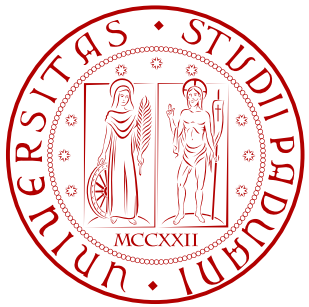
Markov Precision: Modelling User Behaviour over Rank and Time

Marco Ferrante[1] Nicola Ferro[2] Maria Maistro[2]

[1]Dept. of Mathematics, University of Padua, Italy
ferrante@math.unipd.it

[2]Dept. of Information Engineering, University of Padua, Italy
ferro@dei.unipd.it, maistro@dei.unipd.it

IIR 2015, 6th Italian Information Retrieval Workshop
Cagliari, 25–26 May 2015



Outline

Motivations and Goals

User Models

- Information Retrieval (IR) evaluation measures, implicitly or explicitly, embed a user model describing how the user interacts with the ranked result list
- User models may be more or less artificial and may be more or less correlated with actual user behaviour and preferences

User Models: Precision

Precision

The user visits exactly n ranks positions and then stops

$$Prec[n] = \frac{1}{n} \sum_{m=1}^{n} a_m$$

where $a_m = 1$, if the document at rank m is considered relevant

User Models: RBP

Rank-Biased Precision (RBP)

The user starts from the top ranked document and with probability p, called persistence, goes to the next document or with probability $1 - p$ stops

$$\mathrm{RBP} = (1 - p) \sum_{i \in \mathcal{R}} p^{i-1}$$

where $\mathcal{R}$ is the set of the ranks of the retrieved relevant documents

User Models: NCP

Normalized Cumulative Precision (NCP)

NCP [1]is the expectation (average) of the precision at the ranks of the retrieved, relevant documents, accordingly to a distribution $p_s(\cdot)$

$$\mathrm{NCP}(p_s) = \mathbb{E}_{p_s}[\mathrm{Prec}(n)] = \sum_{n=1}^{+\infty} p_s(d_n)\mathrm{Prec}(n) \ .$$

[1] S. E. Robertson, E. Kanoulas, E. Yilmaz: *A New Interpretation of Average Precision*. In SIGIR 2008, pp. 689–690. ACM Press, USA

User Models: AP

Average Precision (AP)

The user will stop his search at a given relevant document in the ranked list (satisfaction point), according to a probability law fixed and independent from the specific run he/she is considering (uniform distribution).

$$\mathrm{AP} = \frac{1}{RB} \sum_{i \in \mathcal{R}} Prec(i) = \frac{r}{RB} \cdot \frac{1}{r} \sum_{i \in \mathcal{R}} Prec(i)$$

where RB is the Recall Base, i.e. the total number of relevant documents, $\mathcal{R}$ is the set of the ranks of the retrieved relevant documents and $r = |\mathcal{R}|$ is the number of relevant retrieved documents

Criticisms to AP

AP is the "gold standard" measure in IR but its user models has some weaknesses:

- it is considered artificial, indeed the uniform distribution which determines the stopping point in a given search is of difficult interpretation, since this means that any relevant document in a ranked list of retrieved documents has the same probability;
- the probability that a user stops his/her search at a given document depends on a probability distribution defined on the whole set of relevant documents assuming a perfect knowledge of the relevance of each document in the collection.

Goals (I)

- More flexible modelling of user behaviour into an evaluation measure
  - not just one user model, even though parametric in some attribute
- Powerful and unifying framework for expressing alternative user models within evaluation measures
- Possibility of bridging towards actual user behaviour

Time

Time-Biased Gain (TBG)

$$\mathrm{TBG} = \sum_{i \in \mathcal{R}} g(i) \exp^{-\frac{\ln 2}{h} T(i)}$$

where $g(i)$ is the gain value of each document, h is a discounting function, $T(i)$ is the expected time spent at each rank i.

- While the time users spend in inspecting and interacting with the result list is a key dimension, it is often overlooked in traditional IR evaluation measures
- TBG focuses on how much time the user spends on each document in order to accumulate gain over rank positions
- TBG embeds a user model similar to the one of RBP and it is complementary to the standard IR measures

Goals (II)

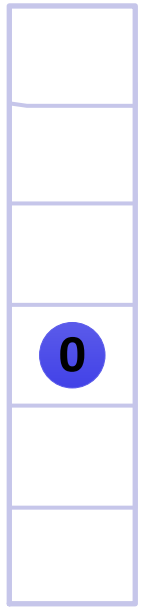
- Unify rank-based and time-based evaluation measures into a single coherent framework
  - avoid to resort to pairs of measures to grasp different angles
- Possibility of embedding alternative user models into time-based evaluation measure

A Markovian Approach to Evaluation Measures

Markovian User Model

We assume that each user:

- starts from a chosen document at rank X_0 and considers this document for a random time, T_0;
- then he/she decides, according to a probability law independent of the random time spent in the first document, to move to another document in the list at position X_1; he/she considers this new document for a random time T_1;
- then he/she moves, independently, to a third relevant document, and so on.



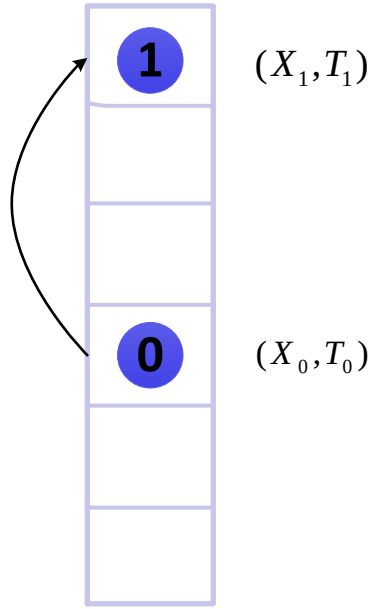
(X_0, T_0)

After a random number of forward or backward movements along the ranked list, the user will end his/her search and we will evaluate the total utility provided by the system to him/her.

Markovian User Model

We assume that each user:

- starts from a chosen document at rank X_0 and considers this document for a random time, T_0;
- then he/she decides, according to a probability law independent of the random time spent in the first document, to move to another document in the list at position X_1; he/she considers this new document for a random time T_1;
- then he/she moves, independently, to a third relevant document, and so on.

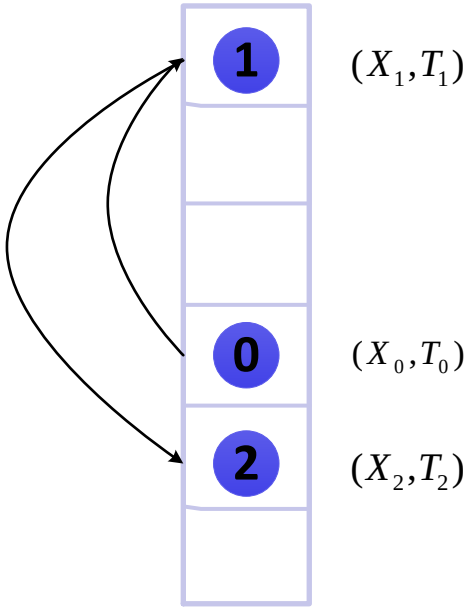


After a random number of forward or backward movements along the ranked list, the user will end his/her search and we will evaluate the total utility provided by the system to him/her.

Markovian User Model

We assume that each user:

- starts from a chosen document at rank X_0 and considers this document for a random time, T_0;
- then he/she decides, according to a probability law independent of the random time spent in the first document, to move to another document in the list at position X_1; he/she considers this new document for a random time T_1;
- then he/she moves, independently, to a third relevant document, and so on.



After a random number of forward or backward movements along the ranked list, the user will end his/her search and we will evaluate the total utility provided by the system to him/her.

Mathematical Framework

Notation:

- $X_0, X_1, X_2, \ldots \in \mathcal{T} = \{1, 2, \ldots, T\}$ the random sequence of document ranks visited by the user;
- $T_0, T_1, T_2, \ldots$ the random times spent, respectively, visiting the first document considered, the second one and so on.

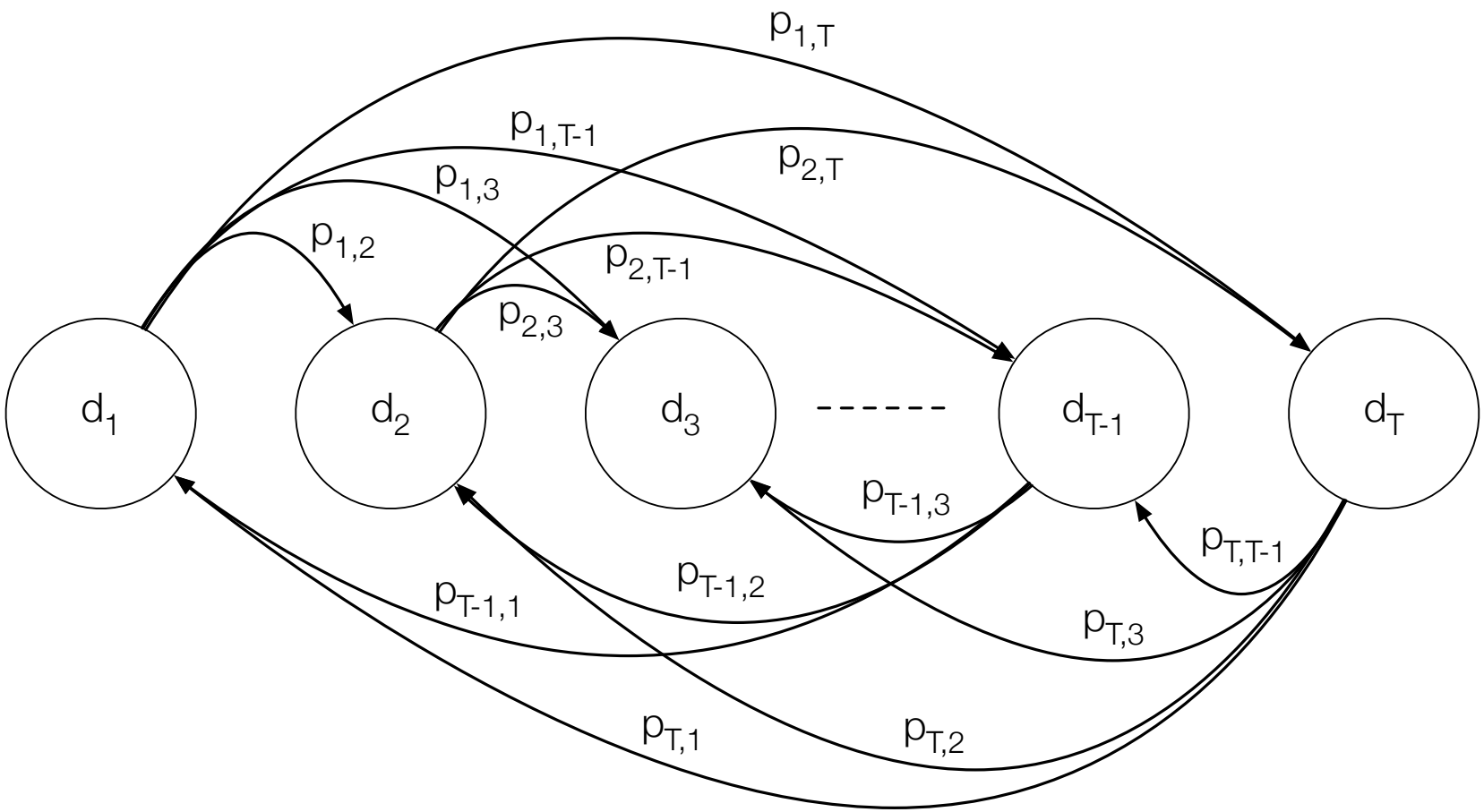
Assumption: the probability to pass from the document at rank i to the document at rank j will only depend on the starting rank i and not on the whole list of documents:

$$\mathbb{P}[X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0] =$$

$$= \mathbb{P}[X_{n+1} = j | X_n = i] = p_{i,j}$$

for any $n \in \mathbb{N}$ and $i, j, i_0, \ldots, i_{n-1} \in \mathcal{T}$.

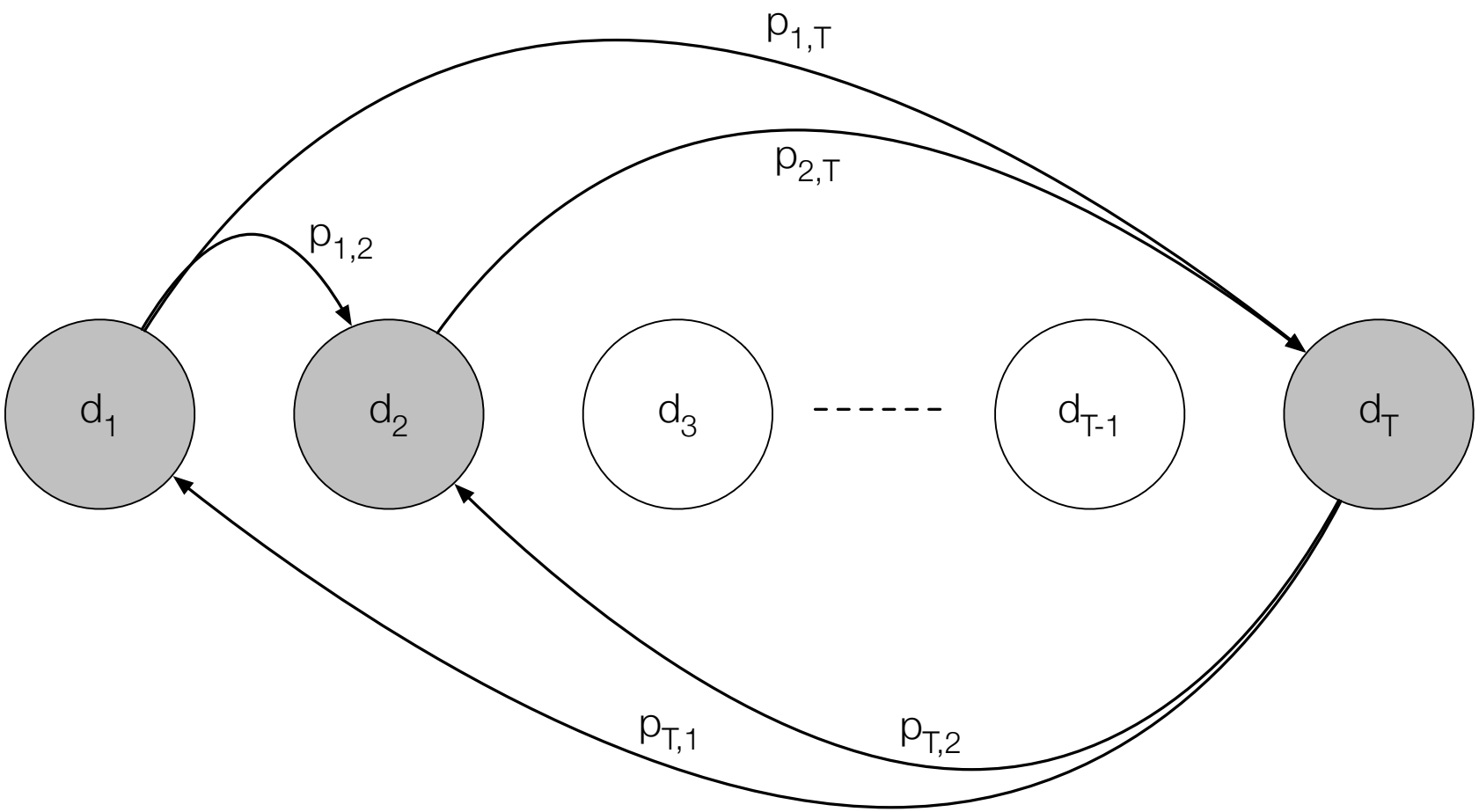
Structure of the Markov Chain



Definition of the Markov Chain

- Fixing a starting distribution λ, the random variables $(X_n)_{n \in \mathbb{N}}$ define a time homogeneous discrete time Markov Chain, with state space $\mathcal{T}$, initial distribution λ and transition matrix $P = (p_{i,j})_{i,j \in \mathcal{T}}$.
- Under the assumption that P is irreducible, i.e. the user can move in a finite number of steps from any document to any other document with positive probability, there exist a unique invariant distribution π, independent from the initial distribution λ, which provides us with the probability of the user being in any given document in the long run, i.e. $n \to \infty$.
  - in practice, the convergence is quite fast and $n \sim 10$ is usually enough
- The invariant distribution is the unique left eigenvector of the eigenvalue 1 of the transition matrix and can be computed solving the linear system $\pi = \pi P$.

Structure of the sub-Markov Chain



Discrete Time Markov Precision

Let $(Y_n)_{n\in\mathbb{N}}$ denotes the sub-chain of $(X_n)_{n\in\mathbb{N}}$ that considers just the visits to the judged relevant documents at ranks $\mathcal{R}$. Note that this sub-chain has in general a transition matrix $\widetilde{P}$ different form P, which can be computed from P by solving a linear system. We define Markov Precision (MP) as:

$$MP = \sum_{i\in\mathcal{R}} \pi_i \text{Prec}(i).$$

i.e. the weighted mean of the precision, at each relevant retrieved document, by the probability of the user visiting that document.

Markov Precision: Considerations

- MP does not depend on the recall base, a knowledge not available to users;
- MP allows users to move forward and backward in the result list;
- The transition matrix allows for plugging several alternative user models into MP, either by pre-defining them or learning them from logs and user interaction data this also allows to compare different user models within the same measure rather than comparing different measures with different user models each other;
- The (fast converging) invariant distribution frees us from assuming that the user starts browsing from a specific document, typically the first one;
- MP relies on the assumption of a memory-less Markovian process this differs from measures like Expected Reciprocal Rank (ERR), which takes into account the history of all the visited documents, but it is similar to measures like RBP, where the decision on whether continuing or not to visit the list is independent from previously visited documents.

Markov Precision and NCP

Recall the NCP model:

$$\mathrm{NCP}(p_s) = \mathbb{E}_{p_s}[\mathrm{Prec}(n)] = \sum_{n=1}^{+\infty} p_s(d_n)\mathrm{Prec}(n) \ .$$

MP fits this model if you use the invariant distribution as probability distribution:

$$p_s(d_n) = a_n \pi_n$$

where $a_n = 1$, if the document at rank n is considered relevant

AP and MP

Consider the following transition probabilities among the relevant documents in a given ranked list:

$$\mathbb{P}[X_{n+1} = j | X_n = i] = \frac{1}{r-1} \quad i, j \in \mathcal{R}, i \neq j$$

where $r = |\mathcal{R}|$ is the number of the relevant retrieved documents.

Then the invariant distribution is $\left(\frac{1}{r}, \frac{1}{r}, \ldots, \frac{1}{r}\right)$ and we obtain

$$MP = \frac{1}{r} \sum_{i \in \mathcal{R}} Prec(i)$$

which is equal to AP once multiplied by the Recall $\frac{r}{RB}$.

AP Markov Model: Considerations

- We explain AP with a slightly richer user model, where the user can move forward and backward among any document;
- MP is not AP unless you provide it with the same amount of information AP knows about the recall base;
- As we will see in the evaluation part, other MP user models are extremely highly correlated with AP, providing further alternative explanations of it.

Continuous time Markov Precision

To obtain a continuous-time Markov Chain, we have to assume that the holding times T_n have all exponential distribution, i.e.

$$\mathbb{P}[T_n \leq t] = \begin{cases} 0 & t < 0 \\ 1 - \exp(-\mu t) & t \geq 0 \end{cases}$$

where μ_i is a positive real number that may depend on the specific state i of the chain the user is visiting at that time.
Let $(X_t)_{t \geq 0}$ be the continuous Markov Chain, with jump chain $(X_n)_{n \in \mathbb{N}}$ and holding times T_n, then the Continuous-Time Markov Precision is:

$$MPcont = \sum_{i \in \mathcal{R}} \widetilde{\pi}_i \text{Prec}(i).$$

where

$$\pi_i = \frac{\pi_i \mu_i^{-1}}{\sum_{j \in \mathcal{R}} \pi_j \mu_j^{-1}}$$

Experimental Evaluation

Experimented User Models

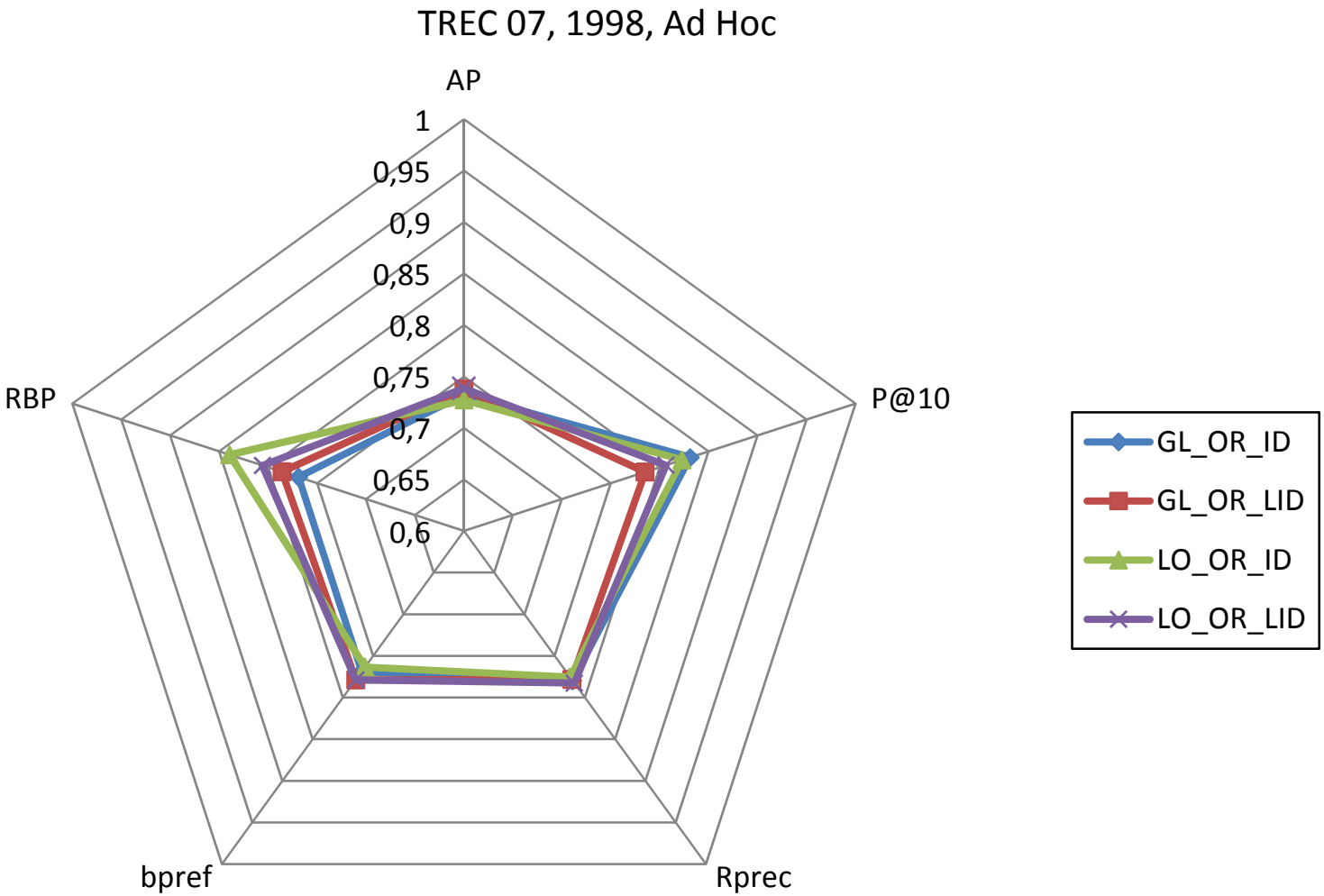
We will analyse three possible choices:

- state space choice: the state space of the Markov chain $(X_n)_{n \in \mathbb{N}}$ is the whole set $\mathcal{T}$, indicated with `AD` (all documents model), or the set $\mathcal{R}$, indicated with `OR` (only relevant documents model);
- connectedness: the nonzero transition probabilities are among all the documents, indicated with `GL` (global model), or only among adjacent documents, indicated with `LO` (local model);
- transition probabilities: the transition probabilities are proportional to the inverse of the distance, indicated with `ID` (inverse distance model), or to the inverse of the logarithm of the distance, indicated with `LID` (logarithmic inverse distance model).

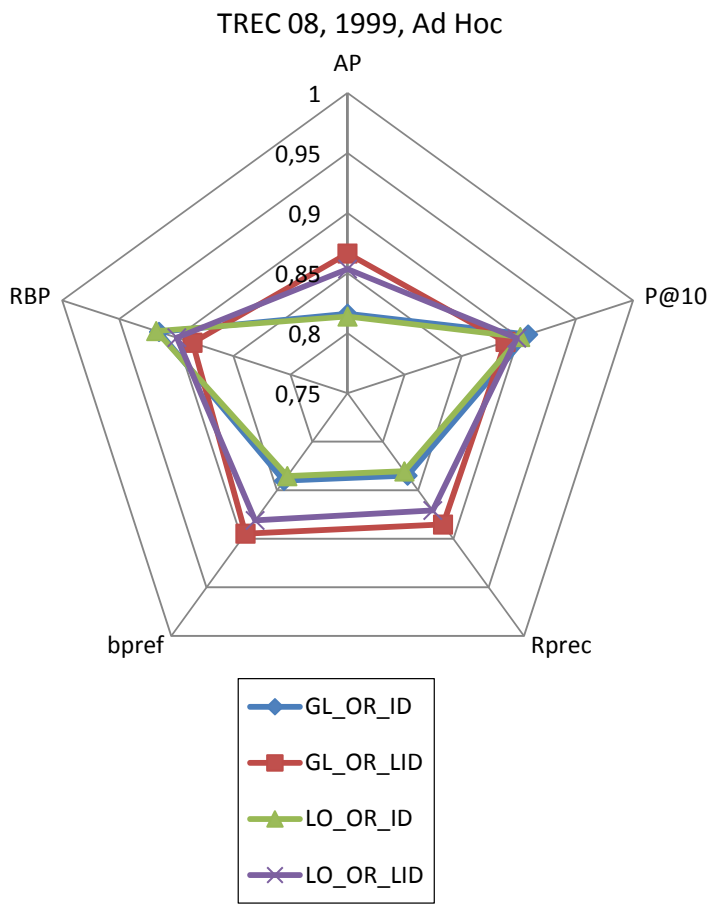
Experimental Collections

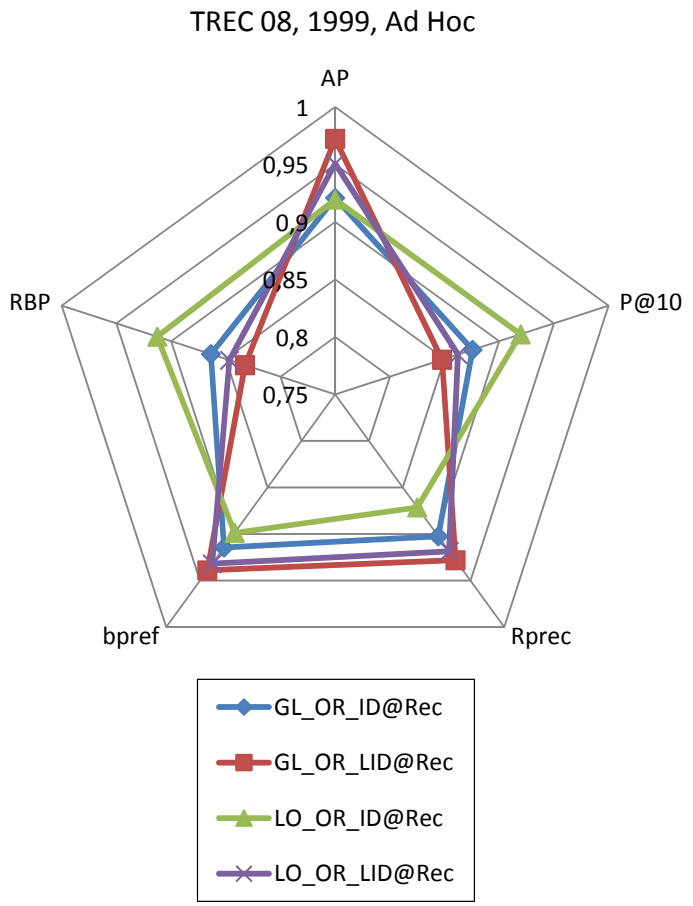
Feature	TREC 7	TREC 8	TREC 10	TREC 14
Track	Ad Hoc	Ad Hoc	Web	Robust
Corpus	Tipster 4, 5	Tipster 4, 5	WT10g	AQUAINT
# Documents	528K	528K	1.7M	1.0M
# Topics	50	50	50	50
# Runs	103	129	95	74
Run Length	1,000	1,000	1,000	1,000
Relevance Degrees	Binary	Binary	3 Grades	3 Grades
Pool Depth	100	100	100	55
Minimum # Relevant	7	6	2	9
Average # Relevant	93.48	94.56	67.26	131.22
Maximum # Relevant	361	347	372	376

Correlation Analysis

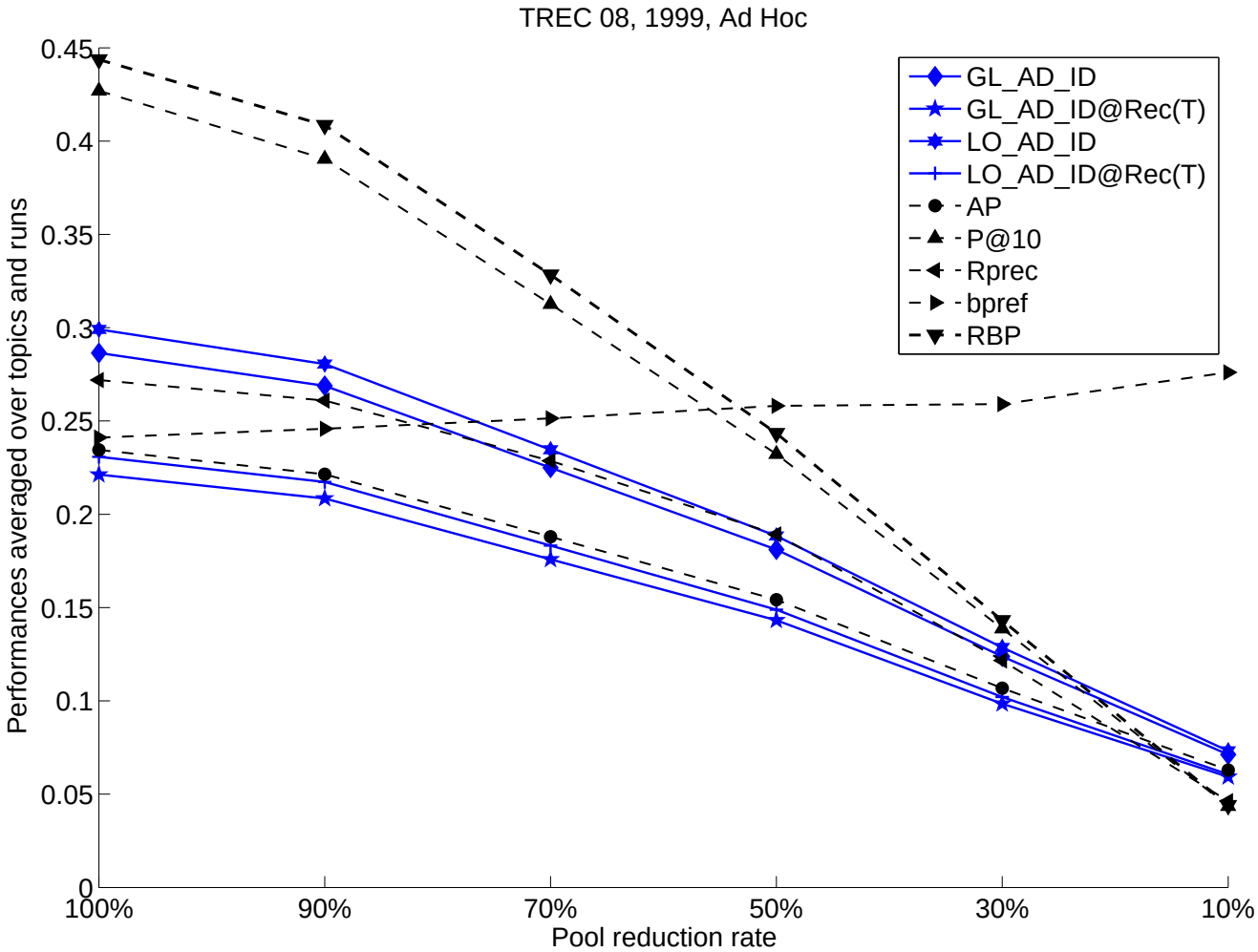


Correlation Analysis

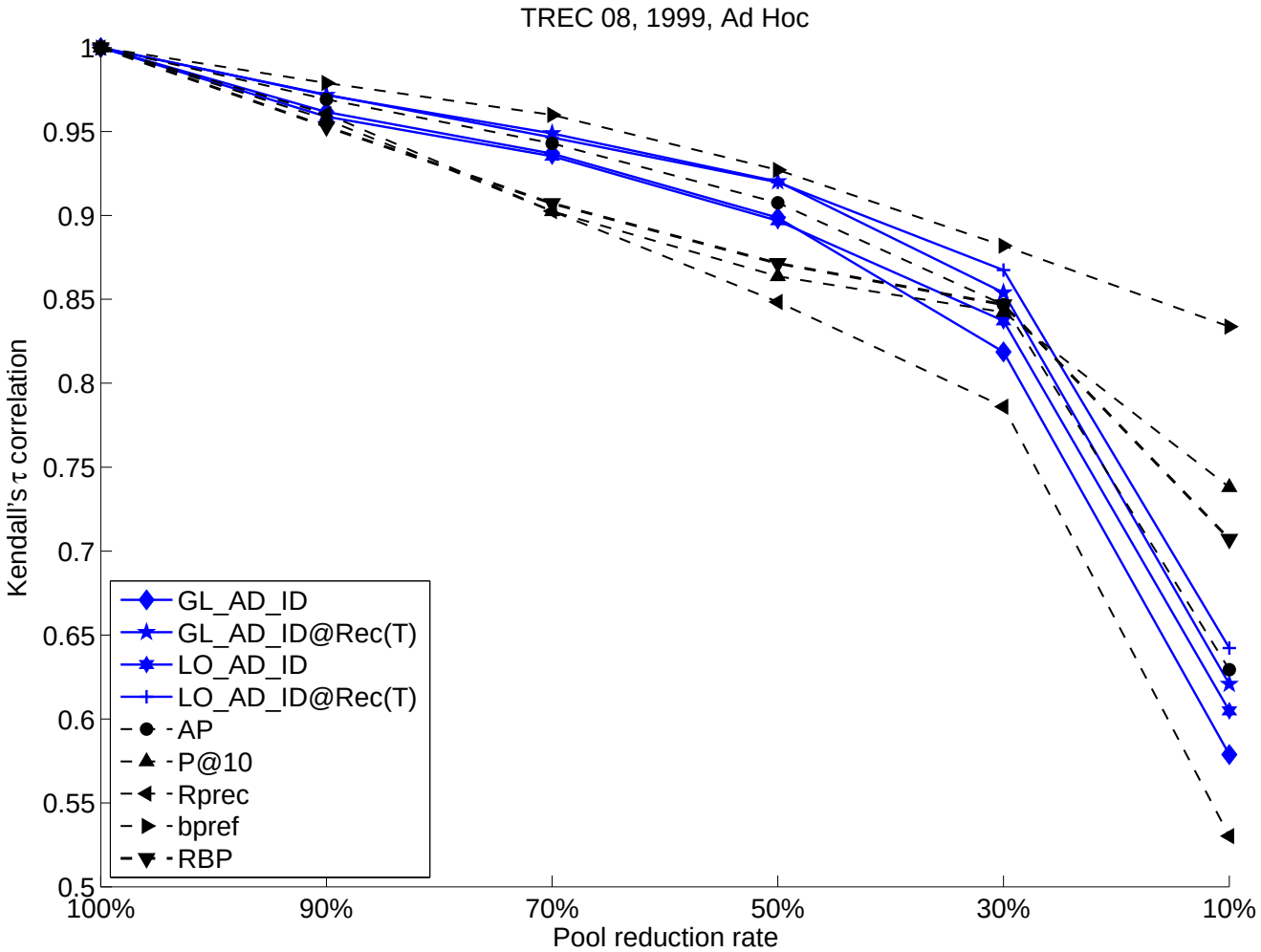




Effect of Incompleteness on Absolute Performances



Effect of Incompleteness on Rank Correlation



Click-Log Data Set

- Click logs made available by Yandex in the context of the Relevance Prediction Challenge;
  - `http://imat-relpred.yandex.ru/en/`
- 340,796,067 records with 30,717,251 unique queries, retrieving 10 URLs each;
  - the training set contains 5,191 assessed queries which correspond to 30,741,907 records;
  - we selected those queries which appear at least in 100 sessions each to calibrate the time.
- On the basis of the click logs, 21% of the observed transitions are backward.

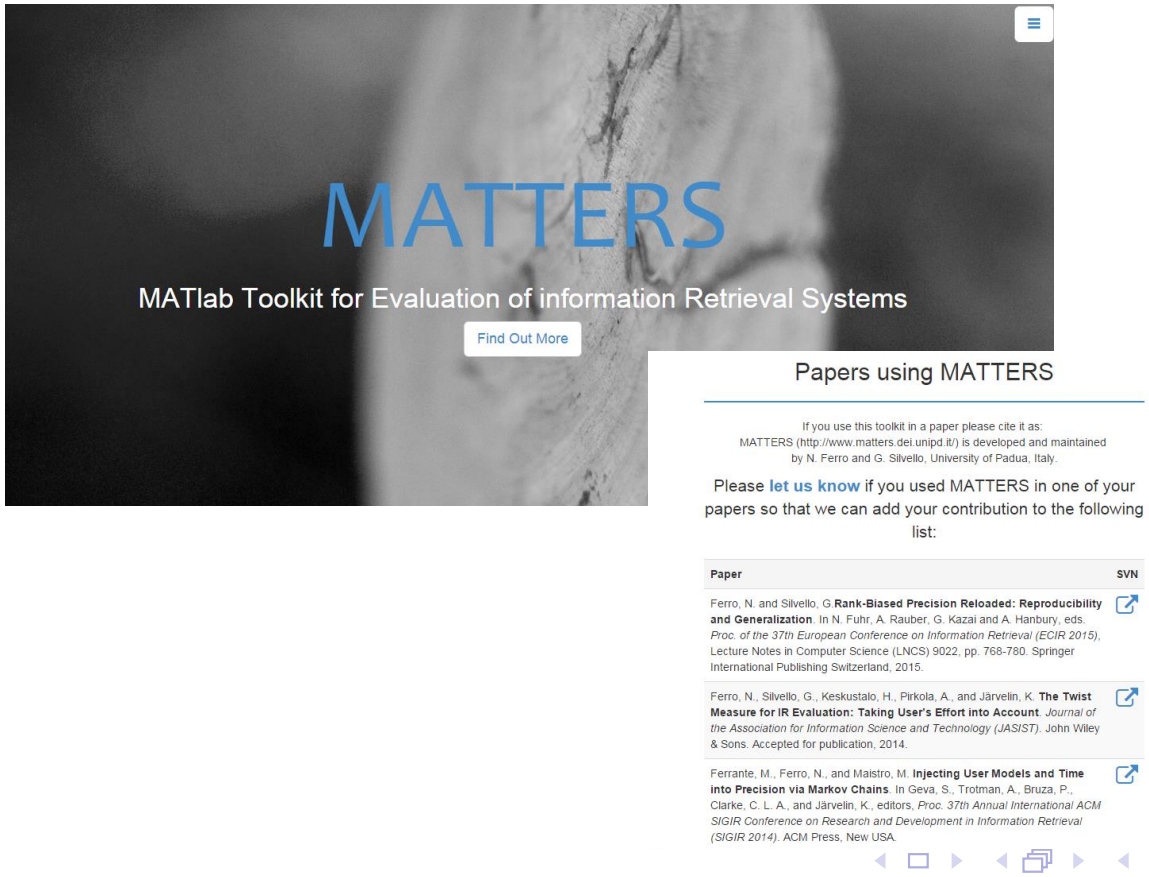
Time Calibration

- We estimated the parameters of the exponential holding times by using an unbiased estimator of the inverse of the sample mean of the time spent by the users visiting these states;
- We used the `GL_AD_ID` model, i.e. considering all the retrieved documents (`AD`), allowing for transition among all of them (`GL`), and with transition probability proportional to the inverse of the distance (`ID`).

Run	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	disc MP	cont MP
(1,1,1,1,0,0,0,1,0,0)	0.2000	0.0357	0.2000	0.0400	0.0056	0.0005	0.0035	0.0017	0.0034	0.0024	0.9205	0.6603
(1,1,1,0,1,0,0,0,1,0)	0.0177	0.0047	0.0037	0.0015	0.0041	0.0031	0.0057	0.0022	0.0061	0.0045	0.8668	0.8710
(1,1,0,1,1,0,0,0,0,1)	0.0056	0.0051	0.0062	0.0031	0.0046	0.0025	0.005	0.0022	0.007	0.005	0.8120	0.8001

Reproducibility

Source code for running the experiments is available at:
http://matters.dei.unipd.it/



Conclusions

We introduce a new measure, Markov Precision (MP), which models the user interaction with the result list via Markov chains:

- it provides a single and coherent framework for plugging alternative user models into the same measure;
- it allows for dealing with both rank-based and time-based evaluation of IR systems;
- it provides us with alternative interpretations of average precision.

Future Work

Markov Precision just scratches the "top of the iceber":

- study different user models, also learned from logs, together with their time-based calibration;
- address the memory-less property of Markov chains in order to take into account also the previously visited documents;
- study its relationship with click-based measures;
- further study of its properties, e.g. discriminative power.


THANK
YOU