Linking Archives Using Document Enrichment and Term Selection (Abstract)*

Marc Bron
m.m.bron@uva.nl

Bouke Huurnink
b.huurnink@uva.nl

Maarten de Rijke
derijke@uva.nl

ISLA, University of Amsterdam
Science Park 904, 1098 XH Amsterdam

1. INTRODUCTION

News, multimedia, and cultural heritage archives are opening up and publishing their content online, enabling users to search for items of interest across multiple archives. With the general public gaining access to archive content, an increasing number of users can be expected to exhibit exploratory behavior [1], rather than the directed search typical of professional users [3]. In order to make archives accessible to the general public, modes of access supporting exploratory behavior should be examined.

One way to enable exploration over (multiple) archives is to create links between individual items. Here we aim to connect an item from one archive to items in another archive that discuss the same or related events. Links to items describing the same event allow users to access different views of the same event, while links to items describing related events allow users to explore interconnected relationships between events. Due to space limitations we only discuss same event linking in this compressed contribution. We focus on a specific instance of the task: linking items from a newspaper archive with rich textual representations to items from a multimedia archive that tend to have sparse annotations. This scenario gives rise to two challenges: first, the targets of our linking task—archived multimedia items—are relatively sparsely annotated which leads to recall problems. Second, the source of a link is a news article in a news archive; such articles may be long and discuss issues that are only indirectly related to the seminal event that triggered the article, thereby potentially giving rise to a precision problem. In this setting we seek answers to the following two research questions: (i) does expanding sparse item representations with text from other sources improve linking performance; and (ii) what effect does modeling reduced versions of the original richly represented source item have on linking performance?

2. APPROACH

Same event linking. Given event e, described by a source item s from a source archive A_s with rich text representations, create links to target items $T = \{t_1, \ldots, t_n\}$ in a target archive A_t, where the event described by each $t_i \in T$ is the *same* as e. We use a definition of event that makes a distinction between *seminal events*, i.e., high impact news events that generate follow-up events, and *related events* that are caused by or predict the seminal event but are not seminal events by themselves.[1] An item is *textually rich* when it contains, on top of human-annotated metadata, textual content. *Sparse* representations only contain human-annotated metadata.

Linking model. We compute the similarity between the textual representation of a fixed source item s and the representation of every potential target item t in the target archive and rank each t accordingly. As similarity function we use the vector space model:

$$\text{sim}(s,t) = \frac{\vec{V}(s)\cdot\vec{V}(t)}{|\vec{V}(s)||\vec{V}(t)|}, \quad (1)$$

here $\vec{V}(s)$ and $\vec{V}(t)$ are vector representations of s and t, the numerator is a dot product of the vectors and $|\vec{V}|$ is the length of $\vec{V}$.

Document expansion. To address sparseness of the representation of a target item t we use other items x for expanding the representation of t. To obtain expansion items x we compute the similarity between $t \in A_t$ and each item in expansion archive A_x and rank its items by similarity, as in (1). The resulting ranked list is cut off at some rank m to yield a list of expansion items; these are then concatenated to t to form an expanded representation of t.

Selecting representative terms. To select terms, we take the top $k\%$ terms from s ranked by their TFIDF score, which is defined as:

$$TFIDF(a) = \frac{c(a,d)}{|d|} \cdot \log\left(\frac{|D|}{|\{d \in D : d \text{ contains } a\}|}\right),$$

where $c(a,d)$ gives the count of term a in document d, $|d|$ is the document length and $|D|$ is the size of the document collection.

Selecting representative entities. To select entities we apply a named entity recognizer [2] based on conditional random fields to the content of all source archive items. We then select the top $k\%$ entities based on their TFIDF value as with the terms.

Date filter. We use a simple date filter that only allows a link from a source item s to a target item t if t's date is within an N day window around the date of s.

3. EXPERIMENTAL SETUP

Evaluation collection. Our evaluation collection consists of a source archive containing textually rich newspaper articles and a target archive of textually sparse television news broadcasts to which we want to link. We single out a set of source items as our test cases for linking, and for each test source item, we have a set of relevance judgments indicating which items in the target archive refer to the

*The full version of this paper appeared in *TPDL 2011*.

[1] http://projects.ldc.upenn.edu/TDT5/Annotation/TDT2004V1.2.pdf

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$10.00.

same seminal event. As our baseline we perform linking using all of the source item's content and metadata without term selection to representations of target items without expansion.

Expanding sparse text representations. We investigate the effect of increasing the number of documents used to expand target. We experiment with three sources of information: the target archive itself (expanding target items with representations from other items in the archive); Wikipedia, the online encyclopedia; and the richly represented, news-focused items in the source archive.

Term selection for rich text representations. Here we evaluate the result of reducing the amount of text in a source item on linking performance. First, we experiment with using newspaper article structure to reduce the source item representation. We then experiment with using only the most representative unique terms and named entities in the source item. Finally, we experiment with using the optimal combination of these options and a date filter.

Evaluation measures and significance testing. We use three evaluation metrics: Mean Average Precision (MAP), precision at rank five (P@5) and mean Reciprocal Rank (MRR). We use a paired t-test to determine significant differences between results, where $^{\triangle}$ or $^{\triangledown}$ ($^{\blacktriangle}$, $^{\blacktriangledown}$) indicate whether a score is significantly higher or lower than the baseline with a significance level of $\alpha < .05$ ($\alpha < .01$).

4. RESULTS

Document expansion. We first contrast different archives for expansion, i.e., the source archive, target archive, and Wikipedia. Table 1 shows the scores when expanding with the optimal number of expansion documents from each archive. We find that expanding with documents from archives other than the source archive does little to improve performance. Expansion with the optimal number of documents from the source archive yields a significant improvement over the baseline. We note that although optimized for MAP, the other early precision metrics follow the same trend in that the optimal number of documents for MAP is also the optimal number for the other metrics. The P5 scores for do improve (by 40.9%), but remain relatively low; this is due to the small number of relevant target items per test item (on average 2.4).

Table 1: Expansion results; significance tested against baseline.

Exp. Model	*detail*	*MAP*	*P5*	*MRR*
baseline	–	.3623	.2000	.4819
n target docs	$n = 3$	.3907	.2227	.4654
n wikipedia docs	$n = 2$	.3964	.2136	.4425
n source docs	$n = 7$	.4949$^{\triangle}$	.2818	.5435

Term selection. On the *source* item side we experiment with different term selection techniques. In this experiment, we link to the original unexpanded target items. Table 2 shows that using only terms from a specific field, e.g., lead or title, improves over using the whole document in terms of absolute scores, but not significantly so. We also select terms and named entities from the content of the *source* item based on their TFIDF score. Table 2 shows that with the optimum of 60% of the terms selected from the content of the source item ($x\%\, terms$), a significant improvement over the baseline is achieved. When considering named entities ($y\%\, ne$) it turns out to be harmful to remove any entities from the representation. The combination of selecting terms and named entities does not improve over selecting terms alone.

Table 2: Selection results; significance tested against baseline.

TS Model	*detail*	*MAP*	*P5*	*MRR*
baseline	–	.3623	.2227	.4820
content	–	.3582	.1955	.4800
metadata	–	.1636$^{\blacktriangledown}$	.0636$^{\blacktriangledown}$	.1863$^{\blacktriangledown}$
title	–	.4157	.2227	.4597
lead	–	.4428	.2318	.5386
x% terms	$x = 60\%$	.5133$^{\triangle}$	.2682	.6390
y% ne	$y = 100\%$	.4374	.2091	.5592
combined	x=60%, y=100%	.4660	.2409	.5849

Further improving linking performance. In order to see how far we can push linking performance we conduct two additional experiments. In the first we combine the best models, i.e., the best term selection is used to find targets and the target items have been expanded with the optimal number of documents. The combination achieves a MAP of .4801, which does not improve over using document expansion (.4949) or term selection (.5133) by itself. We find that for items where document expansion helps, term selection has relatively poor performance, and vice versa. This fits the intuition that term selection and expansion have opposite effects: one makes an item's event description more specific, while the other broadens the description. Depending on the source item only one of the effects may be desired. Our second experiment is with a date filter that restricts target items to a period of 14 days around the date of the source item. This results in a baseline MAP score of .5689 and scores of .7263 and .7397 MAP for the best document expansion and term selection models, respectively. Scores for all models go up, including the baseline, but the same significant differences in performance remain between the baseline and the best models.

5. CONCLUSIONS

We use a retrieval approach to link items from a news paper archive with very rich text descriptions to videos in a multimedia archive with relatively sparse annotations. We find that expanding *target* items with documents from other sources improves performance. Using expansion documents from the source archive is most effective however, as the content has the same focus as the target archive. Additionally we find that reducing the number of terms in the *source* item representation is effective. The reduced items are more robust to topic drift and form a better match for the short event descriptions in the target archive.

Acknowledgements. This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430 (GALATEAS), by the PROMISE Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.061.814, 612.061.815, 640.004.802, 380-70-011 and by the Center for Creation, Content and Technology (CCCT).

6. REFERENCES

[1] M. Bron, J. van Gorp, F. Nack, and M. de Rijke. Exploratory search in an audio-visual archive: Evaluating a professional search tool for non-professional users. In *EuroHCIR 2011: 1st Eur. Worksh. Human-Comp. Interact. and Inf. Retr.*, July 2011.

[2] J. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05*, pages 363–370. ACL, 2005.

[3] B. Huurnink, L. Hollink, W. van den Heuvel, and M. de Rijke. Search behavior of media professionals at an audiovisual archive: A transaction log analysis. *J. American Soc. Information Science and Technology*, 61(6):1180–1197, 2010.