A Probability Ranking Principle for Interactive IR

Norbert Fuhr

Outline

Motivation

- The classical PRP
- Questioning the PRP assumptions
- Interactive Retrieval

The classical PRP

- Task: Retrieve relevant documents for given query
- Optimum retrieval is achieved when documents are ranked according to descending probability of relevance!

Assumptions:

1. Relevance of a document to a query is independent of other documents
2. Scanning through the ranked list is the major task of the user (and the only one considered)

Questioning the PRP assumptions

- Relevance depends on documents the user has seen before
- Relevance judgment is not the most laborious task in interactive retrieval

Interactive Retrieval

- User has a rich set of interaction possibilities
  - (re)formulate query
  - selection based on summaries of various granularity
  - select related terms from list
  - follow document link
  - relevance judgment
- Information need changes during a search

No theoretic foundation for constructing IIR systems!

Approach

- Requirements for an IIR-PRP
- Basic Assumptions
- Abstraction: Situations with Lists of Choices

Requirements for an IIR-PRP

- Consider the complete interaction process
- Allow for different costs for different activities
- Allow for changes of the information need

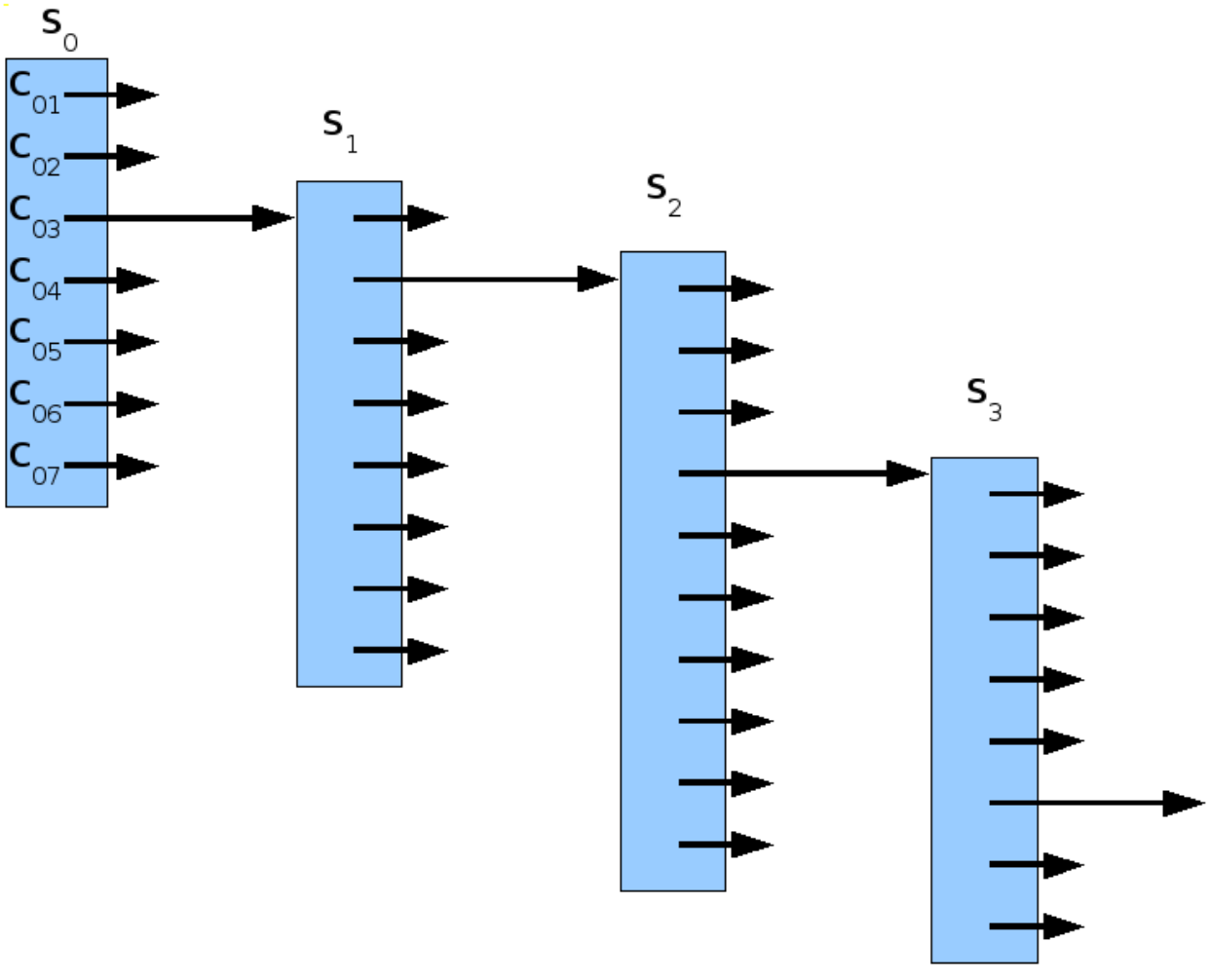
Basic Assumptions

- Focus on a functional level of interaction (usability issues disregarded here)
- System presents list of choices to the user
- Users evaluate choices in linear order
- Only positive decisions/choices are of benefit for a user

Examples of decision lists

- ranked list of documents
- list of summaries
- list of document cluster
- KWIC list
- list of expansion terms
- links to related documents
- ...

Abstraction: Situations with Lists of Choices



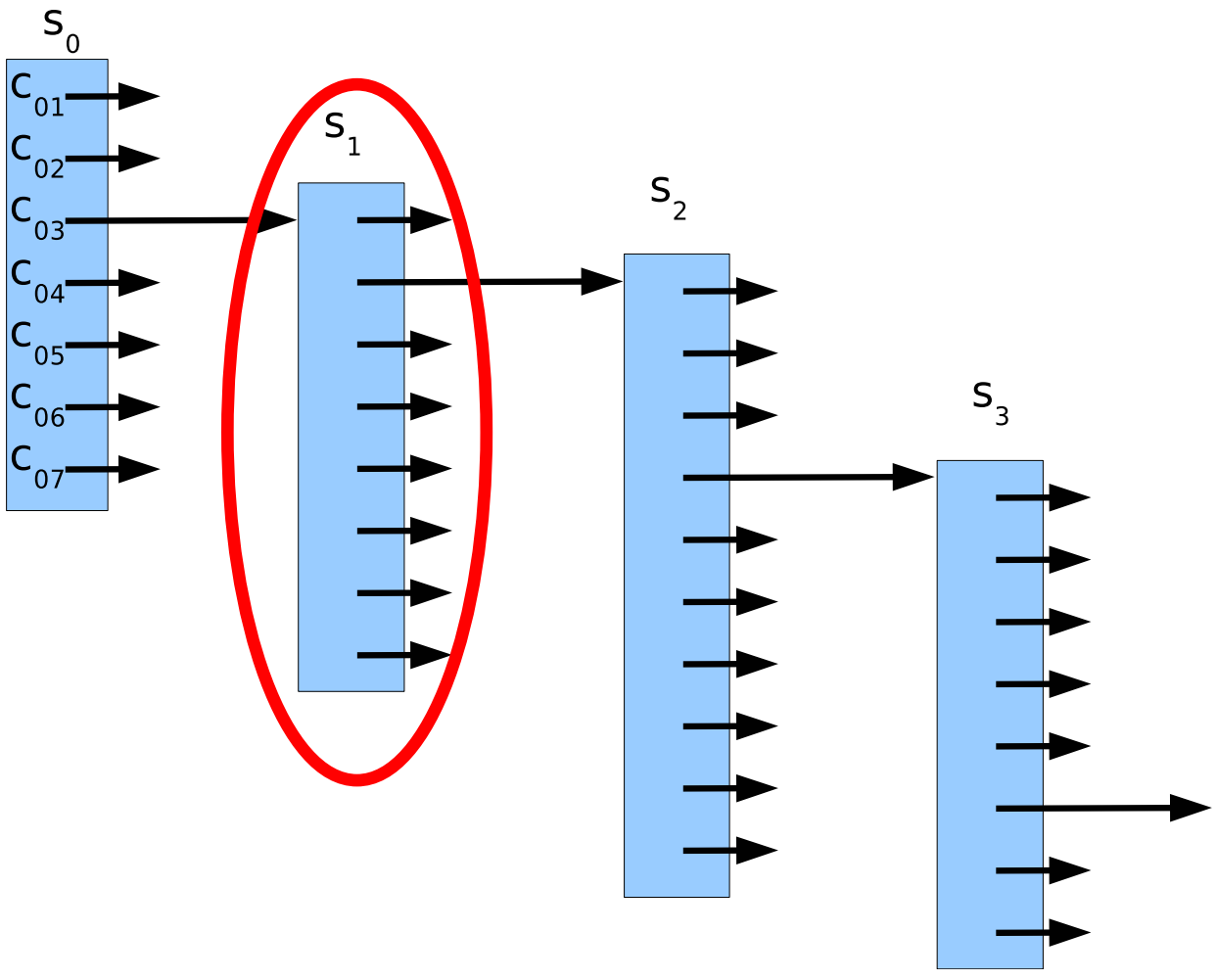
The Model

- Choices
- Expected Benefit of a choice
- Selection lists
- Ranking of choices

Basic ideas

- A user moves from situation to situation
- In each situation s_i, the user is presented a list of (binary) choices $< c_{i1}, c_{i2}, \ldots, c_{i,n_i} >$
- The user decides about each of these choices sequentially
- The first positive decision moves the user to a new situation s_j

Probabilistic model focusing on single situation



Expected Benefit of a choice

p_{ij} probability that the user will accept choice c_{ij}

e_{ij} < 0: effort for evaluating the choice c_{ij}

a_{ij} > 0: resulting benefit from positive decision

Expected benefit of choice c_{ij}

$$E(c_{ij}) = e_{ij} + p_{ij} a_{ij}$$

Example

Web search: 'Java' $\rightarrow$ n_0=290 mio. hits

System proposes extension terms:

term	n_i	p_{ij}	a_{ij}	$p_{ij}a_{ij}$
program	195 mio	0.67	0.4	0.268
blend	5 mio	0.02	4.0	0.08
island	2 mio	0.01	4.9	0.049

benefit $b_{ij} = \log \frac{n_0}{n_i}$

Expected benefit of a choice list

situation s_i with list of choices $r_i =< c_{i1}, c_{i2}, \ldots, c_{i,n_i} >$

expected benefit of choice list:

$$
\begin{aligned}
E(r_i) &= e_{i1} + p_{i1}a_{i1} + \\
&\quad (1 - p_{i1})\,(e_{i2} + p_{i2}a_{i2} + \\
&\quad (1 - p_{i2})\,(e_{i3} + p_{i3}a_{i3} + \\
&\quad \ldots \\
&\quad (1 - p_{i,n-1})\,(e_{in} + p_{in}a_{in})\ ))\\
&= \sum_{j=1}^{n} \left(\prod_{k=1}^{j-1} (1 - p_{ik}) \right) (e_{ij} + p_{ij}a_{ij})
\end{aligned}
$$

Ranking of choices

Consider two subsequent choices c_{il} and $c_{i,l+1}$

$$E(r_i) = \sum_{\substack{j=1 \\ l \neq j \neq l+1}}^{n} \left(\prod_{k=1}^{j-1} (1 - p_{ik}) \right) (e_{ij} + p_{ij} a_{ij}) + t_i^{l,l+1}$$

where

$$t_i^{l,l+1} = (e_{il} + p_{il} a_{il}) \prod_{k=1}^{l-1} (1 - p_{ik}) + (e_{i,l+1} + p_{i,l+1} a_{i,l+1}) \prod_{k=1}^{l} (1 - p_{ik})$$

analogously $t_i^{l+1,l}$ for $< \ldots, c_{i,l+1}, c_{il}, , \ldots >$

Difference between alternative rankings

$$
\begin{aligned}
d_i^{l,l+1} &= \frac{t_i^{l,l+1} - t_i^{l+1,l}}{\prod_{k=1}^{l-1}(1-p_{ik})} \\
&= e_{il} + p_{il}a_{il} + (1-p_{il})(e_{i,l+1} + p_{i,l+1}a_{i,l+1}) - \\
&\quad \left(e_{i,l+1} + p_{i,l+1}a_{i,l+1} + (1-p_{i,l+1})(e_{il} + p_{il}a_{il})\right) \\
&= p_{i,l+1}(e_{il} + p_{il}a_{il}) - p_{il}(e_{i,l+1} + p_{i,l+1}a_{i,l+1})
\end{aligned}
$$

Necessary condition for the maximum expected benefit of the list: $d_i^{l,l+1} \stackrel{!}{\geq} 0$, which leads to

$$
a_{il} + \frac{e_{il}}{p_{il}} \geq a_{i,l+1} + \frac{e_{i,l+1}}{p_{i,l+1}}
$$

PRP for Interactive IR

$$a_{il} + \frac{e_{il}}{p_{il}} \geq a_{i,l+1} + \frac{e_{i,l+1}}{p_{i,l+1}}$$

$\rightsquigarrow$ **Rank choices by decreasing values of**

$$\varrho(c_{ij}) = a_{il} + \frac{e_{il}}{p_{il}}$$

Discussion

- IIR-PRP vs. PRP
- Parameter estimation

IIR-PRP vs. PRP

$$a_{il} + \frac{e_{il}}{p_{il}} \geq a_{i,l+1} + \frac{e_{i,l+1}}{p_{i,l+1}}$$

Assumptions for classical PRP:

1. constant effort for each document $e_{ij} = -E, E > 0$
2. constant benefit from each relevant document $a_{il} = B$

$$\begin{aligned} B - \frac{E}{p_{il}} &\geq B - \frac{E}{p_{i,l+1}} \\ \Rightarrow p_{il} &\geq p_{i,l+1} \end{aligned}$$

$\rightsquigarrow$ Classic PRP still holds!

IIR-PRP: Observations

$$\text{Rank choices by} \quad a_{ij} + \frac{e_{ij}}{p_{ij}}$$

- p_{ij} 'probability of relevance' still involved
- tradeoff between effort e_{ij} and benefit a_{ij}
- difference between PRP and IIR-PRP due to variable values for e_{ij} and a_{ij}
- IIR-PRP looks only for the first positive decision

Parameter estimation

1. Selection probability p_{ij}:
   focus of many IR models,
   but models for dynamic info needs required
2. Effort parameter e_{ij}:
   most research needed
3. Benefit b_{ij}:
   - information value ?
   - saved effort

Conclusion and Outlook

Conclusion and Outlook

- Current IIR systems lack theoretic foundation
- Proposed IIR-PRP as a new framework model
- As with classical PRP, specific model have to be developed
- Other uses of IIR-PRP:
  - Verification of (implicit) assumptions underlying existing IIR systems
  - Definition of new evaluation metrics for IIR system components