Part I
Tracking Content

The first part of the thesis focuses on the first research theme: *tracking content*. A key process within content tracking is link generation where links need to be established between a source document (the document we want to track), and other documents within the same or other domains. This process can be repeated for every document in the collection for generating a hierarchy of links starting from the source document. In the setting of timestamped collections, such as social media, this hierarchy can be laid out over time, resulting in an analogous representation to a discussion thread as witnessed in fora or commenting facilities and as such can reveal the evolution of the discussion around a source document. Analysis of this hierarchy of content over time can prove helpful for a range of tasks that we have reviewed in Chapter 2.

Our focus in the next two chapters is to develop robust methods for cross-domain link generation. We concentrate on two particular domains, these of online news and social media. Our choice is motivated from the symbiotic relationship of the two. One of the challenges here is to bridge the vocabulary gap between edited (news) and unedited (social media) content; we will elaborate more on this in the next chapter. In particular, in Chapter 3 we undertake the task of discovering implicitly linked social media utterances for a news article using an IR approach: a source article is regarded as a query which is issued to an index of social media utterances. Then, the retrieved utterances are candidate links to the source article. Our main concern is on how to model the source article for bridging the vocabulary gap between the two domains. For this purpose, we use content from utterances across several social media platforms that are explicitly linked to the article (i.e., they provide a hyperlink to it). In Chapter 4 we revisit this task with a focus on social media utterances that repost verbatim or near-verbatim a source article. We use the same IR approach as before, however, this setting is very different from ad hoc IR because of the different average query length of queries. In ad hoc retrieval, query lengths are typically a small fraction of the length of the documents to be retrieved, however, here, the query (source article) and the utterances to be retrieved are of similar length. This characteristic violates the assumption behind the widely used query likelihood model in IR, i.e., sampling query terms from a document with replacement. We provide a remedy by sampling query terms without replacement.