

# A probabilistic user model for information retrieval

A1, A2

## Abstract

A general level probabilistic user model for information retrieval is derived from the basic problem definition. The model is shown to work on various example scenarios and extensions such as new modes of feedback are discussed.

## 1 Introduction

## 2 Definition

### 2.1 Problem

An information retrieval system needs to solve the problem of incomplete knowledge of user's information need. Some key components must be defined in order to realize any such system.

First component is a known collection  $C$  of retrieval items. Each item in the collection  $d_i \in C$  must have a quantification: Call such quantification the feature representation  $x_i \in \mathcal{X}$  with some [metric] space  $\mathcal{X}$ .

For example, a piece of music can be described by the artist, title and record labels, each represented as value 1 of a corresponding binary variable in the joint space of all possible artist, title and record names.

Second component is the representation of information need. This can be formalized by equating the information need  $\theta$  with a distribution in the feature space. For example with some parametric distribution family we would set  $P(event A) = \int_A f_\theta(x)dx$  where the parameters define the

Last component is the representation of relevance, expressed as a variable  $r \in \mathbb{R}$ , often defined only on the unit interval. A typical choice interprets the value 0 as 'not relevant' and 1 as 'relevant'. Binary choice is often made, with the downside of losing e.g. interpretation of a value 0.5 as 'irrelevant', and so forth.

### 2.2 Solution

The three components are now connected to make the dependency graph of a solution. Note that the minimal solution graph has at least two edges, as any less would render a component independent of the other two and our problem definition redundant.

The complete graph is not useful: consider the two possible cases of effect. If the user's information need would affect the feature representation, the change of a user would violate the first assumption of a known collection. Vice versa if the feature representation should only affect the system's internal operation and not affect what the user is looking for. Therefore the solution has two edges, connecting the information need with the items through relevance.

## 3 Properties

In this section we have interpretation of higher level quantities. Session, information drift, precision-recall, feedback.

Session:

Precision-recall:

Feedback: implicit, explicit, pseudo

Figure 1: Graphical representation of the minimal component solution.

## **4 Examples**

## **5 Discussion**