Doxels in context for retrieval: from structure to neighbours *

Delphine Verbyst
LIG - Université Joseph Fourier
385 rue de la Bibliothèque
38041 Grenoble Cedex 9, France
Delphine.Verbyst@imag.fr

Philippe Mulhem
LIG - CNRS
385 rue de la Bibliothèque
38041 Grenoble Cedex 9, France
Philippe.Mulhem@imag.fr

ABSTRACT

We propose in this paper a new way of considering retrieval of structured documents, by exploiting non-structural relations between structured document elements (doxels). These relations may be defined by human beings (e.g. by the authors of the documents for navigation or reference purposes), but may also be created by the information retrieval system (e.g. using kNN). Unlike Pagerank or HITS that separate features link and content features, we integrate these two aspects in defining a relative specificity and a relative exhaustivity between doxels. We use these features, as well as the doxel content, in a comprehensive matching process. One concern here is to facilitate the exploration of the result space by selecting the relevant doxels, and by indicating potential good neighbours to access from one doxel. Results of experiments on the INEX2005 test collection are presented.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval - Retrieval Models

General Terms
Theory, Experimentations

Keywords
XML retrieval, relational indexing, exhaustivity, specificity

1. INTRODUCTION

A query result of a classical information retrieval system is a one dimensional list of unrelated documents in which a user navigates. When considering retrieval of structured documents, taking into account the structure, both in the matching process and the presentation of results, is a must [3].

In this paper, we focus on the use of inter-relations between parts of structured documents as a mean to navigate between documents. These different ways to navigate among query results complicate the exploration of the result space. We demonstrate here that integrating non structural relations between document elements (doxels) in the indexing and retrieval processes increases the quality of the retrieval of structured documents. In the following, such relations between doxels will be referred to as the context of the doxels.

Our assumption is that document parts are not only relevant because of their content, but also because they are related to other document parts that answer the query. In some way, we revisit the Cluster Hypothesis of van Rijsbergen [19], by considering that the relevance value of each document is impacted by the relevance values of related documents.

One of our concerns is to take into account the user needs in term of a query result exploration: results may be different if a user wants specific documents (searching in [20]), and if he wants to navigate in documents (scanning in [20]). When we consider navigation in a result space, we need to tackle the elements that are presented as entry points for the navigation. This point has been studied in works on best entry points [8]. Our concern is also to provide such good entry points, but we also propose to provide best navigational links to explore the result space. To do that, we first build content based inter-relations between doxels, then characterize these relations using relative exhaustivity and specificity at indexing time, and finally integrate these elements during the matching process.

We performed experiments on the INEX 2005 collection, showing the effectiveness of our matching in context process. In particular, we showed that the use of relative exhaustivity and relative specificity values between doxels can help achieve better performance. We also found that the relative exhaustivity seems a more important feature to integrate than the relative specificity for the INEX 2005 collection.

The rest of this paper is organized as follows: Section 2 briefly introduces related works. The doxel space is described in detail in Section 3, in which we propose a document model using the context. Section 4 introduces our matching in context process. We set up the experiments and give the corresponding results in Section 5.

2. RELATED WORK

Different research directions have been proposed to improve the use of non structural information such as contextual information.

*This work is supported by Orange-France Telecom.
On the one hand, let’s consider works on non structural links, such as Web links or similarity links. Algorithms such as Pagerank [1] or HITS [7] use the link structure of a network of web pages to assign weights to each page in the network. They allow respectively to calculate the popularity scores of web pages or to classify web sites in hubs and authorities. However, (a) they separate features coming from links and from content, and (b) the relations are not used to help navigation in the result set. Savoy, in [14], has demonstrated, on non structured documents, that integrating different relations during matching improves the quality of the results. However, he has done so without qualifying the strength of the inter-relations. In [18], Smucker and Allan show that similarity links may help navigation in the result space. In [17] the same authors propose to build query-biased similarity links between documents in order to increase the quality of results.

On the other hand, lots of works are based on structural links between parts of documents. Lalmas and Rölleke [12] combine content element and weighted structural links by introducing an accessibility dimension to retrieve the best document units. In [5], the content of a doxel depends of the content of all preceding doxels according to the reading order of doxels in document. This approach ranked 64th/106 at INEX 2006. Schenkel and Theobald in [15] compute the score of an element as the concatenation of the text contents of all the node descendants in document order. This method produced good results at the INEX 2006 Focussed Task. The structure is used in yet another way by Lu, Robertson and MacFarlane in [9]: they first use a document level retrieval and then set a cut-off for the retrieved results to avoid too short elements to be retrieved. This work achieved the highest performance at the INEX 2006 Focussed Task.

Existing works show the importance of integrating links, either existing a priori or created a posteriori, between documents or parts of documents to improve results in information retrieval. We propose here a document model that integrates structured documents and inter-related doxels to ease the exploration of the result set.

3. DOXEL SPACE

In this paper, we assume that composition links are meaningful to express the content of doxels. We also model non compositional relational aspects between doxels.

3.1 Doxel content

The representation of the content of doxel $d_i$ is a vector generated from a usual vector space model using the whole content of the doxel: $d_i = (w_{i,1}, ..., w_{i,k})$. Such a representation has proved to give good results for structured document retrieval [4]. The weighting scheme retained is a simple tf.idf, with idf based on the whole corpus and with the following normalizations: the tf is normalized by the max of the tf of each doxel, and the idf is log-based, according to the document collection frequency.

3.2 Doxel context

Let’s consider two content inter-related structured documents, with the first one, $D_1$, about “secure code distribution” and the second one, $D_2$, about “code signing”. They share information on “code” verification. If a user looks for all the information about secure code distribution, the system should indicate that the link above is a relevant part of the query result. If the user only wants to have general informations about secure code distribution, $D_1$ is highly relevant, $D_2$ is less relevant, and moreover, the system should indicate that the link between $D_1$ and $D_2$ is not interesting for this query result. To characterize the relations between doxels, we propose to define relative exhaustivity and relative specificity between doxels. These features are inspired from the definitions of specificity and exhaustivity proposed at INEX 2005 [11]. Consider a non-compositional relation from the doxels $d_1$ to the doxel $d_2$:

- The relative specificity of this relation, noted $Spe(d_1, d_2)$, denotes the extent to which $d_2$ focuses on the topics of $d_1$. For instance, if $d_2$ deals only with elements from $d_1$, then $Spe(d_1, d_2)$ should be close to 1.
- The relative exhaustivity of this relation, noted $Exh(d_1, d_2)$, denotes the extent to which $d_2$ deals with all the topics of $d_1$. For instance, if $d_2$ discusses all the elements of $d_1$, then $Exh(d_1, d_2)$ should be close to 1.

The values of these features are in $[0, 1]$. We could think that these features behave in an opposite way: When $Spe(d_1, d_2)$ is high, then $Exh(d_1, d_2)$ is low, and vice versa.

Relative specificity and relative exhaustivity between two doxels are extensions of the overlap function [13] of the index at INEX 2005 [11]. Consider a non-compositional relation from the doxels $d_1$ and $d_2$:

1. We estimate a priori values of the exhaustivity and the specificity of $d_1$ and $d_2$, based on a vector where weights are $tf$ or $t.f.idf$:

$$Exh_{ap}(d_1, d_2) = \frac{\sum_i w_{m,i} \cdot w_{2,i}}{\sum_i w_{1,i} \cdot w_{2,i}}$$

$$Spe_{ap}(d_1, d_2) = \frac{\sum_i w_{1,i} \cdot w_{2,i}}{\sum_i w_{1,i}^2}$$

where: $w_{m,i} = \begin{cases} w_{m,i} & \text{if } w_{n,i} \leq 1 \\ \sqrt{w_{m,i} \cdot w_{n,i}} & \text{otherwise.} \end{cases}$

$w_{m,n,i}$ ensures that the scores are in $[0, 1]$.

The above a priori values do not take into account query information. To do so, we introduce new values as follows:

2. We estimate $Exh_{sq}(d_1, d_2, q_{pre})$ and $Spe_{sq}(d_1, d_2, q_{pre})$ as relative exhaustivity and specificity values of inter related doxels $d_1$ and $d_2$, for each query $q_{pre}$ of $Q_{pre}$, and we compute a differential between the doxels, for all the queries, so as to obtain $Exh_{sq}(D_1, D_2, Q_{pre})$ et $Spe_{sq}(D_1, D_2, Q_{pre})$, the relative exhaustivity and relative specificity values on the sample set.
This approach is inspired by Callan and Connel [2]. They describe some resources by using pre-defined queries; one resource is an unspecified information retrieval system. Once the resources are described, a meta search engine can weight their results at query processing time. For our purpose, we reuse this idea by: (a) using \( q_{\text{pre}} \) queries each composed of one randomly selected term in the vocabulary of the collection, (b) computing the doxels exhaustivity and specificity for the query. This “query-based sampling” method allows us to compute only a few queries instead of all the possible queries, which would be too expensive or impossible. The number of queries used to do this sampling is a parameter of our approach, and the best way to define such query set is to use logs of the system.

We finally compute the relative exhaustivity and specificity as a combination of the \textit{a priori} values and values obtained on sample sets of queries, with the following formula:

\[
\begin{align*}
\text{Exh} (d_1, d_2) &= \lambda \cdot \text{Exh}_{\text{ap}}(d_1, d_2) \\
&\quad + (1 - \lambda) \cdot \text{Exh}_{\text{ap-set}}(d_1, d_2, S_{\text{Qpre}}) \\
\text{Spe} (d_1, d_2) &= \lambda \cdot \text{Spe}_{\text{ap}}(d_1, d_2) \\
&\quad + (1 - \lambda) \cdot \text{Spe}_{\text{ap-set}}(d_1, d_2, S_{\text{Qpre}})
\end{align*}
\]

where \( \lambda \in [0.5, 1] \).

In the above formulas, the larger the query set is, the closer \( \lambda \) is to 0.5, and the smaller the query set is, the closer \( \lambda \) is to 1.

We intend to use relative specificity and exhaustivity to present only relevant relations between relevant doxels of the result set for a given user and a given request.

### 3.3 Example

In the INEX2005 Ad Hoc Collection, let’s consider two doxels:

- the doxel \( \text{dox}_1 \) corresponding to the path `/article[1]/body[1]/sec[6]` of the document file `xml/co/1997/r6026.xml`, which is about secure code distribution and
- the doxel \( \text{dox}_2 \) corresponding to the path `/article[1]/body[1]/sec[6]/s1[3]/s3` of the document file `xml/co/2000/s5033.xml`, which is about code signing.

When no query sample set are considered, we compute \( \text{Exh}_{\text{ap}}(\text{dox}_1, \text{dox}_2) = 0.37 \) as the relative exhaustivity value and \( \text{Spe}_{\text{ap}}(\text{dox}_1, \text{dox}_2) = 0.12 \) as the relative \textit{a priori} specificity values for the two doxels \( \text{dox}_1 \) and \( \text{dox}_2 \). This is coherent with the fact that \( \text{dox}_2 \) deals with a lot of the content of \( \text{dox}_1 \), and that \( \text{dox}_2 \) talks also a lot of other subjects outside the scope of \( \text{dox}_1 \).

### 4. MATCHING IN CONTEXT MODEL

As we have characterized the doxel context, the matching process should return doxels relevant to the user’s information needs regarding both content and structure aspects, and considering the context of each relevant doxel.

We define the matching function as a linear combination of a standard matching result without context and a matching result based on relative specificity and exhaustivity. The relevant status value \( RSV(d, q) \) for a given doxel \( d \) and a given query \( q \) is thus given by:

\[
RSV(d, q) = \alpha \cdot RSV_{\text{content}}(d, q) + (1 - \alpha) \cdot RSV_{\text{context}}(d, q),
\]

where \( \alpha \in [0, 1] \) is experimentally fixed. \( RSV_{\text{context}}(d, q) \) is the score without considering the set of neighbours \( V_d \) of \( d \) (i.e. cosine similarity) and

\[
RSV_{\text{context}}(d, q) = \sum_{d' \in V_d} \beta \cdot \text{Exh}(d, d') + (1 - \beta) \cdot \text{Spe}(d') RSV_{\text{content}}(d', q) \text{,}
\]

where \( \beta \in [0, 1] \) is used to privilege exhaustivity or specificity.

The matching in context model computes scores with both content and context dimensions to complete our model. But at this point, the organization of the result set is not yet done: lists of results should include the neighbours which improve the \( RSV \) for a given query, to allow the user to browse this result set.

### 5. EXPERIMENTS AND RESULTS

We want to evaluate the usefulness of neighbours in structured document retrieval. To do so, we have considered the INEX2005 Ad Hoc Collection (16000 documents, 29 queries). We implemented a vector space model based on usual \( \text{tf.idf} \). The \( RSV_{\text{context}} \) function is the cosine and the neighbourhood of a doxel is defined by its 4 Nearest Neighbours [16]. The relations are built offline, the process being speed-up by the use of a clustering techniques. We generate the 4 nearest neighbours because it would have been too difficult to compute the neighbours for the 9 millions doxels, and because of the large vectors dimensions. We generate randomly 500 referent vectors, one vector per cluster. Each doxel belongs to the cluster corresponding to the nearest referent vector.

The evaluation is proposed with the INEX2005 Ad Hoc task (OVERLAP=on, QUANT=gen, TASK=Cofocused). It is the Content Only Focussed Task, assuming the user prefers a single element that is the most relevant. The generalised (gen) function allows different degrees of relevance, by considering not only fully specific and highly exhaustive doxels as relevant. The overlap on means that the evaluation considers only non overlapping doxels.

#### 5.1 \( \alpha \) and \( \beta \) parameters

We need first to fix \( \alpha \) and \( \beta \) to compare our model to those of INEX2005 participants. We present in table 1 the obtained results in terms of normalised cumulated gain at 10, 25 and 50 and mean average effort precision measure, for different values of \( \alpha \) and \( \beta \). The XCG measures are an extension of the Cumulative Gain (CG) based measures, and include the user-oriented measures of normalised extended cumulative gain and the system-oriented effort-precision/gain-recall measures. The MAep measure is the non-interpolated mean average effort/precision, calculated by averaging the effort-precision values obtained for each rank where a relevant document is returned [10].
Table 1 presents results for 5 “idf” runs and the result of the paired bilateral Student’s t-test, which is a common test in information retrieval [6]. The statistically significant results at the 0.05 level compared to the best results, highlighted in bold font, are underlined in the table. the \((1,x)\) run represents the baseline without neighbours; the \((0.5,0.25)\) run outperforms significantly the results of the \((0.5,0.25)\) run, for the nXCG at 10 doxels measure (+11.6%) but nXCG at 25 and 50 doxels are the same; the \((0.75,0.25)\) run outperforms significantly the results of the \((0.5,0.25)\) run and of the baseline, for the nXCG at 25 doxels measure (+16.3% and +26.5%) but nXCG at 10 and 50 doxels are the same; the \((0.75,0.75)\) run shows a significant increase of the results of the \((0.75,0.25)\) run and of the baseline for the nXCG at 50 doxels measure (+6.2% and +19.1%) but nXCG at 10 and 25 doxels are the same. \(\alpha\) and \(\beta\) parameters have to be chosen between the last three runs of the table. We decide in the following to consider the MAep measure: the best score is obtained for the \((0.75,0.75)\) run.

These results show that exhaustivity is a more important factor than specificity for users needs on the INEX2005 collection, for the queries of this pool. As a result, we decide to fix \(\alpha = 0.75\) and \(\beta = 0.75\).

### 5.3 Our results vs. official INEX 2005 runs

To compare our system to the systems of participants of INEX 2005 campaign, we consider a run with parameters \(\alpha = 1\) and \(\beta = x\). This way, we avoid all the context part of our system. It is reduced to a simple \(tf.idf\) based system. Non-interpolated mean average effort-precision MAep scores 0.0286 for INEX2005 Ad Hoc (OVERLAP=on, QUANT=gen, TASK=Co.Focussed): our system would have been ranked at the 36th position over 44 different systems.

This baseline corresponds to second part results of INEX 2005 campaign. Indeed at INEX campaign, people do not use \(tf.idf\) measures to compute the RSV score; system are build on BM-25 for instance which has been proved to give better results. Our aim in this article is not to outperform existing systems, but to show the interest of modeling context, and particularly creating relations between doxels. Our model shows that the use of inter-relations and relative exhaustivity and specificity values can improve the results: using such a model on a more advance RSV should allow one to perform better and to challenge other systems.

### 6. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a contribution to interrelated structured document retrieval based on a non compositional relational indexing. Relative specificity and exhaustivity values have been computed for neighbours in a collection to characterize the relations between doxels. These values are used by the query processing. The evaluation describes the results of our proposal on the INEX2005 Ad Hoc (Co.Focussed) task for cumulated gain at 25 and 50 doxels. The results obtained are significantly better when using the context of doxels.

At a theoretical level, we plan to provide a better way to deal with the query sampling method in a way to compute accurate relative exhaustivity and specificity values between doxels.

We also plan to thoroughly study our system to explore the potential benefit of different doxel neighbourhood configurations (number of neighbours, doxel size, doxel types) on other INEX collections, typically INEX2007. We will tune the parameters by trying new values for \(\alpha\) and \(\beta\), which describe respectively how important the content is compared to the context, and how to consider exhaustivity compared to specificity. More specifically, we will define how to improve the results even at the very first doxels by integrating relational aspects in a probabilistic model.

<table>
<thead>
<tr>
<th>Run((\alpha, \beta))</th>
<th>nXCG@10</th>
<th>nXCG@25</th>
<th>nXCG@50</th>
<th>MAep</th>
</tr>
</thead>
<tbody>
<tr>
<td>((1, x))</td>
<td>0.1021</td>
<td>0.079</td>
<td>0.0886</td>
<td>0.0286</td>
</tr>
<tr>
<td>((0.5, 0.25))</td>
<td>0.1049</td>
<td>0.0859</td>
<td>0.0993</td>
<td>0.0335</td>
</tr>
<tr>
<td>((0.5, 0.75))</td>
<td><strong>0.1171</strong></td>
<td>0.0908</td>
<td>0.1012</td>
<td>0.0357</td>
</tr>
<tr>
<td>((0.75, 0.25))</td>
<td>0.1092</td>
<td><strong>0.0999</strong></td>
<td>0.1035</td>
<td>0.0359</td>
</tr>
<tr>
<td>((0.75, 0.75))</td>
<td>0.1121</td>
<td>0.0999</td>
<td><strong>0.1055</strong></td>
<td><strong>0.0367</strong></td>
</tr>
</tbody>
</table>
Table 2: nXCG for INEX2005 Ad Hoc.

<table>
<thead>
<tr>
<th>Run((\alpha, \beta))</th>
<th>nXCG@10</th>
<th>nXCG@25</th>
<th>nXCG@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, x)</td>
<td>0.1021</td>
<td>0.079</td>
<td>0.0886</td>
</tr>
<tr>
<td>(0.75, 0.75)</td>
<td>0.1121 (+9.8%)</td>
<td>0.0999 (+25.3%)</td>
<td>0.1055 (+19.1%)</td>
</tr>
<tr>
<td>(0.75, 0.75)</td>
<td>0.0989 (-3.1%)</td>
<td>0.0779 (-1.4%)</td>
<td>0.0828 (-6.5%)</td>
</tr>
</tbody>
</table>

Table 3: nXCG for topic 207 of INEX2005 Ad Hoc.

<table>
<thead>
<tr>
<th>Run((\alpha, \beta))</th>
<th>nXCG@10</th>
<th>nXCG@25</th>
<th>nXCG@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, x)</td>
<td>0.54</td>
<td>0.44</td>
<td>0.37</td>
</tr>
<tr>
<td>(0.75, 0.75)</td>
<td>0.64 (+18.3%)</td>
<td>0.49 (+9.7%)</td>
<td>0.39 (+5.4%)</td>
</tr>
</tbody>
</table>

7. REFERENCES