INTRODUCTION

INFORMATION EXTRACTION & KNOWLEDGE MANAGEMENT

Information is vast and broadcast is fast:
- The web with its sheer unmeasurable speed of broadcasting new data and its vast quantity of available knowledge is the first choice for information seekers. It is a user’s privilege to read or skim a webpage or bookmark it for later reference, but considering that the human memory can be deceptive, it also is a user’s obligation to keep information ordered and easily accessible for later reference.

Manual instrumentation methods exist:
- Concept maps, or mind maps provide the necessary methodology and have been implemented in a multitude of prolific, computerized toolkits, which go beyond simple bookmarking.
  → Manual compilation is not maintainable

Managing knowledge:
- Induce taxonomic or ontological knowledge, i.e. general relations between concepts, purely from text
- Identify named entities and relations between them and show them in a so-called network of named entities
  - Magpie or ESpotter highlight named entities in websites
- With STORYFINDER, we aim to support the user to quickly grasp the key concepts of a webpage, make it easily accessible for later usage, and put the new information into relation with previously visited web pages.
  a) highlighting within webpage
  b) separate graph-based view with entities and their relations
  → grasp the “bigger picture”

ABSTRACT

- We present STORYFINDER, a user application which consists of a browser plugin and a web server backend with the goal to highlight and manage the information contained in web pages by combining techniques from natural language processing and visual analytics. Named entities and keywords are extracted and stored per web page and used in a graph of entities and their relations. The scope can be set to a single web page, multiple web pages, or the complete set of analyzed web pages in a user’s history.

SCHEMA

Web browser plugin:
- Listens and reacts to user events
- Initiates analysis on the server
- Provides side pane view of interactive webpage

Server backend:
- Analyzes webpage
- Extracts metadata: Named entities, keywords

Interactive webpage:
- Provides graph view of articles
- Adheres to standards of Visual Analytics

Information is fully editable:
- Add or remove named entities
- Add or remove relations
- Label primary relations, i.e. generally valid
- Label secondary relations, i.e. valid for a particular sentence

TECHNOLOGY STACK

- Docker containers are used for NLP components and Database integration
  → ensures scalability
- RESTful communication between components
  → enables modular design and exchangeability
- Websockets technology is used for communication from server to client, i.e. the interactive webpage
  → enables asynchronous events