Implicit feedback learning in semantic and collaborative information retrieval systems

by Gérard Dupont\textsuperscript{1, 2}

written under the direction of

Sébastien Adam\textsuperscript{1}, Yves Lecourtier\textsuperscript{1}, Bruno Grilheres\textsuperscript{1, 2}, Stephan Brunessaux\textsuperscript{2}

\textsuperscript{1} Laboratoire d'Informatique de Traitement de l'Information et des Systèmes (LITIS) - Saint-Étienne-du-Rouvray, France

\textsuperscript{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
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)”


Variables:

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

\[ \vec{D}_i = \begin{pmatrix}
\text{word}_1 - 0.24 \\
\text{word}_2 - 0.4 \\
\text{word}_3 - 0.1 \\
\vdots \\
\text{word}_N - 0.1
\end{pmatrix} \]

**Query term vectors**

\[ \vec{Q} = \begin{pmatrix}
\text{word}_1 - 0 \\
\text{word}_2 - 1 \\
\text{word}_3 - 0 \\
\vdots \\
\text{word}_N - 1
\end{pmatrix} \]

**Example of similarity formula**

(normalized cosinus)

\[
\text{score} (\vec{D}_i, \vec{Q}) = \frac{\langle \vec{D}_i, \vec{Q} \rangle}{\| \vec{D}_i \| \cdot \| \vec{Q} \|} = \frac{\sum_{k=1}^{N} d_{i,k} \times q_k}{\sqrt{\sum_{k=1}^{N} d^2_{i,k}} \times \sqrt{\sum_{k=1}^{N} q^2_k}}
\]

- 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
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
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
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
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 approaches combine explicit and implicit data.

Explicit
- clicking
- scrolling
- command uses: saving, printing, sending, ...
- reading time/eye's tracking
- querying/input
- rating

Indicators

Behaviours
- examine
- retain
- reference
- annotate
- create

User's model
- interests
- preferences
- usage patterns
- immediate information needs

Implicit

are measured by

are inferred through
Feedback learning and search in context in 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.
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
Focus on learning using behavior measurements as feedback
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

$$X(i,j) = \text{measure of behavior } j \text{ on element } i$$

$$X = \begin{pmatrix}
m_{1,1} & m_{1,2} & m_{1,3} & m_{1,4} \\
m_{2,1} & m_{2,2} & m_{2,3} & m_{2,4} \\
m_{3,1} & m_{3,2} & m_{3,3} & m_{3,4}
\end{pmatrix}$$
Search context with implicit feedback

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

\[ \text{Rel} = \begin{pmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} \\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} \\ r_{3,1} & r_{3,2} & r_{3,3} & r_{3,4} \end{pmatrix} \]

\[ \text{Irr} = \begin{pmatrix} s_{1,1} & s_{1,2} & s_{1,3} & s_{1,4} \\ s_{2,1} & s_{2,2} & s_{2,3} & s_{2,4} \\ s_{3,1} & s_{3,2} & s_{3,3} & s_{3,4} \end{pmatrix} \]

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

ex: a weighted vector of terms reflecting current interests.
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 problem

Query expansion and/or suggestion as a multi objective optimization problem: finding the “best query” regarding multiple criteria and constraints

\[ \tilde{f}(\tilde{x}) = [f_1(\tilde{x}), \ldots, f_i(\tilde{x}), \ldots, f_n(\tilde{x})] \]

with \( \tilde{g}(\tilde{x}) \geq 0 \iff \{ g_1(\tilde{x}) \geq 0, \ldots, g_i(\tilde{x}) \geq 0, \ldots, g_m(\tilde{x} \geq 0) \} \)

Criteria examples: Precision, Recall, Diversity, Novelty

\[ P = \frac{N_{\text{relevant results}}}{N_{\text{results}}} \quad R = \frac{N_{\text{relevant results}}}{N_{\text{relevant doc}}} \]

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

\[ \tilde{Q} = \begin{pmatrix} \text{word}_1 - 0 \\ \text{word}_2 - 1 \\ \text{word}_3 - 0 \\ \vdots \\ \text{word}_N - 1 \end{pmatrix} \]

Search context which "activate" some parts of the vocabulary

\[ \tilde{C} = \begin{pmatrix} \text{word}_1 - 0,0 \\ \text{word}_2 - 0,89 \\ \text{word}_3 - 0,5 \\ \vdots \\ \text{word}_N - 0,01 \end{pmatrix} \]

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


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
Information Processing Competence Center

Enhanced IRS

Information retrieval system

Query

Semantic Information Extraction

Query representation

Document representation

Similarity computation

 Ranked document

Document base

Knowledge base

Learning engine

User models

User logs