A Content-Based Link Detection Approach using the Vector Space Model

Junte Zhang[1] and Jaap Kamps[1,2]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] ISLA, Faculty of Science, University of Amsterdam

Abstract. Link detection can be seen as a special application of Focused Retrieval. This paper presents a content-based link detection approach using the Vector Space Model. We present our results, and conclude by discussing the merits and deficiencies of our approach.

1 Introduction

This paper reports on our participation in the Link The Wiki (LTW) track of INEX. LTW is aimed at detecting or discovering missing links between a set of Wikipedia topics, and the remainder of the collection, hence effectively establishing cross-links between those documents using IR techniques. Existing links were removed from the topics, making these documents 'orphans' that could be linked to potential 'fosters'. This means that hypertext has be constructed automatically. Many hypertext systems have been based on the *Dexter Hypertext Reference Model* [2], and subsequently our system as outlined in this paper is also compliant with this model.

LTW consisted of two tasks. The first task was a continuation of the track of last year with the detection of links between whole files. The second task used 50 selected orphan topics, and went further than link detection on the document level, as links had to be established between spans of characters within one document and spans of characters with another document. The latter is here a *Best Entry Point* (BEP), i.e. the best point where the user can start reading in a document, which makes link detection particularly a special application of Focused Retrieval. A maximum of 5 BEPs per anchor value was allowed. What both tasks had in common was that it consisted of 2 sub-tasks; the detection of links from an 'orphan' (outgoing) and to an 'orphan' (incoming).

Detected links are treated as uni-directional hyperlink arcs. The issue of link density and link repetition as mentioned in [3] has not been addressed, henceforth we restricted our experimentation to detecting unique cross-links between documents. In Sections 2 and 3, we present our approaches. The results are presented and discussed in Section 4, and we conclude with our findings in Section 5.

2 Detection of Document-to-Document Links

We employ a content-based (and thus collection-independent) approach with IR techniques as previously outlined in [1]. This means we do not rely on learning,

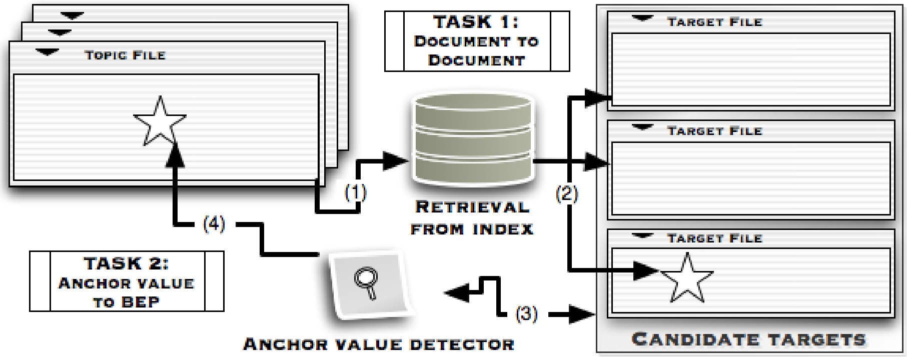

Fig. 1. System overview of a content-based link detection approach.

heavy heuristics or existing link structures in the Wikipedia, and *only* use the orphaned topics as evidence. An overview of our system is depicted in Fig. 1.

We adopt a *breadth m–depth n* technique for automatic text structuring for identifying the anchor values and links, i.e. a fixed m number of documents accepted in response to a query (step 1) and a fixed n number of iterative searches (step 2). The similarity on the document level and text segment level (substrings of a line) is used as evidence. We used the whole document (i.e. full-text content) as a query (not only title), because in prior experiments we found that this performed best. The standard Vector Space Model (VSM) implementation of Lucene was used for retrieval, i.e., for a collection D, document d, and query q:

$$sim(q,d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} \cdot coord_{q,d} \cdot weight_t, \qquad (1)$$

where $tf_{t,X} = \sqrt{\mathrm{freq}(t,X)}$; $idf_t = 1 + \log \frac{|D|}{\mathrm{freq}(t,D)}$; $norm_q = \sqrt{\sum_{t \in q} tf_{t,q} \cdot idf_t{}^2}$; $norm_d = \sqrt{|d|}$; and $coord_{q,d} = \frac{|q \cap d|}{|q|}$.

Before the actual link detection starts, some pre-processing is done by extracting for each topic the title enclosed within the `<name>` tag and storing that in a hash-table for substring matching. We do not apply case-folding, but we do remove any existing disambiguation information put between brackets behind the title. Only titles of > 3 characters length are considered.

We do not assume that links are reciprocal or bi-directional, so we have different approaches for detecting outgoing and incoming links. A threshold of 250 was set for both types of links, and repeated links were not allowed. Links also appear locally within a document to improve the navigation there, but this was outside the scope of the LTW track. So there is (a) an *outgoing link* for an 'orphan' topic when the title of a 'foster' document occurs in the orphan topic. and (b) there is an *incoming link* for an orphan when the title of the orphan occurs in a foster document. We describe the following 2 runs:

a2a_1 The whole orphan document is used as a query. The pool of plausible 'foster' (candidate) documents is the top 300 returned by this query.

a2a_3 The whole orphan document is used as a query. The pool of plausible candidate links is the top 500 of the ranked list.

3 Detection of Anchor-to-BEP Links

The Anchor-to-BEP task is based on a hypertext mechanism called *anchoring* [2]. The actual *anchor value* had to be specified using the File-Offset-Length (FOL) notation, which at the same time serves as the *anchor identifier* [2]. At the same time, the BEP of the outgoing link had to be provided. For all of these runs, we assume that the BEP is always the start of the document (i.e. offset = 0). Multiple links per anchor were only computed for the run `a2bep_5`.

a2bep_1 The whole orphan document is used as query, and the top 300 results is used to find potential cross-links.

a2bep_3 The whole orphan document is used as query. The top 50 ranking documents is harvested. Each of these documents is used again as a query to retrieve its top 6 results; resulting in 300 foster documents.

a2bep_5 This run is similar to the first Anchor-to-BEP run, but we expanded this run by allowing more than 1 BEP for each anchor. We use the depth-first strategy, and the broader-narrower conceptualization of terms by re-grouping the extracted list of titles based on a common substring. For example, the anchor value *"Gothic"* could refer to the document "Gothic", but also to documents with the titles *"Gothic alphabet"*, *"Gothic architecture"*, *"Gothic art"*, *"Gothic Chess"*, and so on.

4 Experimental Results and Discussion

Our results are evaluated against the set of existing links (in the un-orphaned version of the topics) as ground truth, both for the sample of 6,600 topics in the first task, as well as the 50 topics in the second task. The results of our runs are depicted in Table 1 for links on the document-level and in Table 2 for the Anchor-to-BEP links. Additionally, the outgoing Anchor-to-BEP links were assessed manually (see Table 3), so there are no results for the incoming links.

Generally, our approach performed better for detecting incoming links than outgoing ones. We achieved the highest early precision for incoming links detection. Table 3 suggests that the existing links in the Wikipedia do not suffice or is a spurious ground truth given the user assessments, where MAP = 0.27653 for Document-to-Document links, and MAP = 0.20790 for Anchor-to-BEP links. For example, when we compare the scores of automatic (Table 2) vs manual evaluation (Table 3) of outgoing links, we see that the actual set of detected links is only a small subset of what users really want.

These results, especially the sub-optimal results for the outgoing links and the general results on the document-level, warrant some reflection on several

Table 1. Document-to-Document runs with Wikipedia as ground truth.

Run	Links	MAP	R-Prec	P@10	Rank
a2a_1	In	0.33927	0.35638	0.57082	15/24
a2a_3		0.35758	0.37508	0.58585	14/24
a2a_1	Out	0.10716	0.17695	0.19061	15/21
a2a_3		0.10174	0.16301	0.17073	19/21

Table 2. Anchor-to-BEP runs with Wikipedia as ground truth.

Run	Links	MAP	R-Prec	P@10	Rank
a2bep_1	In	0.23495	0.25408	0.80400	7/27
a2bep_3		0.15662	0.16527	0.77400	23/27
a2bep_5		0.23495	0.25408	0.80400	8/27
a2bep_1	Out	0.09727	0.20337	0.27400	20/30
a2bep_3		0.09106	0.18296	0.32800	23/28
a2bep_5		0.14262	0.24614	0.47000	14/30

Table 3. Anchor-to-BEP runs based on manual assessments.

Run	Links	MAP	R-Prec	P@10	Rank
Wikipedia	Out	0.20790	0.31258	0.45996	1/28
a2bep_1		0.05557	0.12511	0.14195	23/28
a2bep_3		0.05181	0.13368	0.18699	24/28
a2bep_5		0.08472	0.16822	0.31773	16/28

limitations of our approach. We did exact string matching with the titles of the candidate foster topics and did not apply case-folding or any kind of normalization. This means we could have incorrectly discarded a significant number of relevant foster documents (false negatives). Moreover, we could missed a significant number of linkable candidates in step 1 due to the limitations of the VSM. Conversely, this means effectively under-generating the incoming and outgoing links, however, for task 1 we over-linked the outgoing links in the topics (see Tables 4 and 5). Interestingly, we found that we can significantly improve the accuracy of the detection of our outgoing links by generating multiple BEPs for an anchor, which partly deals with the issue of underlinking.

5 Conclusions

In summary, we continued with our experimentation with the Vector Space Model and simple string processing techniques for detecting missing links in the Wikipedia. The link detection occurred in 2 steps: first, a relevant pool of foster (candidate) documents is collected; second, substring matching with the

Table 4. Number of Document-to-Document links.

Links	Measure	Qrel	a2a_1	a2a_3
In	Mean	35.66	17.04	19.66
	Median	21	9	9
Out	Mean	36.31	109.09	123.76
	Median	28	95	110

Table 5. Number of Anchor-to-BEP links.

Links	Measure	Qrel (auto)	a2bep_1	a2bep_3	a2bep_5
In	Mean	278.32	62.96	23.2	62.96
	Median	134	29.5	17	29.5
Out	Mean	79.18	36.82	25.72	26.02
	Median	62	41	24	26.5

list of collected titles to establish an actual link. We used entire orphaned documents (full-text) as query, with the idea to use all textual content as maximal evidence to find 'linkable' documents.

Clearly, we showed the limitations of this full-text approach based on the VSM, especially on the document level. A content-based full-text approach is not competitive against anchor-based approaches, however, a content-based approach adheres most strictly to an obvious assumption of link detection, namely that documents do not already have existing links as evidence and these cannot be used to 're-establish' links, which is not necessarily equal to 'detection'.

A competitive content-based link detection approach that discovers high quality links is needed, for example for detecting links in legacy or cultural heritage data. The impact of link detection on those datasets and domains will be large (for users and systems), since there are no such links yet (which would enable new navigation and search possibilities), and the alternative is expensive manual linking. To improve our approach, we are considering to experiment more with the granularity in a document to find focusedly link candidates (besides title and whole document), such as on the sentence level.

Acknowledgments This research is supported by the Netherlands Organization for Scientific Research (NWO) under project number 639.072.601.

Bibliography

[1] K. N. Fachry, J. Kamps, M. Koolen, and J. Zhang. Using and detecting links in wikipedia. In *INEX 2007*, pages 388–403. Springer-Verlag, 2008.

[2] F. Halasz and M. Schwartz. The Dexter hypertext reference model. *Commun. ACM*, 37(2):30–39, 1994.

[3] J. Zhang and J. Kamps. Link detection in XML documents: What about repeated links? In *SIGIR 2008 Workshop on Focused Retrieval*, pages 59–66, 2008.