A Probabilistic Automaton for the Dynamic Relevance Judgements of Users

Peng Zhang, Ulises Cerviño Beresi
School of Computing
The Robert Gordon University
United Kingdom
p.zhang1, prs.cervino-beresi@rgu.ac.uk

Dawei Song
School of Computing
The Robert Gordon University
United Kingdom
d.song@rgu.ac.uk

Yuexian Hou
School of Computer Sci & Tec
Tianjin University
China
yxhou@tju.edu.cn

ABSTRACT

Conventional information retrieval (IR) evaluation relies on static relevance judgements in test collections. These, however, are insufficient for the evaluation of interactive IR (IIR) systems. When users browse search results, their decisions on whether to keep a document may be influenced by several factors including previously seen documents. This makes user-centred relevance judgements not only dynamic but also dependent on previous judgements. In this paper, we propose to use a probabilistic automaton (PA) to model the dynamics of users' relevance judgements. Based on the initial judgement data that can be collected in a proposed user study, the estimated PA can further simulate more dynamic relevance judgements, which are of potential usefulness for the evaluation of IIR systems.

Category and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Formal Model

Keywords: Interactive IR, Dynamic Relevance Judgement, Probabilistic Automaton, Simulation

1. INTRODUCTION

Relevance judgements in TREC test collections are static. However, in an interactive IR (IIR) environment, when users inspect the results of a search, relevance judgements become dynamic and dependent on each other. Consider the following two scenarios, the first one involving two complementary documents. Each document provides a portion of the solution but in combination they both provide a full solution to a user's problem. Either document in isolation is likely to be judged partially relevant (if not irrelevant), however, when combined, both are likely to be judged relevant. The second scenario is related to the comparison effects between two relevant documents, one being slightly more relevant than the other. Should the less relevant one be encountered first, the user is likely to want to keep it. When faced with the more relevant one, however, this decision may change.

Despite the recognition for the dynamic nature of relevance judgements [2], to the best of our knowledge, little attention has been paid to their formal modelling in terms of the judgement interference, i.e., the interference among relevance judgements for different documents. In this paper, we use the probabilistic automaton (PA) to model the changing process of judgement scores, as the result of the judgement interference. Specifically, the states of the PA can represent users' judgement states, which correspond to judgement scores. Thus the dynamic changes among judgement scores are modelled by the transitions among the PA states. Additionally, each symbol α in the alphabet Σ of the PA denotes (or defines) one type of interference, and the corresponding interference effects are reflected by the PA states transitions, for which the transition probabilities are represented in the transition matrix $M(\alpha)$. Suppose that the judgement for a document d_i is interfered by the judgement for another document d_j. Our method is that after the judgement for d_j, a symbol α will be generated and input to the PA, which triggers d_i's judgement changed, and the change obeys the transition probabilities in the $M(\alpha)$.

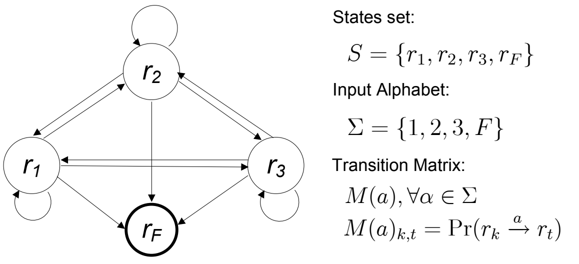
r_2
r_1
r_3
r_F
States set:
$S = \{r_1, r_2, r_3, r_F\}$
Input Alphabet:
$\Sigma = \{1, 2, 3, F\}$
Transition Matrix:
$M(a), \forall \alpha \in \Sigma$
$M(a)_{k,t} = \Pr(r_k \xrightarrow{a} r_t)$

Figure 1: A probabilistic automaton (PA) with three judgement states r_1, r_2 and r_3, and one final state r_F.

In this paper, we focus on the inter-document judgement interference. Nonetheless, the proposed method can be generalised to incorporate other factors that may affect relevance judgements, such as evolving information needs or contexts, by encoding these factors into the PA alphabet.

2. DYNAMIC RELEVANCE JUDGEMENT

2.1 Modelling Judgement Interference

The probabilistic automaton (PA) is a generalisation of Markov Chains [4]. In general, the PA can have any finite number of states, which means that it can model the changing process of any finite number of judgement scores. For simplicity, suppose there are three judgement scores, i.e. 1, 2 and 3, where higher score means higher relevance.

As shown in Figure 1, the PA is a 5-tuple (S, Σ, M, $\mathbf{w}$, r_F). The states (also called judgement states) r_1, r_2 and r_3 in S correspond to the graded scores 1, 2 and 3, respectively. The states distribution $\mathbf{w}$ is a row vector that represents the probabilities of all judgement states, where the k^{th} element of $\mathbf{w}$ denotes the probability of the state r_k . We neglect the final state r_F in $\mathbf{w}$ to make the description simpler. For each symbol $\alpha \in \Sigma$, the $M(\alpha)$, which is a stochastic matrix (i.e.,

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.
.

its each row sums to 1), represents transition probabilities among judgement states.

Let states distribution $\mathbf{w}_i$ be the initial states distribution of a document d_i to represent its static judgement (e.g., TREC judgement). The judgement of d_i, as interfered by the judgement of another document d_j, can be computed as:

$$\mathbf{w}_i^{\alpha} = \mathbf{w}_i \times M(\alpha). \tag{1}$$

where $\mathbf{w}_i^{\alpha}$ denotes the interfered judgements of d_i, and the α is the symbol generated after the judgement of d_j.

For example, if d_i's static judgement score is 2, then the $\mathbf{w}_i = [0, 1, 0]$, meaning that 100% probability of the state r_2. After judging d_j with the score 1, suppose the input symbol corresponds to d_j's judgement score, i.e., $\alpha = 1$ is generated, and accordingly the $M(\alpha) = \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0 & 0.8 & 0.2 \\ 0 & 0 & 1 \end{bmatrix}$, where the $M(\alpha)_{k,t}=\Pr(r_k \xrightarrow{\alpha} r_t)$ denotes the transition probability from the state r_k to r_t. Based on Formula 1, the interfered judgement result $\mathbf{w}_i^{\alpha}$ is [0, 0.8, 0.2]. This means that after judging a less relevant d_j with score 1, the user may heighten the judgement score of d_i, i.e., the user has 0.8 probability of keeping his original judgement (score 2) for d_i, and 0.2 probability of changing d_i's score from 2 to 3.

This example may be over simplistic to assume that the symbols correspond exactly to judgment scores. One should consider the inter-document dependency, or other factors, e.g. the task and context, in the alphabet encoding.

2.2 Simulating Dynamic judgements

Based on a TREC collection $\mathcal{C}$, our aim is to simulate the dynamic judgements of every $d_i \in \mathcal{C}$, assuming the users are judging a list of documents $d_1 \cdots d_n$. It could be extremely complex to consider all the possible judgement interferences. Therefore, we design and run a user study on the inter-document judgement interference of the representative document pairs $(d_i,\ d_j)$, in order to obtain the judgement transition matrices $(\forall \alpha) M(\alpha)$. The user study will be discussed in the next subsection. Here we assume that transition matrices are available and present how to use them to simulate the dynamic judgements of d_i.

Suppose the judgement of each d_j in the list $d_1 \cdots d_n$ generates an input symbol $\alpha(d_j)$ for the PA. After the user judged the documents $d_1 \cdots d_p$, the dynamic judgements of d_i at the position p $(p < n)$, can be computed by:

$$\mathbf{w}_i^{x_p} = \mathbf{w}_i \times M(x_p). \tag{2}$$

where the string $x_p = \alpha(d_1) \cdots \alpha(d_p)$, and the $M(x_p) = \prod_{j=1}^{p} M(\alpha(d_j))$ is also a transition matrix. Note that the dynamic judgement modeling is also connected to the design of novel evaluation metrics considering user behavior [5, 3].

2.3 User Study Methodology

This proposed user study is to collect the initial judgement interference data for the simulation task described in Section 2.2. We need to study the inter-document judgement interference of each document pair $(d_i,\ d_j)$, selected from TREC collections. The selection process should consider two factors between d_i and d_j, i.e., the document dependency (e.g., similarity) and judgement score difference, which are possibly related to the judgement interference. We plan to adopt the statistical document dependency, rather than the document dependency measured by ourselves or users, since we aim at using the collected interference data on selected documents to simulate dynamic judgements for a large number of other documents, for which it is too expensive to obtain the human-measured document dependency. We stick to use binary judgement in the initial stage, since it can reduce the efforts and randomness of user evaluation. The document pairs will be selected and assigned to several categories, and under the same category, all document pairs have the same document dependency degree and the same judgement score difference. We let each category correspond to each symbol α in the PA. We then study the pairwise-document judgement interference in each category, to learn the transition matrix $M(\alpha)$ for the corresponding symbol α.

For each document pair $(d_i,\ d_j)$, we need to consider two situations, i.e., $rank(d_i) > rank(d_j)$ and $rank(d_i) < rank(d_j)$. In the first situation, users would be presented with d_i first and be asked to judge its relevance with respect to an information need (represented for instance as a request or a simulated work task [1]). Once d_i was judged, users would be asked to judge d_j's relevance. To close the circle, users would be asked to judge d_i again, i.e. to reconsider their previous judgement. In the second situation, two separate user groups would be involved. Users in the first group would be asked to judge d_j first and then judge the d_i, while in the second group, users would be only asked to judge the relevance of d_i. Collected data in both situations could help us to study how the judgement of d_i can be interfered by the judgement of d_j. We will further investigate the user study design under two situations in the future.

3. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed to use a probabilistic automaton (PA) to model the dynamics of user-centred relevance judgements. We then further presented how to use PA, which can be trained through a user study, to simulate more dynamic judgement data, potentially useful for the IIR evaluation. In the future, we will adopt more appropriate strategies for encoding the alphabet Σ to include other factors that affect relevance judgement. We also plan to refine the user study methodology and carry out an extensive user study, of the best possible quality, as we can. Our ultimate goal is to simulate an IIR evaluation including the derivation of the appropriate evaluation metrics.

4. ACKNOWLEDGMENTS

We would like to thank anonymous reviewers for their constructive comments. This work is supported in part by the UK's EPSRC grant (No.: EP/F014708/2).

5. REFERENCES

[1] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3):8–3, 2003.

[2] S. Mizzaro. Relevance: The whole (hi)story. *Journal of the American Society for Information Science*, 48:810–832, 1996.

[3] C. Olivier, M. Donald, Z. Ya, and G. Pierre. Expected reciprocal rank for graded relevance. In *CIKM '09*, pages 621–630, 2009.

[4] W.-G. Tzeng. Learning probabilistic automata and markov chains via queries. *Machine Learning*, 8:151–166, 1992.

[5] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Incorporating user behavior information in ir evaluation. In *SIGIR 2009 Workshop on Understanding the user- Logging and interpreting user interactions in information retrieval*, page 3, 2009.