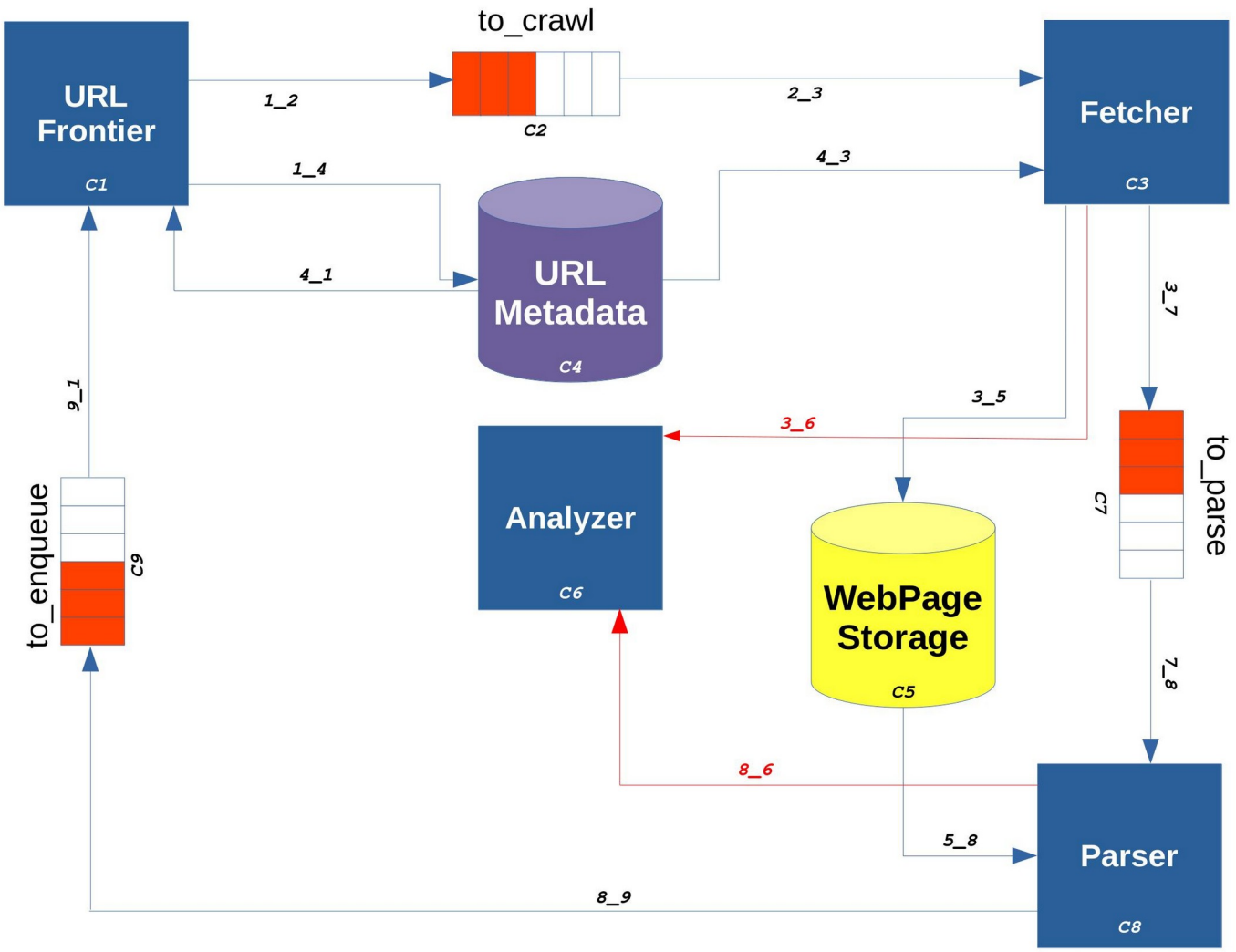
معماری خزش‌گر وب

علی رشیدی

نمای کلی


URL Frontier
C1
to_crawl
1_2
C2
2_3
Fetcher
C3
1_4
4_3
URL Metadata
C4
4_1
3_7
9_1
3_5
3_6
Analyzer
C6
WebPage Storage
C5
to_parse
C7
to_enqueue
C9
7_8
8_6
5_8
Parser
C8
8_9

کامپوننت‌ها

- C1: URL Frontier
  - URLهایی که Parser در صف ورودی گذاشته را برمی‌دارد
  - در مورد URLهای ورودی تصمیم‌گیری می‌کند که پردازش شوند یا خیر، و ضمن اختصاص ID و ذخیره در C4 در صف خروجی قرار می‌دهد
- C2: to_crawl queue
  - آدرس‌هایی که باید خزش شوند در این صف قرار می‌گیرند
- C3: Fetcher
  - از صف ID را برمی‌دارد و دانلود می‌کند و در WebPage Storage می‌گذارد
  - آی.دی را در صف خروجی می‌گذارد تا Parse شود. در ضمن اطلاعاتی مانند حجم را به Analyzer می‌دهد

کامپوننت‌ها: ادامه

- C4: URL Metadata storage
  - ذخیره اطلاعات آدرس‌ها. در حال حاضر به صورت جفتی به شکل ID:URL
  - از Redis استفاده می‌شود
- C5: WebPage Storage
  - محلی بر روی دیسک که صفحات ذخیره می‌شوند. چون صرفا فایل html دانلود می‌شود، فایل‌ها به عنوان <id>.html ذخیره می‌شوند
- C6: Analyzer
  - این سرویس اطلاعاتی که برای تحلیل نتایج مورد نیاز است را دریافت می‌کند و گزارش تولید می‌کند

کامپوننت‌ها: ادامه

- C7: to_parse queue
  - آدرس‌هایی که دانلود شده‌اند و باید parse شوند در این صف قرار می‌گیرند
- C8: Parser
  - شناسه صفحات را از صف ورودی برمی‌دارد، لینک‌ها را استخراج می‌کند
  - اطلاعات لازم را به analyzer می‌فرستد و لینک‌های مربوط به هاست خودمان را در صف می‌گذارد
- C9: to_enqueue queue
  - آدرس‌هایی که توسط parser کشف شده‌اند و باید خزش شوند
  - C1 در مورد آن‌ها تصمیم می‌گیرد

اتصالات

- 1_2
  - Publish URL ID to “to_crawl” queue
- 1_4
  - Update URL Metadata (in our case, just insert a new ID:URL pair)
- 2_3
  - Direct the message to C3
- 3_5
  - Save downloaded webpage with the name <id>.html
- 3_6
  - Send webpage size and compression flag to the analyzer

اتصالات (ادامه)

- 3_7
  - Publish a message with the ID of the fetched URL to be parsed
- 4_1
  - C1 can read from URL Metadata storage to check for uniqueness
- 4_3
  - Fetcher reads the URL based on the ID read from the queue
- 5_8
  - Parser reads the stored webpage to parse
- 7_8
  - The queue directs the URL ID to the parser

اتصالات (ادامه)

- 8_6
  - Parser sends data about the outlinks to the analyzer
- 8_9
  - Parser puts the discovered URLs in the “to_enqueue” queue
- 9_1
  - The URLs are directed to C1 to check whether they have to be crawled or not