Jens Gäbeler

Vorhaben einer Masterarbeit im Online-Masterstudiengang Medieninformatik

Betreuung: Prof. Dr.-Ing. Nils Jensen & Prof. Dr. rer. nat. habil. Torsten Sander

Ostfalia
Hochschule für angewandte Wissenschaften
Fakultät Informatik

Entwicklung einer Steuerungskomponente zur Priorisierung von Aufträgen für verteilte Webcrawls

Motivation

Mithilfe des Open Web Index (OWI) möchte die Open Search Foundation eine europäische Alternative zu den kommerziellen Webindizes der Firmen Google, Microsoft, Baidu und Yandex schaffen. Dieser Index als öffentliche Infrastruktur soll von europäischen Organisationen genutzt werden können, um vielfältige Dienste erstellen und anbieten zu können.

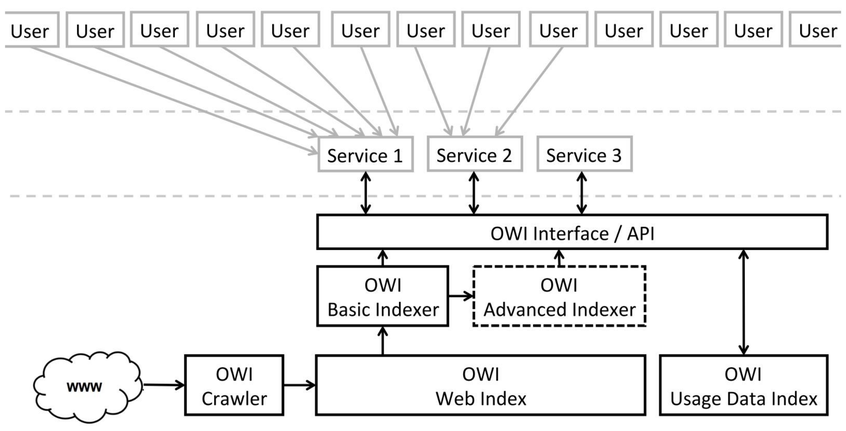

Abbildung 1: Modularer Aufbau der Dienste des OWI [1]

Die geplante modulare Bauweise des OWI (siehe Abbildung 1) ermöglicht die Konzentration auf einzelne Aspekte, wie zum Beispiel das in der Masterthesis zu bearbeitende Thema der Priorisierung und Verteilung von Crawlaufträgen.

Fragestellung

Vor der Erstellung einer Steuerungskomponente zur Priorisierung entstehen die folgenden wissenschaftlichen Fragen:

1. Welche Arten der Priorisierung von Crawler-Warteschlangen gibt es?
2. Gibt es Priorisierungstechniken explizit für verteilte Crawler?
3. Welche geotopologische Ansätze gibt es bereits?
4. Welche Vergleiche gibt es und wie erfolgt die Evaluation?

Abbildung 2: Forschungsfragen

Vorgehensweise

Die Masterthesis durchläuft geplant 5 Phasen (siehe Tabelle 1). Davon finden die ersten beiden Phasen parallel statt, die weiteren Phasen laufen konsekutiv.

Tabelle 1: Phasen der Masterthesis

Phase	Methode	Ziel
1. Akademische Analyse	SLR & Kriterienaufnahme	Katalog mögl. Priorisierungs- und Evaluationstechniken
2. Techn. Vorbereitung	Bereitstellung der techn. Umgebung	Erstellung einer Minimalversion
3. Entwurf	Umsetzung auf Basis von 1)	Funktionsentwurf & Prozessentwurf
4. Realisierung	Umsetzung auf Basis von 3)	Prototyp
5. Evaluation	Vergleich	Ranking

Zeitplanung

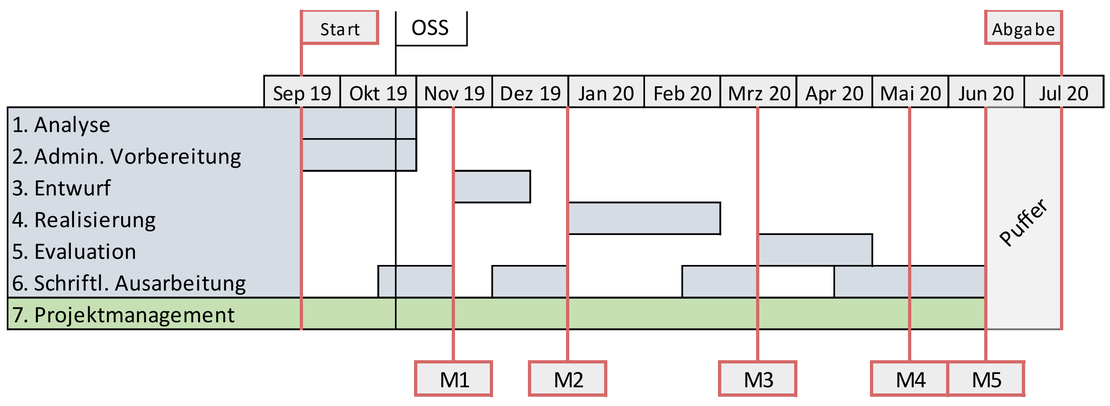

Abbildung 3: Zeitplanung der Masterthesis

Der Beginn der Masterarbeit ist der 19.09.2019. Das späteste Abgabedatum ist der 23.07.2020. Dies entspricht einer Bearbeitungszeit von 44 Wochen in Teilzeit. In Abbildung 3 sind die geplanten Arbeitspakete und deren Zeiten dargestellt.

Stand der Arbeit

Organisatorisch

Die Literaturrecherche erweist sich als umfangreicher als während der initialen Planung gedacht. Daher wurde der Zeitplan für die ersten beiden Phasen um zwei Wochen verlängert. Im Gegenzug wird der zeitliche Rahmen der Realisierung gekürzt. Abbildung 3 zeigt die aktuelle Version.

Wissenschaftliche Erkenntnisse (Phase 1)

Mit der Methode SLR (Systematic Literature Review) werden zu den Forschungsfragen passende wissenschaftliche Arbeiten gesucht, kategorisiert und ausgewertet. Zum aktuellen Stand wurden die folgenden relevanten Arbeiten identifiziert:

Tabelle 2: Arbeiten zur Beantwortung der Forschungsfragen (FF)

Autor	Titel	Jahr	FF
Cui, Y.; et al.	Unsupervised Domain Ranking in Large-Scale Web Crawls	2018	2,4
Baker, M. R.; Akcayol, M. A.	Priority Queue Based Estimation of Importance of Web Pages for Web Crawlers	2017	1
Zhang, L.; et al.	DGWC: Distributed and generic web crawler for online information extraction	2016	2,4
Tran, G.; et al.	A Random Walk Model for Optimization of Search Impact in Web Frontier Ranking	2015	1,4
Le Quoc, D.; et al.	UniCrawl: A Practical Geographically Distributed Web Crawler	2015	3,4
Gupta, S.; Bhatia, K. K.	CrawlPart: Creating Crawl Partitions in Parallel Crawlers	2013	2
Seyed, M. M.; et al.	A Brief History of Web Crawlers	2013	4
Yadav, D.; et al.	An Approach to design incremental parallel Webcrawler	2012	2
Ahmadi-Abkenari, F.; Selamat, A.	A clickstream-based web page significance ranking metric for Web crawlers	2011	4
Soon, L.-K.; Lee, S. H.	Reducing Redundant Web Crawling Using URL Signatures	2010	2

Des Weiteren wurden die folgenden Grundlagenwerke zur umgebenden Recherche des Themas ermittelt:

Tabelle 3: Literatur zur Beantwortung von Fragen zum Thema Web Crawling

Autor	Titel	Jahr
Cambazoglu, B. B.; Baeza-Yates, R.	Scalability Challenges in Web Search Engines	2015
Feitelson, D. G.	Workload Modeling for Computer Systems Performance Evaluation	2015
Zheng, S.	Effective methods for web crawling and web information extraction	2011
Liu, B.; Menczer, F.	Web Crawling	2011
Olston, C.; Najork, M.	Web Crawling	2010
Weise, T.	Global optimization algorithms-theory and application	2009
Manning, C. D., et al.	Introduction to Information Retrieval	2008
Castillo, C.	Effective web crawling	2004

Technische Bereitstellung (Phase 2)

Die OWI Architektur wurde analysiert und folgender Einhängepunkt der zu entwickelnden Steuerungskomponente identifiziert:

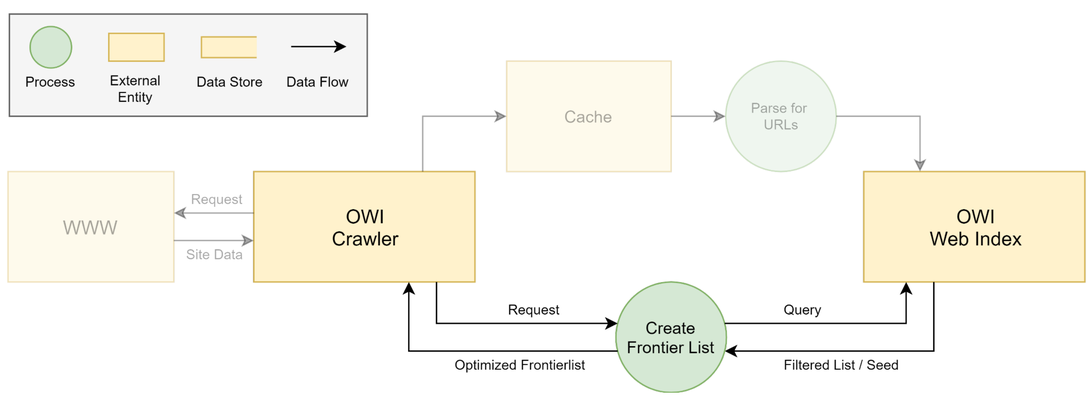

Abbildung 4: Verortung der Steuerungskomponente (Create Frontier List)

Für die Realisierung wird *Python* und das Microframework *Flask* verwendet. *Python* eignet sich durch die umfangreichen Funktionen seiner Standardbibliothek und durch eine Vielzahl an Bibliotheken zur Berechnung (z.B. *Numpy*, *SciPy*), Datenhaltung (z.B. *Pandas*) und bei Bedarf Künstlicher Intelligenz (z.B. *SciKit*, *Keras*, *TensorFlow*). Das Microframework *Flask* eignet sich sehr für die Erstellung einer Anwendung, die über eine *REST API* zur Verfügung gestellt wird.

Quelle [1] Huss, N.; Lewandowski, D.; Sander-Beuermann Wolfgang; Ude, A.: Proposal for an Open Web Index, Hannover 2017. https://openwebindex.eu/wp-content/uploads/2019/01/Open_Web_Index_proposal.pdf, abgerufen am 09. August 2019