# Improving Information Retrieval Evaluation via Markovian User Models and Visual Analytics

Maria Maistro
Department of Information Engineering
University of Padua, Italy
*maistro@dei.unipd.it*

**To address the challenge of adapting experimental evaluation to the constantly evolving user tasks and needs, we develop a new family of Markovian Information Retrieval (IR) evaluation measures, called Markov Precision (MP), where the interaction between the user and the ranked result list is modelled via Markov chains, and which will be able to explicitly link lab-style and on-line evaluation methods. Moreover, since experimental results are often not so easy to understand, we will develop a Web-based Visual Analytics (VA) prototype where an animated state diagram of the Markov chain will explain how the user is interacting with the ranked result list in order to offer a support for a careful failure analysis.**

*Evaluation, Markov Precision, User Model, Visual Analytics*

## 1. INTRODUCTION

Nowadays information and its retrieval are fundamental and pervasive in everyday life of each person and the methods of accessing it and the systems themselves are changing rapidly. Hence, the quantity and heterogeneity of available information is rapidly increasing as well as the complexity of user tasks and needs are performing. This calls for increasingly sophisticated IR methods and systems which, in turn, need advanced evaluation techniques to be properly conceived, designed and developed.

In particular, IR systems operate using a best match approach: in response to an often vague user query, they return a ranked list of documents ordered by the estimation of their relevance to that query. In this context effectiveness, meant as "the ability of the system to retrieve relevant documents while at the same time suppressing the retrieval of non-relevant documents" (Rijsbergen 1979), is the primary concern. Since there are no a-priori exact answers to a user query, experimental evaluation based on effectiveness is the main driver of research and innovation in the field. Indeed, the measurement of system performances from the effectiveness point of view is basically the only mean to determine which are the best approaches and to understand how to improve IR systems.

Today the available evaluation methods can be divided in two distinct categories: the batch and the on-line methods. The batch methods are based on models that consider a hypothetical user and depict his/her behaviour in an abstract way. Hence, the user does not interact with the system and the system is evaluated through controlled experiments. The main disadvantage of these methodologies is that they are mostly focused on the algorithmic and system side and they may be somewhat "artificial" by abstracting away too much of the user interaction. On the other hand, these methods have the advantage of being reproducible and scalable.

The other family of evaluation methods, the on-line methods, bases its strategy on user studies and analysis of interaction data, such as search logs, to investigate and consequently connect the behaviour of the user with the system. These methods undoubtedly have the advantage of taking into account the user's needs and to interact with him/her, but they are more expensive because they require the involvement of real users, they are time consuming and they cover many different disciplines. Furthermore on-line methods are not easily reproducible or scalable.

In batch evaluation, many different measures, e.g. precision and recall, have been created to determine in a rigorous way when one ranked list of documents is better sorted than another one. However, if the value of precision or the


*Proceedings of the 6th Symposium on Future Directions in Information Access 2015.*

value of recall increase, does user satisfaction also increase? It is therefore fundamental to envision new evaluation methodologies capable of linking on-line and batch strategies and of providing a better fit with actual user needs and behaviour. This is needed to provide a more accurate estimation of system performances, which is crucial to cope with ever increasing information resources and rapidly evolving user tasks.

Moreover, both batch and on-line evaluation methods produce huge amounts of experimental and scientific data from which it is not so obvious how to infer useful information. Statistical tools (Savoy 1997) and other recent techniques, for instance VA (Keim et al. 2010), can play a key role in coping with and understanding such large amounts of experimental data. Thus, new evaluation methodologies should also comprise powerful VA techniques to support researchers and developers in analysing, exploring and understanding experimental results in order to more effectively improve IR systems.

## 2. OBJECTIVES

As previously discussed, there are two methods that allow for the evaluation of an IR system: batch and on-line methods. An evaluation metric including both these aspects represents an innovation with respect to the state-of-the-art because:

- up to now there is not a measure capable of connecting the lab-style and the on-line evaluation methods and to merge them in a single tool, also accounting for the time dimension;
- from the scientific point of view this will provide the possibility of fostering an in-depth analysis of user behaviour models;
- From the engineering point of view, a more powerful measure, able to better grasp and explain the interaction between the system and its users, will provide a valuable support for the design and the development of next generation IR systems.

To reach the purpose of defining a class of measures which can be used with both batch and on-line strategies, we plan to rely on the Markov chains framework (Norris 1998) as proposed in our work (Ferrante et al. 2014). Furthermore, regarding the management and the visualisation of the experimental data we will use VA techniques, which are a quite new idea to the IR field (Angelini et al. 2014), and which allow the experimental results to be more efficiently and effectively explained.

Therefore, we can summarize the aim of this research project in four main objectives:

**Definition of the Markov measures** : a new family of metrics based on Markov chains that can be used both with the batch and the on-line evaluation methods;

**Analysis of the properties of the Markov measures** both from the mathematical and experimental point of view;

**Design and development of a prototype web application** that uses techniques of VA to represent experimental results;

**Evaluation of the web prototype** with domain experts and examples of its possible applications.

## 3. METHODOLOGIES

In this Section we will describe the methodologies adopted to reach the four principal objectives and the first research results.

### 3.1. Definition of the Markov Measures

Even though Markov chains are an intuitive and robust tool, up to now they have never been applied to the field of evaluation in IR, except for our paper (Ferrante et al. 2014). In particular, we define a new class of evaluation measures called MP, by representing each position in a ranked result list with a state in a Markov chain and the transition probabilities among the states allow us to model the different and complex user interaction in scanning the ranked result list, e.g. forward and backward movements or jumps.

Firstly we introduce some notation that we use through this section. Let us consider a ranked list of $T$ documents, let $\mathcal{R}$ be the set of the ranks of the relevant documents and $RB$ the recall base, i.e. the total number of judged relevant documents. We 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 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 by assuming 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 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 $\lambda$, the random variables $(X_n)_{n\in\mathbb{N}}$ define a time homogeneous discrete time Markov Chain, with state space $\mathcal{T}$, initial distribution $\lambda$ 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 $\mu_i$, where $\mu_i > 0$. 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 $P$ is 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 evaluate the average of the precision, computed at the ranks of the relevant documents visited by the user, as

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

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 $\mathcal{R}$. Clearly, this 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

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

where $\pi$ is the (unique) invariant distribution of the Markov chain $(Y_n)_{n\in\mathbb{N}}$. Note that MP is defined without knowing the recall base $RB$, but just the ranks of the judged relevant documents in the given run.

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

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

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

Therefore, 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 (batch measure), while when we consider the continuous-time Markov chain, we also embed the information about the time spent by the user in visiting a document(on-line measure).

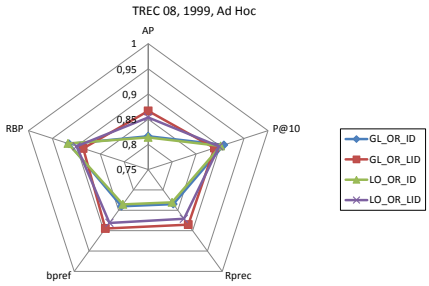
### 3.2. Analysis of the Properties of the Markov Measures

From the mathematical point of view we concentrate on the analysis of the invariant distribution of Markov chains. We will study how the shape of the invariant distribution depends on the relative position of the relevant documents. For instance, if in a ranked result list the order of some documents is modified we would like to approximate this change with a mathematical estimate which can predict the tendency of both the invariant distribution and the value of the measure. This estimate would be useful to tackle the significant problem represented by the cost of the experimental campaigns; if it were possible to judge only a small part of the great number of documents and to predict how the lack in the relevance judgements would affect the measure, then the employment of the resources would be less expensive.

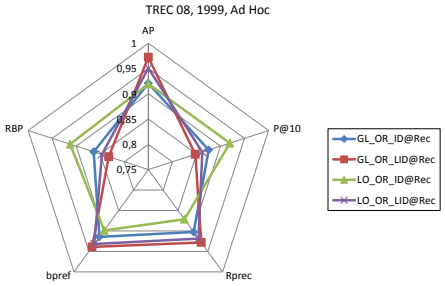
From the experimental point of view we compare MP to the other evaluation measures (Average Precision (AP), P@10, Rprec, Rank-Biased Precision (RBP), and Binary Preference (bpref)); we conducted a correlation analysis and we studied its robustness to pool downsampling on the following data sets: TREC 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 (Serdyukov et al. 2012). The full source code of the software used to conduct the experiments is available for download[1] in order to ease comparison and verification of the results.

Firstly, we computed the Kendall $\tau$ correlation (Kendall 1945) between MP and the performance measures of direct comparison. As a general trend MP tends not to have high correlations with the other evaluation measures (Figure 1a), indicating that it takes a different angle from them (users move forward and backward in the result list); but if we provide MP with the same amount of information AP has, i.e. we rescale MP by recall, the correlation with AP increases in almost all cases (Figure 1b). Then we analyse the effect of reducing the pool size on the absolute average performances and on the Kendall $\tau$ correlation. Concerning the absolute average performances MP shows a consistent behaviour over all

[1] `http://matters.dei.unipd.it/`




***(a)*** *Kendall $\tau$ correlation between different instantiations of MP and the other comparison measures.*




***(b)*** *Kendall $\tau$ correlation between different instantiations of rescaled MP and the other comparison measures.*

the collections: its absolute average values decrease as the pool reduction rate increases. If we consider the effect on the correlation,MP models tend to perform comparably to AP and, when provided with the same information about the recall base, they consistently improve their performances.

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 compared the values of continuous-time MP and discrete-time MP, concluding that the continuous-time version depends heavily on the calibration of the holding times.

### 3.3. Development of a Web Prototype

As mentioned in Section 1, experimental evaluation generates huge amount of scientific data that need to be effectively and efficiently analysed, explored and understood. We plan to develop a prototype web application that allows the user not only to visualize the experimental results but to interact with them too. We will use VA tools (Keim et al. 2010), which integrate the user in the data mining process, and, through an efficient visualization of the information, he/she can interact with, modify and enhance the analysis of the data in order to detect the weak points of his/her system and to improve it.

Concerning the architecture of the prototype the Markovian user models and measures, as well as the data mining steps, will be performed using Matlab on the server side while, on the client side, an advanced web application will exploit information visualisation and VA techniques in order to engage and provide the user with an intuitive representation of the experimental results.

## 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 investigation of the robustness of MP e.g. discriminative power (Sakai 2006), and the developmet of the web prototype.

## 5. ACKNOWLEDGMENTS

This work has been partially supported by ELIAS, a research networking programme (09-RNP-085) project (`http://elias-network.eu`) of the European Science Foundation (ESF).

## REFERENCES

M. Angelini, N. Ferro, G. Santucci, G. Silvello, *VIRTUE: A Visual Tool for Information Retrieval Performance Evaluation and Failure Analysis*. In Journal of Visual Languages and Computing, volume 25, issue 4, pages 394–413, August 2014.

M. Ferrante, N. Ferro, M. Maistro, *Injecting User Models and Time into Precision via Markov Chains*. Proc. 37th Annual International ACM SIGIR, 2014, pages 597-606. ACM Press, New York, USA.

D. Keim, J. Kohlhammer, G. Ellis and F. Mansmann, *Mastering the Information Age Solving Problems with Visual Analytics*. Eurographics Association, Germany, 2010.

M. G. Kendall, *The Treatment of Ties in Ranking Problems*. Biometrika, 33(3):239–251, 1945.

A. Moffat and J. Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM TOIS*, 27(1):2:1–2:27, 2008.

J. R. Norris. *Markov chains*. Cambridge University Press, UK, 1998.

C. J. Van Rijsbergen, *Information Retrieval*. Butterworth-Heinemann Newton, USA, 1979.

T. Sakai, *Evaluating Evaluation Metrics based on the Bootstrap*. SIGIR 2006, pages 525–532.

J. Savoy, *Statistical Inference in Retrieval Effectiveness Evaluation*. Information Processing and Management, Vol. 33, No. 4, pages 495–512, 1997.

P. Serdyukov, N. Craswell, and G. Dupret. WSCD2012: Workshop on Web Search Click Data 2012. In WSDM, pages 771–772. ACM, 2012.