A probabilistic user model for information retrieval

A1, A2

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.

Introduction

Definition

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 θ with a distribution in the feature space. For example with some parametric distribution family p_θ we would set $P(event\ A) = \int_A p_\theta(x)dx$.

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.

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 (Figure 1) is not useful: consider the two possible cases of interaction. 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 the feature representation should not affect what the user is looking for, but affect only the system's internal operation. Therefore the solution has two edges, connecting the information need with the items through relevance (Figure 2).

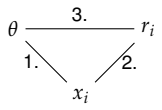

Figure 1: Complete graph.

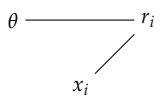

Figure 2: Solution graph.

Having fixed the dependency structure of the three components, the next step is to argue effect direction. Consider the interaction between θ and r_i (Figure 3). Natural direction of effect is that infromation need defines what is the relevance of an item. The other direction would imply that the items have some user independent state of relevance, and hence the system would not be providing the user relevant documents, but penalize users with non-relevant needs (T: not good... need more thought).

Similarly in Figure 4, if the fixed collection of items is modelled by their relevances, it implies that each document would have a different representation when the document's relevance is changed.

The final solution proposed models the relevance variable as a function of information need and item features (Figure 5). In probabilistic form the model is written as $p(r_i|\theta, x_i)$.

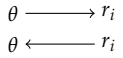

Figure 3: Possible effect directions between information need and relevance.

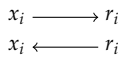

Figure 4: Possible effect directions between documents and relevance.

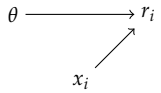

Figure 5: Final solution.

Properties

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

Definitions of consepts in IR systems

A session is a bounded continuous period of system use during which the information need is not changed. Assumption of a constant information need θ over the session implies constant item relevances and the IR problem is solved by statistical estimation using all data gathered during the session. The problem of session boundary detection is solved by change-point detection in the value of θ.

Alternative to a fixed θ is a drifting or shifting information need, reflecting the user's learning or remembering over time. To define a session in such a case requires an extra variable Θ for the task involving the retrieval. Then $(\theta_t)_{t=1}^{T}$ is an (autocorrelated) realisation of Θ over time, i.e. a process approach is needed.

The information need is not possible to be observed in full as this would negate the need for retrieval ('retrieve exactly this' is redundant). Querying, item or feature rating and other data collection mechanisms built to the user interface are inprecise and constrained ways for the system to get hints of the information need θ. Each mechanism is to be modelled individually as a variable dependent on the information need. For example, query strings and relevance feedback clicks on ranked items are both stochastic expressions of the same underlying θ.

History data, borrowing from other users, pseudo feedback.

Evaluation of the retrieval: precision-recall, diversity, etc.

The conditional nature of the solution: User's information need given the feature space and collection. Does it match the problem.

Latent variable model and Bayesian inference

Examples

Discussion