Time-aware Approaches to Information Retrieval

Nattiya Kanhabua
Norwegian University of Science and Technology
nattiya@idi.ntnu.no

Abstract

In this thesis, we address major challenges in searching *temporal document collections*. In such collections, documents are created and/or edited over time. Examples of temporal document collections are web archives, news archives, blogs, personal emails and enterprise documents. Unfortunately, traditional IR approaches based on term-matching only can give unsatisfactory results when searching temporal document collections. The reason for this is twofold: the contents of documents are strongly time-dependent, i.e., documents are about events happened at particular time periods, and a query representing an information need can be time-dependent as well, i.e., a *temporal query*.

Our contributions in this thesis are different time-aware approaches within three topics in IR: content analysis, query analysis, and retrieval and ranking models. In particular, we aim at improving the retrieval effectiveness by 1) analyzing the contents of temporal document collections, 2) performing an analysis of temporal queries, and 3) explicitly modeling the time dimension into retrieval and ranking.

Leveraging the time dimension in ranking can improve the retrieval effectiveness if information about the creation or publication time of documents is available. In this thesis, we analyze the contents of documents in order to determine the time of non-timestamped documents using temporal language models. We subsequently employ the temporal language models for determining the time of implicit temporal queries, and the determined time is used for re-ranking search results in order to improve the retrieval effectiveness.

We study the effect of terminology changes over time and propose an approach to handling terminology changes using time-based synonyms. In addition, we propose different methods for predicting the effectiveness of temporal queries, so that a particular query enhancement technique can be performed to improve the overall performance. When the time dimension is incorporated into ranking, documents will be ranked according to both textual and temporal similarity. In this case, *time uncertainty* should also be taken into account. Thus, we propose a ranking model that considers the time uncertainty, and improve ranking by combining multiple features using learning-to-rank techniques.

Through extensive evaluation, we show that our proposed time-aware approaches outperform traditional retrieval methods and improve the retrieval effectiveness in searching temporal document collections.

Available online at: `http://www.idi.ntnu.no/research/doctor_theses/nattiya.pdf`