feedback learning concept :

ex : “Rocchio” Algorithm established in the 70's

EADS DEFENCE & SECURITY

VITALAS

Explicit vs Implicit feedback

Experimental (and operational) studies have shown that users are reluctant to provide explicit feedback on documents

Use of implicit (behavioral) indicators to fill the gap

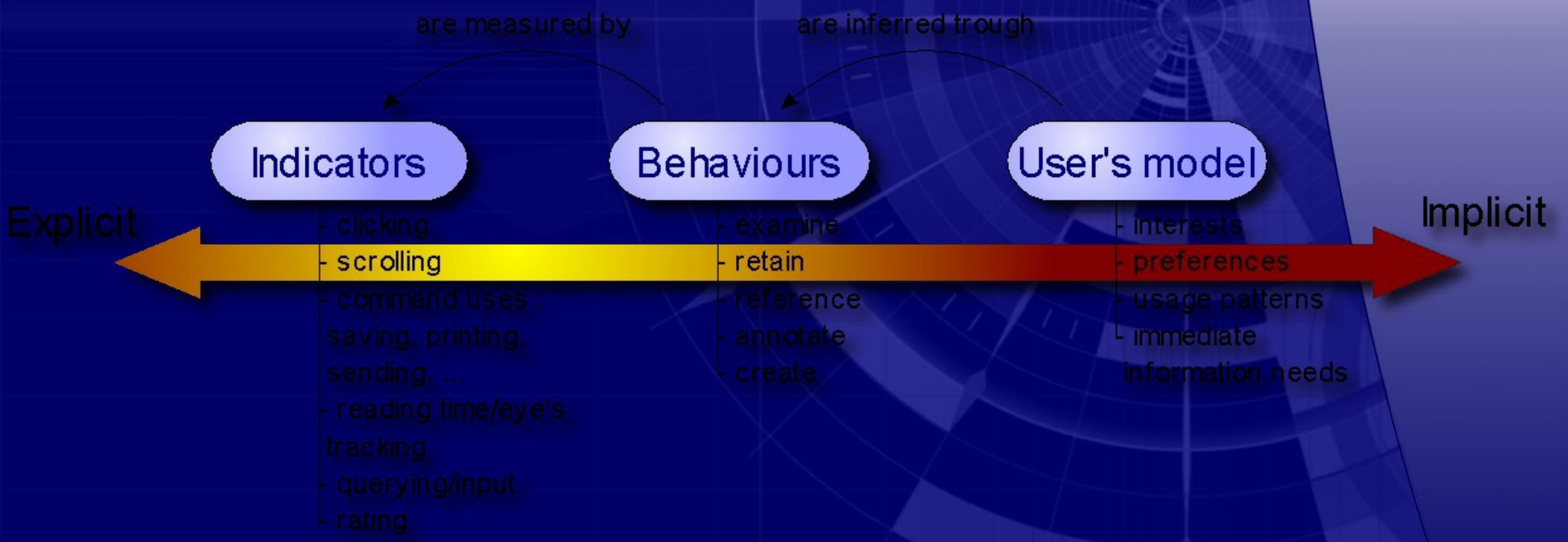
ex : reading time, scroll behavior, click trough data, ...

Implicit data are known :

- to raise privacy issues (but solutions exist)
- to be noisy (or biased, but issues are raised on explicit bias too)
- to be easy to gather in large amount

Explicit vs Implicit feedback

Hybrid approches combines explicit and implicit data.



EADS DEFENCE & SECURITY

VITALAS

Feedback learning and search in context in VITALAS

EADS DEFENCE & SECURITY

VITALAS

Analysis of search logs

Search in context and relevance feedback : starting with search log data study in collaboration with CWI.

Research issues are :

- Do users of professional images IRS have the same behavior as classic users from state-of-the-art studies ?
- Are they advanced searchers ?
- Can we detect specific behavior pattern ?

Aim is to select the right approach.

EADS DEFENCE & SECURITY

VITALAS

First experiments

Experiments (conducted by CWI) with “implicit collaboration” using past search sessions :

- query suggestion
- term suggestion
- results suggestion

Experiments (conducted by EADS) on using implicit feedback data to infer document interests/relevance :

- interaction events tracking in web based GUI
- framework to learn search context in WebLab platform
- optimization framework for query/term suggestion

EADS DEFENCE & SECURITY

VITALAS

Focus on learning using behavior measurements as feedback

EADS DEFENCE & SECURITY

VITALAS

Search context with feedback

Explicit and implicit feedback have advantages and drawbacks. It is better to combine feedback through a common framework.

Measurements of current user behavior to extract interests :

- Time spent on reading a document
- Selection of terms in abstract
- Click on a link after reading its description
- Explicit rating of items

Matrix X of measurements per documents/parts of documents

X(i,j) = measure of behavior j on element i

$$X = \begin{pmatrix} m_{1,1} & m_{1,2} & m_{1,3} & m_{1,4} \\ m_{2,1} & m_{2,2} & m_{2,3} & m_{2,4} \\ m_{3,1} & m_{3,2} & m_{3,4} & m_{3,4} \end{pmatrix}$$

EADS DEFENCE & SECURITY

VITALAS

Search context with implicit feedback

- Using past search history to learn relevance pattern in behavior measurements

Rel= matrix of behavior patterns on relevant documents

$$Rel=\begin{pmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} \\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} \\ r_{3,1} & r_{3,2} & r_{3,4} & r_{3,4} \end{pmatrix}$$

Irr = matrix of behavior patterns on irrelevant documents

$$Irr=\begin{pmatrix} s_{1,1} & s_{1,2} & s_{1,3} & s_{1,4} \\ s_{2,1} & s_{2,2} & s_{2,3} & s_{2,4} \\ s_{3,1} & s_{3,2} & s_{3,4} & s_{3,4} \end{pmatrix}$$

- Classic supervised learning problem which enables the computation of current search context.

ex: a weighted vector of terms reflecting current interests.

EADS DEFENCE & SECURITY

VITALAS

Searching in context

Using the search context to enhance user experience while in a search session :

- Query expansion and/or suggestion to help users to define their needs
- Changing the similarity and ranking algorithm to personalize the behavior of the system to the user and its current needs
- Adapting the presentation of results
- Providing tools to interact/explore the corpus (to provide more accurate data for implicit relevance feedback)

Searching in context : a multi objective optimisation probl

Query expansion and/or suggestion as a multi objective optimization problem

: finding the “best query” regarding multiple criteria and constraints

$$\vec{f}(\vec{x}) = \left\{ f_1(\vec{x}), \ldots, f_i(\vec{x}), \ldots, f_n(\vec{x}) \right\}$$

$$\text{with} \quad \vec{g}(\vec{x}) \geq 0 \Leftrightarrow \left\{ g_1(\vec{x}) \geq 0, \ldots, g_i(\vec{x}) \geq 0, \ldots, g_m(\vec{x} \geq 0) \right\}$$

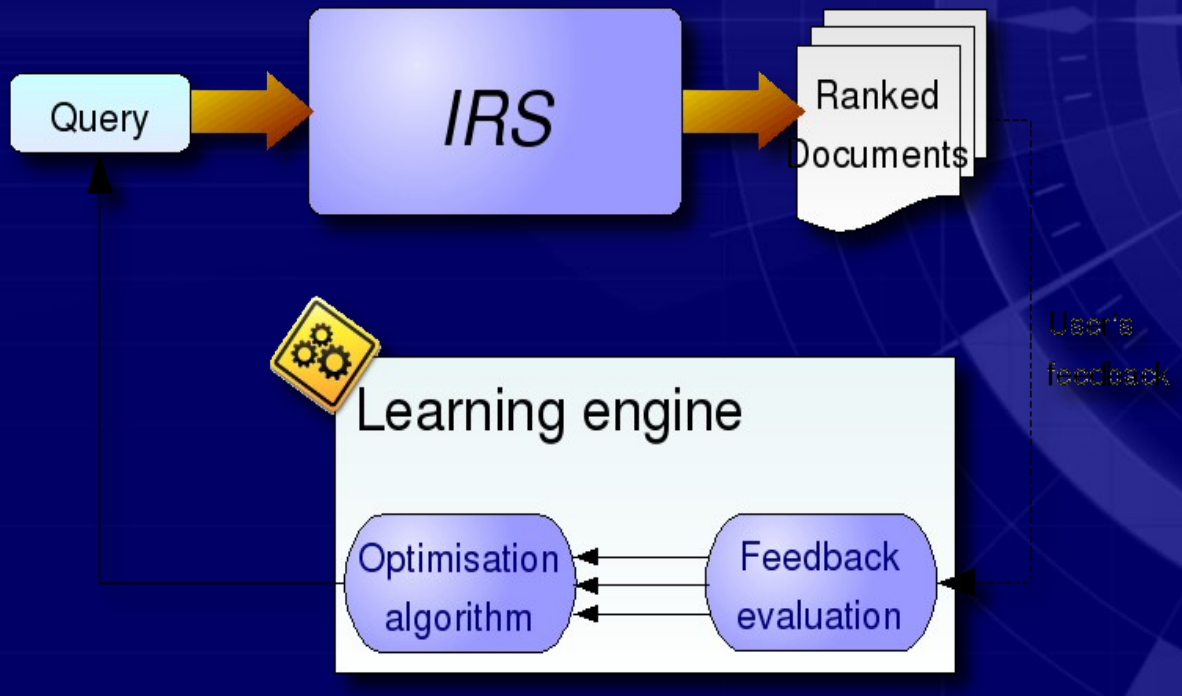
Criteria examples : Precision, Recall, Diversity, Novelty

$$P = \frac{N_{relevant\,results}}{N_{results}} \qquad R = \frac{N_{relevant\,results}}{N_{relevant\,doc}}$$

Adapted and personalized for each user or community of users

Evolutionary algorithm for query expansion/suggestion

Evolutionary algorithms to optimize the first user query :



1. Given a query of N terms
2. Rank document
3. Evaluate criteria through user's feedback
4. Optimise the query vector to maximize the criteria

Evolutionary algorithm for query expansion/suggestion

Difficulties :

- High dimensionality of term space

  reduced through the use of search context learned from feedback

- Combinaison advanced query operator

  use of genetic programming to compute advanced queries

Multiple level of impact :

- Query suggestion (with or without complex syntax)
- Term suggestion to disambiguate with context
- Implicit rewriting with “push” of new documents

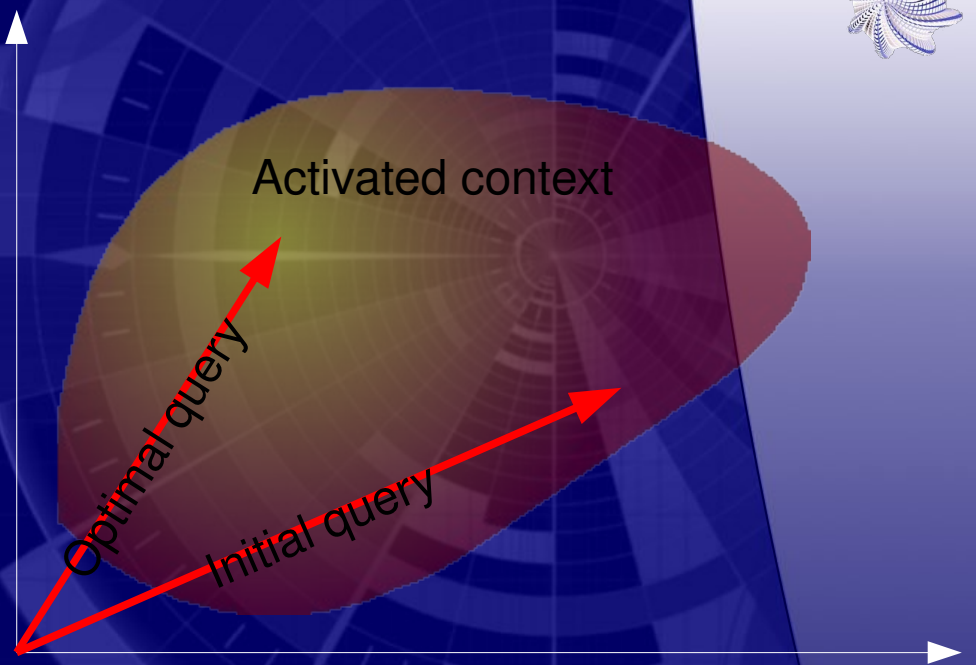
Evolutionary algorithm for query expansion/suggestion

User query vector in the whole vocabulary of N terms

$$\vec{Q} = \begin{pmatrix} word_1 - 0 \\ word_2 - 1 \\ word_3 - 0 \\ \dots \\ word_N - 1 \end{pmatrix}$$

Search context which ”activate” some parts of the vocabulary

$$\vec{C} = \begin{pmatrix} word_1 - 0{,}0 \\ word_2 - 0{,}89 \\ word_3 - 0{,}5 \\ \dots \\ word_N - 0{,}01 \end{pmatrix}$$



Geometrical representation of search space, initial user query and search context used to limit the searched area

EADS DEFENCE & SECURITY

VITALAS

Expanding context using semantic and collaboration

Use of semantic knowledge bases to expand the context : changing the information model to concept space

- Classification/clustering problem in a graph or a hierarchy of semantic concepts
- Use of Word Sense Disambiguation (WSD) techniques

Knowledge representation comes out users past searches : use of collaborative search experiences and/or external bases (ontologies, wikipedia...)

Conclusion and future work

EADS DEFENCE & SECURITY

VITALAS

Future research paths

- Finalisation of Logs analysis
- Exploration of feedback learning approaches
  - Implicit relevance learning with already existent data : Experimentation of state-of-the-art approaches in IR based on statistics and of pattern recognition approaches
  - Query expansion with evolutionary algorithms to optimise query weights with operators
  - Extension of collaboration to enlarge user search context by using user model similarity matching and developing new paradigm of collaboration
- Evaluation of proposed approaches
- Integration within WebLab platform (VITALAS V2 ?)

Bibliography

[Bottraud2003] Bottraud, J.C. and Bisson, G. and Bruandet, M.F., "Apprentissage des profils pour un agent de recherche d'information", Actes de CAP 2003, 2003.

[Claypool2001] Claypool, Mark and Le, Phong and Waseda, Makoto and David, Brown, "Implicit Interest Indicators", 2001.

[Crestani1998] Crestani, F. and Lalmas, M. and Rijsbergen, C. and Campbell, I., "Is this document relevant?...Probably: A survey of probabilistic models in information retrieval", ACM Computing Surveys, Vol. 30, no. 4, pp., Dec., p.528-552, 1998.

[Gaussier2003] Gaussier, Éric and Stéfanini, Marie-Hélène, "Assitance intelligente à la recherche d'informations", Hermès science (Ed.), Traité des sciences et techniques, 2003.

[Joachims2002] Joachims, Thorsten, "Optimizing search engines using clickthrough data", ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2002.

[Kelly2004] Kelly, Diane, "Understanding implicit feedback and document preference: A naturalistic user study", 2004.

[Manning2007] Manning, Christopher D. and Prabhakar, Raghavan and Schütze, Hinrich, "An introduction to information retrieval", Cambridge University Press (Ed.), 2007.

[Middleton2004] Middleton, Sturat E. and Shadbolt, Nigel R. and Roure, David C. De, "Ontological User Profiling in Recommender Systems", ACM Transactions on Information Systems (TOIS), Vol. 22, p.54-88, 2004.

[Bottraud2004] Bottraud, Jean-Christophe, "Un assistant adaptatif pour la recherche d'information : AIRA", 2004.

[Radlinski2007] Radlinski, Filip and Joachims, Thorsten, "Active Exploration for Learning Rankings from Clickthrough Data", 2007.

[White2004] White, Ryen W., "Implicit Feedback for Interactive Information Retrieval", 2004.

About the author

- Research engineer at EADS DS
- PhD thesis since November 2006 (in collaboration with LITIS laboratory)
- Involved in VITALAS (EC project 2007/2009)
  - EADS DS as software architect
  - Personal involvement in “search in context”
- Research interests :

  information retrieval, search engine, Web intelligence, information extraction, semantic extraction, machine learning, evolutionary algorithm, swarm algorithm, optimisation

Enhanced IRS