XML-IR: Coverage as Part of Relevance

Johan List and Arjen P. de Vries

Center for Mathematics and Computer Science (CWI)
Data Mining and Knowledge Discovery (INS1)
P.O.Box 94079, 1090GB Amsterdam, The Netherlands
{j.a.list, a.p.de.vries}@cwi.nl

Abstract

Relevance is a multidimensional concept, not only consisting of linguistic-only properties but also enriched by various other relevance dimensions that are largely orthogonal to the *topicality* (i.e. content-based relevance) of a document. The question is how to capture such dimensions of relevance effectively in a retrieval model.

In this paper we propose a model where we regard additional relevance dimensions independent (given a document instantiation). The independence assumption is made because it is very difficult to predict influence of relevance dimensions a-priori. The model also reflects our belief that modeling of additional knowledge with prior probabilities (in a probabilistic setting) is a counter-intuitive approach because of 1) the orthogonality of additional relevance dimensions and 2) the difficulty to reliably (re-)estimate dimension models, due to possible 'noise' introduced by non-dimension related priors.

Also, relevance feedback needs to be able to handle multiple dimensions of relevance effectively. Feedback in the model is done with dimension-specific feedback sets.

We can only report informally on the results of our model; based on the experimental scenarios performed, the model is appearing to perform very well, although quantitative assessments using an assessed collection are necessary to confirm this and draw further conclusions.

1 Introduction

We believe a clear distinction should be made between topicality and 'relevance', where topicality (i.e. content-based or linguistic similarity) is an approximation of relevance and can be seen as only a single dimension of relevance. An information need can include a variety of extra dimensions, not necessarily all linguistic in nature. Mizarro (Mizarro, 1998) gives examples of such dimensions, including comprehensibility (style or difficulty of the text) and quantity (how much information does the user want; this is measured in a.o. the size of documents and the number of documents returned to the user). Our aim is to capture such dimensions of relevance effectively in our retrieval model.

A closely related issue is the notion of 'coverage', as e.g. used in the INEX XML Retrieval initiative. Coverage is defined as how much of the document component is relevant to the topic of request. Estimating the right amount of coverage for a search request plays a significant role in the case of structured document retrieval where the desirable retrieval unit is not known a-priori. Effective determination of the retrieval unit is a key issue which distinguishes structured document retrieval from traditional retrieval (where the retrieval unit is fixed a-priori).

To further illustrate the retrieval unit problem, consider a short motivating example. Let us assume we have a document consisting of a section with three subsections, and each subsection containing five paragraphs. Now, the system that estimates topicality identifies three relevant paragraphs in a subsection. The open question is then whether to return the three separate paragraphs, or the single subsection containing these as well as the remaining two (possibly irrelevant) paragraphs. The additional context provided by the full subsection may be more desirable for a user than the individual three paragraphs in isolation.

Assume a user is trying to solve the retrieval unit question and decides to use coverage as an additional relevance dimension. For modeling coverage, the user decides to regard coverage as a function of both topicality of document components and the size of document components (size being an aspect of the quantity dimension). The user reasons that:

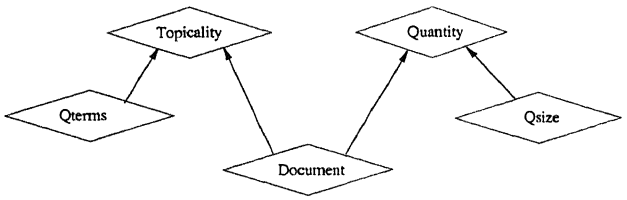

Figure 1: Encoding of additional relevance dimensions. Note that *Qterms* and *Qsize* denote information given by the query (query terms and preferred component size).

- the shorter the document component is, the more likely it will not contain enough information to fulfill the information need;
- the longer the document component is, the more likely it is that distilling the topically relevant information will take substantial more reader effort.

Now, when a user is ranking a document collection with regard to coverage, a ranking is performed against a combination of both topicality relevance and quantity relevance (where the user uses document component size as a representation of quantity). In probabilistic terms we are calculating the probability of complete relevance of a document component, given topicality relevance and quantity relevance.

More generally, we propose a probabilistic model where, given a document instantiation, we regard the dimensions of relevance independent based on the assumption that without user interaction, we cannot say anything about the influence of each dimension on user satisfaction. Traditional information retrieval uses only a single dimension of this model, namely topicality. The model is visualized in Figure 1.

2 Retrieval Model

2.1 Modeling Additional Relevance Dimensions

Firstly, for modeling additional relevance dimensions, we need a probabilistic description. The model in Figure 1 leads to the following. When $P(R_t|D_d)$ is the probability of topical relevance given document d and $P(R_q|D_d)$ is the probability of quantity relevance given document d, then we can calculate a joint probability of 'complete' relevance or user satisfaction as:

$$P(D_d,R_t,R_q,Q_{terms},Q_{size}) = P(R_t|D_d,Q_{terms})P(R_q|D_d,Q_{size})P(D_d)$$

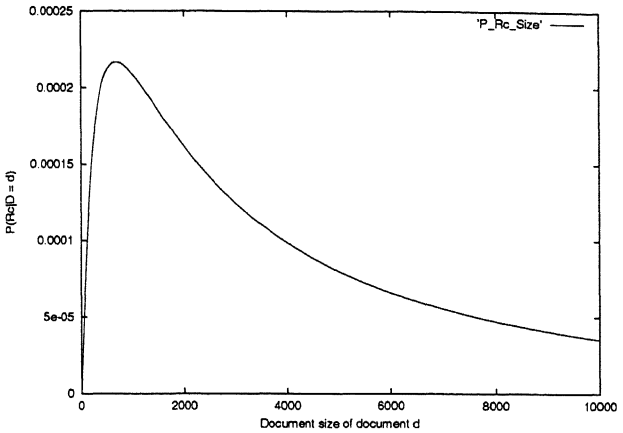

Figure 2: The log-normal distribution used for modeling the quantity dimension

Looking at the motivating example in Section 1 and especially the user reasoning for modeling the quantity dimension, we decided to use a log-normal distribution. It is a distribution characterized by both a steep slope at the start and a long tail (as can be seen from Figure 2). The steep slope at the start reflects the 'punishing' behavior we want to model for (extremely) short document components. The long tail reflects that we do want to punish long document components, but not as harshly as extremely short ones (since these still might be useful, even while taking more reader effort to distill the relevant information).

Secondly, we need the modeling parameter for the distribution itself. We have chosen for component size, but other possibilities include:

- the depth of the document component in the tree structure, where we want to penalize components present deep in the trees (generally small components and too specific) or components present high in the trees (generally large components and too broad);
- the number of children of a document component. A short document component containing a large amount of children highly likely contains a diversified mix of information and a could be less desirable for a user than a more homogeneous component.

Lastly, we need to integrate additional dimensions of relevance into our retrieval model. A general approach for modeling additional knowledge is using the prior probabilities (in a probabilistic setting). For ex-

ıle, Westerveld et al. (Westerveld et al., 2001) used strategy successfully to increase the likelihood of ing entry pages in a Web retrieval task. Also, a ır on document length improved retrieval perfor- ıce at TREC-style experiments (Hiemstra, 2000), ɔd on the assumption that longer documents have a ıer probability of containing relevant information. ince we regard additional dimensions of rele- ɔe independent given an document instantiation, model in Figure 1 also reflects our belief that ıeling other dimensions of relevance through ır probabilities is a somewhat counter-intuitive roach. For the rest of the discussion, note that have used a language modeling approach for ıeling topicality (see subsection 2.2).

'irstly, non-linguistic dimensions of relevance are gely) orthogonal to the topicality, estimated by language model. The orthogonality assumption ıeled by research performed in the user modeling relevance areas (see a.o. (Belkin et al., 1982a), lkin et al., 1982b), (Bruce, 1994), (Barry, 1994). ıin, the common thread in this work is the fact relevance is a multidimensional concept, of which cality is only a single one. Mizarro (Mizarro, 8) names other, possible non-topical dimensions *:tract characteristics of documents*, constructed in- endently from the particulars of the database or ection at hand. In other words: other, non-topical ensions are constructed independently from the guage models present in the documents of a col- ion.

econdly, encoding additional knowledge in prior ɔabilities makes it more difficult to reliably esti- ɔ dimension models, due to the possible noise non- ension related prior probabilities introduce.

Topicality Modeling

model used for describing topicality of docu- ıts is a probabilistic model, the statistical language ıel described by Hiemstra (Hiemstra, 2000). The n idea of this model is to extract and to compare ument and query models and determine the proba- :y that the document generated the query. In other ds, the statistical language model extracts linguis- nformation and is suited for modeling of the topi- ty dimension of the information need.

ı deriving document models for all of the doc- ɜnts in the collection, we regarded every subtree ɜent in the collection as a separate document. The ɔability of topical relevance $P(R_t|D_d, Q_{terms})$ where Q_{terms} consists of the set of query terms $\{T_1, \cdots, T_n\}$ is calculated with:

$$P(R_t|D_d, Q_{terms}) = P(R_t|D_d, T_1, \cdots, T_n) = P(D_d) \prod_{i=1}^{n} P(I_i) P(T_i|I_i, D_d)$$

where $P(I_i)$ is the probability that a term is important (the event I has a sample space of $\{0, 1\}$).

We follow the reasoning of Hiemstra (Hiemstra, 2000) to relate the model to a weighting scheme (tf.idf-based). After some manipulation of the model we get:

$$P(D_d, T_1, \cdots, T_n) \propto P(D_d) \prod_{i=1}^{n} (1 + \frac{\lambda P(T_i|D_d)}{(1-\lambda) P(T_i)})$$

As estimators for $P(D_d$, $P(T_i|D_d)$ and $P(T_i)$ we used:

$$P(D_d) = \frac{1}{n} \tag{2.1}$$

$$P(T_i|D_d) = \frac{tf_{i,d}}{\sum_i tf_{i,d}} \tag{2.2}$$

where n is the number of documents, $tf_{i,d}$ is the term frequency of term i in document d and $\sum_i tf(i, d)$ is the length of document d.

For $P(T_i)$ we used:

$$P(T_i) = \frac{df_i}{\sum_i df_i} \tag{2.3}$$

where df_i is the document frequency of term i.

Filling in the likelihood estimators gives us the following model for topicality (with a constant λ for all terms):

$$P(R_t|D_d, Q_{terms}) = P(R_t|D_d, T_1, \cdots, T_n) \propto \sum_{i=1}^{n} \log(1 + \frac{\lambda}{1-\lambda} \frac{tf_{i,d}}{\sum_i tf_{i,d}} \frac{\sum df_i}{df_i})$$

We used a very simple query model resulting in query term weights represented with $tf_{i,q}$, the term frequency of term i in query q.

3 Relevance Dimensions and Relevance Feedback

Explicit relevance feedback is the main entry point for learning additional relevance dimensions, since we cannot assume a system has knowledge of other relevance dimensions at the initial query stage.

We make a distinction between relevance feedback as a *directed* process, in the case of a user identifying relevant documents and feeding that information back to the system or relevance feedback as an *undirected* process, in the case of taking top-ranked documents and using that collection for query term expansion.

For the purpose of relevance feedback, let us assume we have a user examining the result set after an initial search and this user is judging the results set on topicality and quantity. We can distinguish three possible decisions by this user when judging a result:

- The user sees the result as correct regarding topicality and not quantity;
- The user sees the result as correct regarding quantity and not topicality;
- The user sees the result as contributing to both relevance dimensions.

If the first situation applies and a user is giving feedback on topicality alone and not on quantity, we simply can re-estimate the language model parameters and disregard quantity influence altogether. If the second situation applies and a user is giving feedback on quantity alone and not topicality, we simply can re-estimate the model parameters of the quantity model and we can leave the language model parameters as they were. The situation becomes more difficult in the third case, when a user is giving feedback based both on topicality and quantity. Feedback in this situation can be visualized with the adapted model of Figure 3. In relevance feedback, the user can be regarded as specifying the probability distributions of topicality and quantity, given that the document *is* 'completely' relevant.

We have only experimented with undirected feedback without further specification of probability distributions for relevance feedback. Then undirected feedback will only increase performance when giving feedback per dimension. To explain this further, consider the user from the motivating example having performed a search. The ranking has been performed on quantity, the combination of topicality and component size. The question if the quantity-ranked set

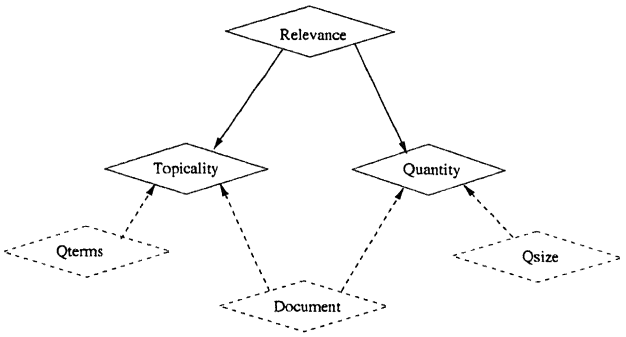

Figure 3: Relevance feedback

of document components can be used effectively for blind feedback.

Compare the document set ranked on quantity with a document set ranked on topicality only. Since it is possible that documents with a lower topicality-only score get a higher rank in the quantity ranking (because of a better size), using the quantity-ranked document set for e.g. topicality feedback will worsen the quality of the (estimated) topicality model parameters.

To update our model given relevance feedback, we perform re-estimation of the separate dimension models. We see the document feedback set as a collection of *content sources*. Each content source is characterized by a collection of properties which map to relevance dimensions. For example, when considering topicality and quantity, we characterize each content source with two properties R_t (topicality) and R_q (quantity). Recall that we consider additional relevance dimensions independent given a document instantiation. We can characterize a document D_r in the feedback set, being characterized by topicality and quantity as:

$$P(D_d, R_t, R_q, Q_{terms}, Q_{size}) = P(R_t|D_d, Q_{terms})P(R_q|D_d, Q_{size})P(D_d)$$

We assigned a uniform distribution to $P(D_d)$ so we can safely leave this out of the model without affecting the ranking. Using the language model for topicality (including a λ_i, varying per term) and the log-normal for quantity gives us for $P(D_d, R_t, R_q, Q_{terms}, Q_{size})$:

$$\left[P(D_d)\prod_{i=1}^{n}\lambda_i P(T_i|D_d) + (1-\lambda_i)P(T_i)\right] P(R_q|D_d, Q_{size})$$

We now want to find the set of model parameters which maximize the likelihood (with r feedback documents):

$$\prod_{e=1}^{r}\Big[P(D_e)\prod_{i=1}^{n}\lambda_i P(T_i|D_e)+(1-\lambda_i)P(T_i)\Big] \\ P(R_q|D_e,Q_{size})$$

When we work the model out further for the topicality and quantity dimensions only (where quantity is modeled by a log-normal distribution) and leave out $P(D)$ since it is uniform, we want to maximize the likelihood L (with r feedback documents):

$$\prod_{e=1}^{r}\Big[\prod_{i=1}^{n}\lambda_i P(T_i|D_e)+(1-\lambda_i)P(T_i)\Big]P(R_q|D_e,Q_{size})$$

or the log-likelihood Λ:

$$\sum_{e=1}^{r}\sum_{i=1}^{n}\log(\lambda_i P(T_i|D_e)+(1-\lambda_i)P(T_i))+ \\ \sum_{e=1}^{r}\log P(R_q|D_e,Q_{size})$$

Due to the independence assumption, we can divide the estimation problem into two subproblems and update each dimension separately (with r feedback documents):

$$\lambda_i^* = \arg\max_{\lambda_i}\sum_{e=1}^{r}\log(\lambda_i P(T_i|D_e)+(1-\lambda_i)P(T_i)) \tag{3.1}$$

$$\{\mu^*,\sigma^*\} = \arg\max_{\{\mu,\sigma\}}\sum_{e=1}^{r}\log P(R_q|D_e,Q_{size}) \tag{3.2}$$

For the first estimation problem in equation 3.1 we can use EM (Hiemstra, 2000). For iteration p we use as E-step (with r feedback documents):

$$k_i = \sum_{e=1}^{r}\frac{\lambda_i^{(p)}P(T_i|D_e)}{(1-\lambda_i^{(p)})P(T_i)+\lambda_i^{(p)}P(T_i|D_e)}$$

and as M-step:

$$\lambda_i^{(p+1)} = \frac{k_i}{r}$$

For the second estimation problem in equation 3.2 we can perform a maximum likelihood estimation procedure for μ and σ as follows. The usual approach to estimation of a log-normal distribution $LN(\mu,\sigma^2)$ is to consider a new data sample Y_i, where $Y_i = \log X_i$, $i = 1\cdots n$. The estimation then becomes the estimation of a normal distribution $N(\mu,\sigma)$ for which we can easily derive the maximum likelihood estimators.

The probability density function for a normal distribution Y with mean μ and standard deviation σ is described by

$$f(Y) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{Y_i-\mu}{\sigma})^2)$$

The likelihood function is given by:

$$\begin{aligned} L &= \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{Y_i-\mu}{\sigma})^2) \\ &= (\frac{1}{\sigma\sqrt{2\pi}})^n\exp(-\frac{1}{2}\sum_{i=1}^{n}(\frac{Y_i-\mu}{\sigma})^2) \end{aligned}$$

The log-likelihood is given by:

$$\begin{aligned} \Lambda &= \log(\frac{1}{\sigma\sqrt{2\pi}})^n\exp(-\frac{1}{2}\sum_{i=1}^{n}(\frac{Y_i-\mu}{\sigma})^2) \\ &= \log(\frac{1}{\sigma\sqrt{2\pi}})^n+\log(\exp(-\frac{1}{2}\sum_{i=1}^{n}(\frac{Y_i-\mu}{\sigma})^2)) \end{aligned}$$

Working this out further gives us:

$$\Lambda = -n\log\sigma-\frac{n}{2}\log 2\pi-\frac{1}{2}\sum_{i=1}^{n}(\frac{Y_i-\mu}{\sigma})^2$$

Taking the partial derivatives with respect to μ and σ gives us:

$$\frac{\partial(\Lambda)}{\partial(\mu)} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2$$

$$\frac{\partial(\Lambda)}{\partial(\sigma)} = -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(Y_i-\mu)^2$$

We can set the partial derivatives to 0 and solve for μ and σ (we know that the original normal function is

Table 1: Experimentation scenarios.

Scenario	Retr. Unit	Dimension(s)
V_1	$\{tr(article)\}$	R_t
V_2	$\{tr(*)\}$	R_t
V_3	$\{tr(*)\}$	R_t, R_q

positive for all values in the range, and we know there is a single (non-local) maximum). This gives us:

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\sigma^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2}$$

4 Experimental Work

We participated in INEX[1] and implemented an XML retrieval system based on Monet, a main-memory database kernel.

With our system, we performed three experimentation scenarios. The first scenario mimicked 'flat-document' retrieval of articles, i.e. retrieval of documents which possess no structure whatsoever. The second scenario regarded all subtrees or transitive closures in the collection as separate documents.

For the third scenario we re-used the result sets of the second run and used the log-normal distribution to model the quantity dimension. To penalize the retrieval of extremely long document components (this in contrast with the language model that assigns a higher probability to longer documents), as well as extremely short document components, we set the mean at 500 (representing a user with a preference for components of 500 words).

In all three scenarios we used the statistical language model of subsection 2.2 to model topicality.

Table 1 summarizes our experimentation scenarios. Note that $tr(c)$ denotes the transitive closure of a document component with root c and $tr(*)$ denotes the transitive closures of all subtrees present in the original XML syntax trees.

[1]XML Retrieval Initiative, see `http://qmir.dcs.qmw.ac.uk/inex/index.html`

5 Conclusions and Future Work

From an informal look into our results, modeling coverage by using a combination of topicality and quantity (in terms of component size), using a subjective probability function for the latter, seems to work pretty well. To be able to make this conclusion more firmly, we need to perform further experiments on coverage estimation, as well as other dimensions of relevance.

For quantitatively backing up our model, we need evaluation results of the runs as well (sadly not available at the time of finishing this paper). We plan to report on the retrieval performance in our INEX workshop paper (List and de Vries, 2003).

In future work, we intend to perform experimentation with relevance feedback and extend the model further for other dimensions and ultimately, for the mapping of user context to retrieval model parameters.

References

C.L. Barry. 1994. User-defined Relevance Criteria: An Exploratory Study. *Journal of the American Society for Information Science*, 45(3):149–159.

N.J. Belkin, R.N. Oddy, and H.M. Brooks. 1982a. ASK for Information Retrieval: Part 1. Background and Theory. *Journal of Documentation*, 38(2):61–71.

N.J. Belkin, R.N. Oddy, and H.M. Brooks. 1982b. ASK for Information Retrieval: Part 2. Results of a Design Study. *Journal of Documentation*, 38(3):145–164.

H.W. Bruce. 1994. A Cognitive View of the Situational Dynamism of User-centered Relevance Estimation. *Journal of the American Society for Information Science*, 45(3):142–148.

D. Hiemstra. 2000. *Using Language Models for Information Retrieval.* Ph.D. thesis, University of Twente, Twente, The Netherlands.

J.A. List and A.P. de Vries. 2003. CWI at INEX 2002 (to appear, January 2003).

S. Mizarro. 1998. How Many Relevances in Information Retrieval? *Interacting With Computers*, 10(3):305–322.

T. Westerveld, W. Kraaij, and D. Hiemstra. 2001. Retrieving Web Pages using Content, Links, URLs and Anchors. In *NIST Special Publication 500-250, The 10th TREC Retrieval Conference (TREC 2001).*