Markov Precision: Modelling User Behaviour over Rank and Time

Extended Abstract of [1]

Marco Ferrante[1], Nicola Ferro[2], and 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

Abstract. We propose a family of new evaluation measures, called Markov Precision (MP), which exploits continuous-time and discrete-time Markov chains and we conduct a thorough experimental evaluation providing also an example of calibration of its time parameters.

1 Introduction

We propose a family of measures of retrieval effectiveness, called Markov Precision (MP) [1] where we exploit Markov chains [2] to inject different user models into precision. We represent each position in a ranked result list with a state in a Markov chain and the different transition probabilities among the states allow us to model the perhaps complex user paths in scanning a ranked result list. The framework we propose is actually more general and it is based on continuous-time Markov chains in order to take into account also the time a user spends in visiting a single document. This gives us a two-fold opportunity: when we consider the discrete-time Markov chain, we are basically reasoning as traditional evaluation measures which assess the utility for the user in scanning the ranked result list; when we consider the continuous-time Markov chain, we also embed the information about the time spent by the user in visiting a document.

Finally, we conduct a thorough experimental evaluation of the MP measure both using standard Text REtrieval Conference (TREC)[3] collections and click-logs with assessed queries made available by Yandex [3]. The results show that MP is comparable to other measures, while the Yandex click-logs allow us to estimate the time spent by the users on the documents.

2 A Markovian User Model

Firstly we will introduce some notation that we will use through the whole paper. Let us consider a ranked list of T documents and let $\mathcal{R}$ be the set of the ranks of the relevant documents. We will denote by RB the recall base, i.e. the total number of judged relevant documents.

[3] http://trec.nist.gov/

We will assume that each user starts from a chosen document, at rank X_0 in the list, and considers this document for a random time T_0, that is distributed according to a known positive random variable. Then he/she decides to move to another document, at rank X_1, and he/she considers this new document for a random time T_1. Successively, he/she moves, independently, to a third document and so on. Hence, we will denote by $X_0, X_1, X_2, \ldots$ the (random) sequence of document ranks visited by the user and by T_0, T_1, T_2 the random times spent visiting each considered document.

We mathematically model the user behaviour in the framework of the Markovian processes [2]. First of all, we will assume that X_0 is a random variable on $\mathcal{T} = \{1, 2, \ldots, T\}$ with a given distribution $\lambda = (\lambda_1, \ldots, \lambda_T)$; so for any $i \in \mathcal{T}$, $\mathbb{P}[X_0 = i] = \lambda_i$. Then, we will assume that 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 visited before. Thanks to this condition and 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.

To obtain a continuous-time Markov Chain, we have to assume that the holding times T_n have all exponential distribution, and conditioned on the fact that $X_n = i$, the law of T_n will be exponential with parameter μ_i, where μ_i is a positive real number. When our interest is only on the jump chain $(X_n)_{n\in\mathbb{N}}$, we simply assume that all these variables are exponential with parameter $\mu = 1$; while when we are also interested in the time dimension, we have to provide a calibration for these exponential variables.

Let us assume hereafter that the matrix P will be irreducible and that after visiting n documents in the list the user will stop his/her search. In order to measure his/her satisfaction, we will evaluate the average of the precision of the ranks of the judged relevant documents visited by the user during their search as

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathrm{Prec}(Y_k) \ .$$

where $(Y_n)_{n\in\mathbb{N}}$ 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 different form P, that we will denote with $\widetilde{P}$.

Clearly, the previous quantity is of little use if evaluated at an unknown finite step n. However, the Ergodic Theorem for the Markov processes approximates this quantity with

$$MP = \sum_{i\in\mathcal{R}} \pi_i \mathrm{Prec}(i) \quad ,$$

where π is the (unique) invariant distribution of the Markov chain $(Y_n)_{n\in\mathbb{N}}$. Hence, we have defined a new family of user oriented retrieval effectiveness measures, called Markov Precision (MP). Note that MP is defined without knowing the recall base RB, but just the ranks of the judged relevant documents in the given run.

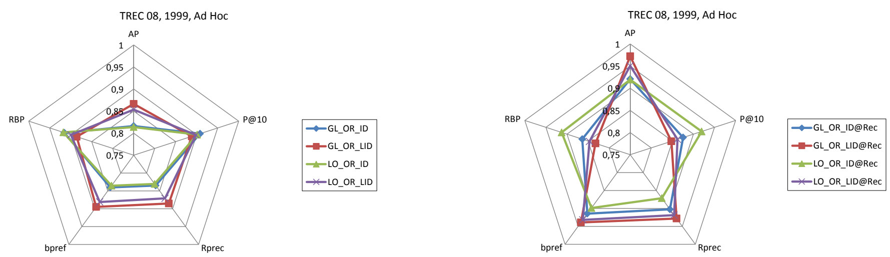

Fig. 1. Kendall τ correlation between different instantiations of MP and the other comparison measures using complete judgements.

In order to include the time dimension, we can replicate the previous computations and define a new measure

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

where $\widetilde{\pi}_i = \frac{\pi_i(\mu_i)^{-1}}{\sum_{j\in\mathcal{R}} \pi_j(\mu_j)^{-1}}$, π denotes again the (unique) distribution of the Markov chain $(Y_n)_{n\in\mathbb{N}}$ and μ_i is the parameter of the holding time in state i.

3 Evaluation

In order to assess MP and compare it to the other evaluation measures (Average Precision (AP), P@10, Rprec, Rank-Biased Precision (RBP), and Binary Preference (bpref)), we conducted a correlation analysis and we studied its robustness to pool downsampling. We used the following data sets: TREC 7 Ad Hoc, TREC 8 Ad Hoc, TREC 10 Web, and TREC 14 Robust. As far as calibration of time is concerned, we used click logs made available by Yandex [3]. The full source code of the software used to conduct the experiments is available for download[4] in order to ease comparison and verification of the results.

Firstly, we computed the Kendall τ correlation between the different models for MP and the performance measures of direct comparison, for all the considered collections; Figure 1 reports the results in the case of TREC 8. We indicate by [OR] the model where the user moves only on relevant documents, [GL] means that the user moves among all the documents, [LO] only among adjacent documents. [ID] denotes that the transition probabilities are proportional to the inverse of the distance, or to the inverse of the logarithm of the distance, indicated with [LID].

As a general trend MP tends not to have high correlations with the other evaluation measures, even if the correlation never drops below 0.70, indicating

[4] http://matters.dei.unipd.it/

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

Table 1. Estimated parameters of the exponential holding times for three runs and values of the discrete-time and continuous-time MP.

that it takes a different angle from them (users move forward and backward in the result list). With regard to the rescaled version of MP by recall (@Rec), the correlation with AP increases in almost all cases. Moreover, if we assume that a user moves with a constant, fixed transition probability, the number of which depends on the number of retrieved relevant document, we will obtain that MP is equal to AP, once rescaled by the recall.

Then we analyse the effect of reducing the pool size on the absolute average performances, over all the topics and runs. MP shows a consistent behaviour over all the collections and for various models: its absolute average values decrease as the pool reduction rate increases.

Furthermore, we study the effect of reducing the pool size on the Kendall τ correlation between each measure on the full pool and the pool at a given reduction rate. MP models tend to perform comparably to AP and, when provided with the same information about the recall base, they consistently improve their performances. For all models, using the log of the inverse of the distance [`LID`] provides more robustness to pool reduction.

Finally, on the basis of the click logs, we can state that 21% of the observed transitions are backward, a fact that validates our assumption that a user moves forward and backward along the ranked list. Moreover, we compute the values of continuous-time MP and we compare them with the discrete-time MP, reported in Table 1, concluding that the continuous-time version depends heavily on the calibration of the holding times.

4 Future Work

Future works concern the investigation of alternative user models able to account also for the number of relevant/not relevant documents visited so far, the possibility of learning the transition probabilities of the Markov chain directly from click-logs, the calibration of time into MP, the adjustment of MP to fact, entity, or attributes focused searches, and the investigation of the robustness of MP.

References

1. M. Ferrante, N. Ferro, M. Maistro. Injecting User Models and Time into Precision via Markov Chains. In SIGIR 2014, pages 597-606. ACM Press, USA. (2014)
2. J. R. Norris. *Markov chains.* Cambridge University Press, UK, 1998.
3. P. Serdyukov, N. Craswell, and G. Dupret. WSCD2012: Workshop on Web Search Click Data 2012. In WSDM, pages 771–772. ACM, 2012.