UNIVERSITEIT VAN AMSTERDAM

Exploratory Search over Semi-structured Documents

H. Azarbonyad

The Web is a major source of information for many users. Automatically grouping and classifying documents on the Web is an important ingredient when trying to support effective search of Web documents. Documents on the Web are a mixture of content (text, image, etc.), structure, and metadata. Beside content, metadata and structure can be helpful in managing documents and supporting exploratory search over them. In this thesis, we aim at empowering content-based approaches—for managing documents and exploratory search—with structure and metadata and study how metadata and structure associated with documents can help to both manage documents more accurately and support different exploratory search tasks.

In the first part of the thesis, we study how we can leverage metadata and structure to manage documents. We define document management as the task of grouping similar documents together and classifying them. We define the similarity from three different angles in the three chapters of the first part. We first start by classifying documents based on their topical similarity and use content, structure, and metadata associated with documents to classify them. We show that a metadata/structure powered classifier performs much better than a classifier that is solely based on the content. We then focus on classifying documents based on their topical diversity. We measure topical diversity by means of topic models. In doing this, we characterize the main drawbacks of topic models when they are applied to the task of measuring topical diversity and propose a hierarchical approach to address these drawbacks. We find that the topic models achieved using our hierarchical approach not only have a superior performance in the topical diversity task compared to the previous approaches, but also are useful in other tasks such as classification and clustering of topically similar documents. Finally, we move to the email domain and study how we can help users automatically manage their tasks created via email. We focus on the task of detecting commitments made in email as it enables digital assistants to help their users recall promises they have made and assist them in meeting those promises in a timely manner. We show that domain bias associated with email corpora has a large negative impact on the performance of commitment detection models. We adapt and use transfer learning methods to remove the domain bias from email corpora and show that by means of transfer learning we can reliably detect commitments even if there is a domain bias.

In the second part of the thesis, we devote two chapters to study how metadata and structure can be useful in two different exploratory search tasks. First, we consider detecting shifts in the meaning of words and study how metadata features such as time and social entities such as political party can be used to discover shifts in a short period of time. We use distributional semantics models to represent the

meaning of words and propose different approaches to measure the difference between the meaning of a term in different semantic spaces. We show that the detected shifts can be helpful in various tasks such as automatic text summarization and ideology detection. Second, we focus on finding similar questions in community question answering forums. Our aim is to use structure and metadata associated with questions beside their content to measure the semantic similarity between questions. We show that a neural model that effectively exploits the heterogeneous information associated with questions can improve the performance of content-based approaches in this task.