LIG at INEX 2007 Ad Hoc Track: Using Collectionlinks as Context

Delphine Verbyst[1] and Philippe Mulhem[2]

[1] LIG - Université Joseph Fourier, Grenoble, France
Delphine.Verbyst@imag.fr
[2] LIG - CNRS, Grenoble, France
Philippe.Mulhem@imag.fr

Abstract. We present in this paper the work[1] of the Information Retrieval Modeling Group (MRIM) of the Computer Science Laboratory of Grenoble (LIG) at the INEX 2007 Ad Hoc Track. We study here the impact of non structural relations between structured document elements (doxels) on structured documents retrieval. We use existing links between doxels of the collection, encoded with the *collectionlink* tag, to integrate link and content aspects. We characterize the relation induced by the *collectionlink* tags with relative exhaustivity and specificity scores. As a consequence, the matching process is based on doxels content and these features. Results of experiments on the test collection are presented. Runs using non structural links overperform a baseline without such links.

1 Introduction

This paper describes the approach used for the Ad Hoc Track of the INEX 2007 competition. Our goal here is to show that the use of non structural links can increase the quality of the results provided by an information retrieval system on XML documents. We consider that handling links between documents in a smart way may help an information retrieval system, not only to provide better results, but also to organize the results in a way to overcome the usual simple list of documents. For INEX 2007, we only show that our approach impacts in a positive way the quality of the results provided.

The use of non structural links, such as Web links or similarity links has been studied in the past. Well known algorithms such as Pagerank [1] or HITS [3] do not integrate in a seamless way the links in the matching process. Savoy, in [6], showed that the use of non structural links may provide good results, without qualifying the strength of the inter-relations. In [7], Smucker and Allan show that similarity links may help navigation in the result space. We want, with the work described here, to go further in this direction.

In the following, the non structural relations between doxels will be referred to as the *context* of the doxels. Our assumption is that document parts are

[1] This work is supported by Orange France Telecom.

N. Fuhr et al. (Eds.): INEX 2007, LNCS 4862, pp. 138–147, 2008.

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 [8], by considering that the relevance of each document is impacted by the values of related documents.

In our proposal, we first build inter-relations between doxels, and then characterize these relations using relative exhaustivity and specificity (see section 3.2) at indexing time. These elements are used later on by the matching process.

The nine officially submitted runs by the LIG for the Ad Hoc track integrate such non structural links. For each of the three tasks (Focused, Relevant in Context, Best in Context) a baseline without using such links was submitted. Taking into account the non structural links outperforms consistently this baseline.

The remaining of this paper is organized as follows: we describe the links that were used in our experiments in part 2, 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. Results of the INEX 2007 Ad Hoc track are presented in Section 5.

2 Choice of Collectionlinks

The idea of considering neighbours was first proposed in [9], in order to facilitate the exploration of the result space by selecting the relevant doxels, and by indicating potential good neighbours to access from one doxel. For this task, the 4 Nearest Neighbours were computed.

The INEX 2007 collection contains several links between documents, like *unknownlinks*, *languagelinks* and *outsidelinks* for instance. We only considered existing relations between doxels with the *collectionlink* tag, because these links denote links inside the collection. We use the $xlink : href$ attribute that indicates the target (file name) of the link. We notice that the targets of such links are only whole documents, and not documents parts; this aspect may negatively impact our expectations compared to our model that supports documents parts as targets. The table 1 shows these relations, with a first document D_1 (file 288042.xml) about "Croquembouche" and a second document D_2 (file 1502304.xml) about "Choux pastry". The third *collectionlink* tag in D_1 links D_1 to D_2; the source of this link is underlined in D_1 in the table and the target is the whole document D_2 which is also underlined. Overall, there are 17 013 512 collectionlinks in the INEX 2007 collection. We applied the following restrictions:

- for each leaf doxel d: the 4 first collectionlinks of d,
- for non-leaf doxels d': the union of 4 first collection links of its leaf doxels direct or indirect components.

With the restrictions above, we only take into account 12 352 989 collectionlinks.

Table 1. An example of collectionlinks from the INEX2007 corpus

Document D_1: file 288042.xml

```
<article>
<name id="288042">Croquembouche</name>
...
<body>A
<emph3>croquembouche</emph3>is a
<collectionlink ... xlink:href="10581.xml">French</collectionlink>
<collectionlink ... xlink:href="57572.xml">cake</collectionlink>
consisting of a conical heap of cream-filled

<collectionlink ... xlink:href=1502304.xml’>choux</collectionlink>

buns bound together with a brittle
<collectionlink ... xlink:href="64085.xml">caramel</collectionlink>
sauce, and usually decorated with ribbons or spun sugar.
...
</body>
</article>
```

Document D_2: file 1502304.xml

```
<article>
<name id="1502304">Choux pastry</name>
...
<body>
<emph3>Choux pastry</emph3>
<emph2>(pte  choux)</emph2>is a form of light
<collectionlink ... xlink:href="67062.xml">pastry</collectionlink>
used to make
<collectionlink ... xlink:href="697505.xml">profiterole</collectionlink>
s or
<collectionlink ... xlink:href="1980219.xml">eclair</collectionlink>
s. Its
<collectionlink ... xlink:href="198059.xml">raising agent</collectionlink>
is the high water content, which boils during cooking, puffing
out the pastry.
...
</body>
</article>
```

3 Doxel Space

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 [2]. 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. To avoid an unmanageable quantity of doxels, we kept only doxels having the following tags: article, p, collectionlink, title, section, item. The reason for using only these elements was because, except for the collectionlinks, we assume that the text content for these doxels are not too small. The overall number of doxels considered by us here is 29 291 417.

3.2 Doxel Context

Consider the two structured documents D_1 and D_2 linked as shown in table 1: they share *apriori* information. If a user looks for all the information about "croquembouche", the system should indicate if the link from D_1 to D_2 is relevant for the query. If the user only wants to have general informations about "croquembouche", 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 [4]. 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. But $Spe(d_1, d_2)$ and $Exh(d_1, d_2)$ could be high both if d_1 and d_2 are encapsulated and deal with the same subject.

We propose to describe relative specificity and relative exhaustivity between two doxels d_1 and d_2 as extensions of the overlap function [5] of their index: these values reflect the amount of overlap between the source and target of the relation. We define relative specificity and relative exhaustivity in formulas (1) and (2) on the basis of the non normalized doxel vectors $w_{1,i}$ and $w_{2,i}$ (respectively for d_1 and d_2).

$$Exh(d_1, d_2) = \frac{\sum_i w_{1,i} \cdot w_{2,i}}{\sum_i w^2_{\oplus 1/w_{2,i}}} \qquad (1)$$

$$Spe(d_1, d_2) = \frac{\sum_i w_{1,i} \cdot w_{2,i}}{\sum_i w^2_{\oplus 2/w_{1,i}}} \qquad (2)$$

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

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

4 Model of Matching in Context

We assume that the matching process should return doxels relevant to the user's information needs, regarding both content, structure aspects, and considering also 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:

$$RSV(d, q) = \alpha * RSV_{content}(d, q) + (1 - \alpha) * RSV_{context}(d, q), \tag{3}$$

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

$$RSV_{context}(d, q) = \sum_{d' \in \mathcal{V}_d} \frac{\beta * Exh(d, d') + (1 - \beta) * Spe(d, d')}{|\mathcal{V}_d|} RSV_{content}(d', q) \tag{4}$$

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. Using a linear combination makes sense, as a doxel may be relevant *per se* without any other relevant context but a relevant context may increase the relevance of a doxel.

5 Experiments and Results

The INEX 2007 Adhoc track consists of three retrieval tasks: the Focused Task, the Relevant In Context Task, and the Best In Context Task. We submitted 3 runs for each of these tasks. For all these runs, we used only the *title* of the INEX 2007 queries as input for our system: we removed the words prefixed by a '-' character, and we did not consider the indicators for phrase search. The vocabulary used for the official runs is quite small (39 000 terms), but was assumed large enough to prove the validity of our proposal.

First of all, we have experimented our system with INEX 2006 collection to fix α and β parameters of formulas (3) and (4). The best results were achieved with a higher value for the exhaustivity than for the specificity. As a consequence, we decide to fix $\alpha = 0.75$ and $\beta = 0.75$ for our expected best results.

5.1 Focused Task

The INEX 2007 Focused Task is dedicated to find the most focused results that satisfy an information need, without returning "overlapping" elements. In our focused task, we experiment with two different rankings.

For the first run, the "default" one, namely *LIG_075075_FOC_FOC* with $\lambda = 0.75$ and $\beta = 0.75$, we rank the result based on matching in context proposed in section 4; overlap is removed by applying a post-processing.

For the second run, we choose to use the results of the Relevant In Context Task to produce our Focused Task results : relevant doxels are ranked by article, and we decide to score the doxels with the score of each corresponding article and list them according to their position in the document, and removing overlapping doxels. This run is called *LIG_075075_FOC_RIC*, and we set $\lambda = 0.75$ and $\beta = 0.75$.

The last run, namely *LIG_1000_FOC_RIC* is our baseline. It is similar to the second run with $\lambda = 1.0$ and $\beta = 0.0$, i.e. it considers only the contents of the doxels.

Table 2. Focused Task for INEX2007 Ad Hoc

Run	precision at 0.0 recall	precision at 0.01 recall	precision at 0.05 recall	precision at 0.10 recall
LIG_075075_FOC_FOC $MAiP = 0.0158$	0.3107	0.1421	0.0655	0.0492
LIG_1000_FOC_RIC $MAiP = 0.0580$	0.3540	0.3192	0.2119	0.1734
LIG_075075_FOC_RIC $MAiP = 0.0647(+11.6\%)$	0.3475 (-1.8%)	0.3144 (-1.5%)	0.2480 (+17.0%)	0.2126 (+22.6%)

We present our results for the focused task in Table 2 showing precision values at given percentages of recall, and in Figure 1 showing the generalized precision/recall curve. These results show that runs based on Relevant In Context approach outperforms the "default" Focused Task run, *LIG_075075_FOC_FOC*: after checking the code, we found a bug that leads to incorrect paths for the doxels, and this bug impacts in a lesser extent the second run. The first column of the Table 2 shows that, considering the Mean Average Interpolated Precision, the *LIG_075075_FOC_RIC* run outperforms the *LIG_1000_FOC_RIC* run by +11.6%, proving that the collectionlinks are usefull. Moreover, in Table 2 and in Figure 1, we see that for the results between 0.05 recall and 0.25 recall, the *LIG_075075_FOC_RIC* performs much better than the *LIG_1000_FOC_RIC*. Our best run is ranked 60 on 79 runs.

5.2 Relevant in Context Task

For the Relevant In Context Task, we take "default" focused results and reordered the first 1500 doxels such that results from the same document are

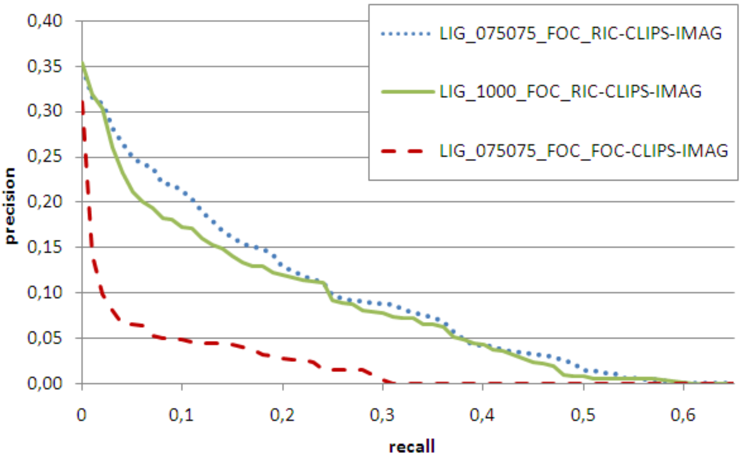

Fig. 1. Interpolated Precision/Recall - Focused Task

clustered together. It considers the article as the most natural unit and scores the article with the score of its doxel having the highest RSV.

We submitted three runs:

- *LIG_1000_RIC* : a baseline run which doesn't take into account the inner collectionlinks to score doxels. We set $\lambda = 1.0$ and $\beta = 0.0$;
- *LIG_075075_RIC* : a retrieval approach based on the collectionlinks use. We set $\lambda = 0.75$ and $\beta = 0.75$;
- *LIG_00075_RIC* : an approach that consider the RSV of a doxel only considering its context: we set $\lambda = 0.0$ and $\beta = 0.75$.

Table 3. Relevant In Context Task for INEX2007 Ad Hoc

Run	gP[5]	gP[10]	gP[25]	gP[50]
LIG_1000_RIC $MAgP = 0.0329$	0.0926	0.0826	0.0599	0.0448
LIG_075075_RIC $MAgP = 0.0424$ (+28.9%)	0.1031 (+11.3%)	0.0957 (+15.9%)	0.0731 (+22.0%)	0.0542 (+21.0%)
LIG_00075_RIC $MAgP = 0.0174$ (-47.1%)	0.0779 (-15.9%)	0.0581 (-29.7%)	0.0401 (-33.1%)	0.0291 (-35.0%)

For the relevant in context task, our results in terms of non-interpolated generalized precision at early ranks $gP[r], r \in \{5, 10, 25, 50\}$ and non-interpolated Mean Average Generalized Precision $MAgP$ are presented in Table 3. Figure 2 shows the generalized precision/recall curve. This shows that using collectionlinks and the doxels content (*LIG_075075_RIC*) improves the baseline by a ratio greater than 11%. The *LIG_00075_RIC* gives bad results, showing that the context of the doxels only is not relevant. In Figure 2, we see that the *LIG_075075_RIC* run is also above the default run. Our best run is ranked 56 on 66 runs.

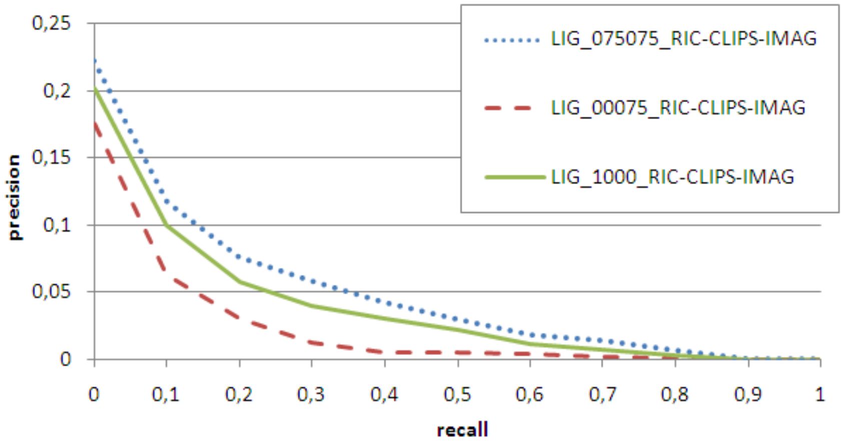

Fig. 2. Generalized Precision/Recall - Relevant In Context task

5.3 Best in Context Task

For the Best In Context Task, we examine whether the most focused doxel in a relevant document is the best entry point for starting to read relevant articles. We take "normal" focused results and the first 1500 doxels belonging to different files. For this task, we submitted three runs:

- *LIG_1000_BIC* : the baseline run which doesn't take into account collectionlinks: we set $\lambda = 1.0$ and $\beta = 0.0$;
- *LIG_075075_BIC* : the retrieval approach based on the use of collectionlinks. We set $\lambda = 0.75$ and $\beta = 0.75$;
- *LIG_00075_BIC* : the approach that uses only the context of doxels to compute their RSV: we set $\lambda = 0.0$ and $\beta = 0.75$.

Table 4. Best In Context Task for INEX2007 Ad Hoc

Run	gP[5]	gP[10]	gP[25]	gP[50]
LIG_1000_BIC $MAgP = 0.0630$	0.1194	0.1176	0.1035	0.0910
LIG_075075_BIC $MAgP = 0.0761$ (+20.8%)	0.1373 (+15.0%)	0.1261 (+7.2%)	0.1151 (+11.2%)	0.0957 (+5.2%)
LIG_00075_BIC $MAgP = 0.0639$ (+1.4%)	0.1303 (+9.1%)	0.1107 (-5.9%)	0.0977 (-5.6%)	0.0819 (-0.1%)

For the best in context task, our results are presented in Table 4 and Figure 3 with the same measures as the Relevant In Context Task results. Our best run is ranked 54 on 71. Conclusions are the same: using collectionlinks and content improves the baseline by a mean average of more than 20%, and the *LIG_00075_BIC* run is consistently below the baseline. There is one result however, the *LIG_00075_BIC* run outperforms the baseline at $gP[5]$ by more than

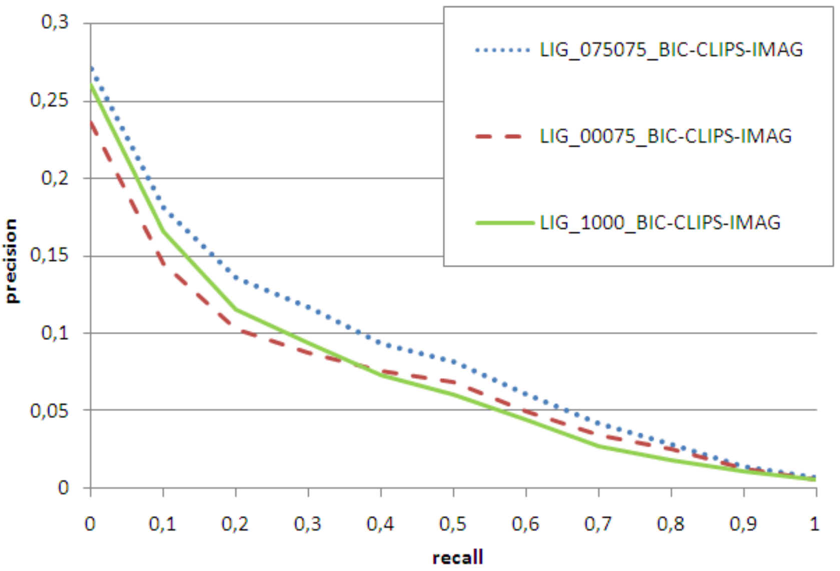

Fig. 3. Generalized Precision/Recall - Best In Context task

9% and in Figure 3 we see than the baseline and the *LIG_00075_BIC* are quite close to each others. This means that the *a priori* links are really meaningful.

6 Summary and Conclusion

We proposed a way to integrate the content of the doxels as well as their context (collectionlinks in INEX 2007 documents). We have submitted runs implementing our theoretical proposals for the different Ad Hoc tasks. For each of the tasks, we showed that combining content and context produce better results than considering content only and context only of the doxels, which is a first step in validating our proposal. According to the official evaluation of INEX 2007, our best runs are ranked in the last third of participants systems, for the Content-Only runs. However, we plan to improve our baseline to obtain better results in the following directions:

- As mentioned earlier, the size of the vocabulary used is too small, leading to query terms out of our vocabulary.
- When submitting our runs for our first participation at INEX competition we found some bugs related to the identifiers of the doxels, so the results were negatively impacted.
- We are working on the integration of negative terms in the query, in a way to get better results.

Since the submission of our official runs, we integrated a larger vocabulary (about 200 000 terms) and corrected our bugs, which led to an increase of 24%

for the MAiP, when using the official evaluation tool released in december 2007 and the version 2.0 of the assessments.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1–7), 107–117 (1998)
2. Fang Huang, D.H., Watt, S., Clark, M.: Robert Gordon University at INEX 2006: Adhoc Track. In: INEX 2006 Workshop Pre-Proceeding, pp. 70–79 (2006)
3. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM 46(5), 604–632 (1999)
4. Piwowarski, B., Lalmas, M.: Interface pour l'evaluation de systemes de recherche sur des documents XML. In: Premiere COnference en Recherche d'Information et Applications (CORIA 2004), Toulouse, France, Hermes (2004)
5. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval, ch. 6, p. 203. McGraw-Hill, Inc., New York (1986)
6. Savoy, J.: An extended vector-processing scheme for searching information in hypertext systems. Inf. Process. Manage. 32(2), 155–170 (1996)
7. Smucker, M.D., Allan, J.: Using similarity links as shortcuts to relevant web pages. In: SIGIR 2007: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 863–864. ACM Press, New York (2007)
8. van Rijsbergen, C.: Information retrieval, 2nd edn., ch. 3. Butterworths (1979)
9. Verbyst, D., Mulhem, P.: Doxels in context for retrieval: from structure to neighbours. In: SAC 2008: Proceedings of the 2008 ACM symposium on Applied computing. ACM Press, New York (2008)