The GoOLAP Fact Retrieval Framework

Alexander Löser, Sebastian Arnold, and Tillmann Fiehn

FG DIMA
Technische Universität Berlin
Berlin, Germany
firstname.lastname@tu-berlin.de

Summary. We discuss the novel problem of supporting analytical business intelligence queries over web-based textual content, e.g., BI-style reports based on 100,000s of documents from an ad-hoc web search result. Neither conventional search engines nor conventional Business Intelligence and ETL tools address this problem, which lies at the intersection of their capabilities. Three recent developments have the potential to become key components of such an ad-hoc analysis platform: significant improvements in cloud computing query languages, advances in self-supervised keyword generation techniques and powerful fact extraction frameworks. We will give an informative and practical look at the underlying research challenges in supporting ”Web-Scale Business Analytics” applications that we met when building GoOLAP, a system that already enjoys a broad user base and over 6 million objects and facts.

4.1 Introduction

Which companies collaborate with Boeing? Are these organizations also collaborating with Airbus or Fokker? Do employees of these companies have a criminal record? Who published the information?

Each day new pages emerge in the Web that may contain a textual representation of facts for answering these questions. Strategic decision makers may frequently research the Web for questions like these. Often answers to these queries might not be published in textual or structured form by Web ‘information aggregators’ like Wikipedia.com, Freebase.com, Trueknowledge.com or Yago [1]. Rather, this rare factual information is hidden in unstructured Web text on a few Web pages of news agencies or blogs. Unfortunately, collecting factual answers from these pages with a general Web search engine is still a dreaded process for a user.

One option to populate a fact base is to crawl a large document collection. For instance Google Squared [2] populates its data base with facts from the large corpus of the general Web. The system extracts these facts from tables, lists and from text with open information techniques. However, in [3] we observed that only a small fraction of a large archive de facto contains factual
information. Hence, strategies that might execute a full scan over the entire archive can drastically waste processing and storage resources (see also [4]). Another option is discovering facts in retrieved pages from ad-hoc keyword search. Unfortunately, this is still a tedious task, since Web search engines do not return aggregated factual information. The heuristic of a search engine user is: type in keywords as queries, ‘extract’ relevant facts from the top-k documents, filter out relevant facts and compile facts into a structured fact table. Therefore the user typically repeats this process multiple times in order to complement missing attribute values and to enhance the chance to discover unseen facts.

**Significance of our approach.** We present GoOLAP, a system that aims to automate the fact retrieval process from Web search engines. GoOLAP has three significant aspects: (1) It provides powerful operators for analytical Web research on textual information, such as augmenting facts for an object, tracing back the textual origin of a fact or comparing factual information for a list of objects. (2) As a natural alternative to crawling a large proportion of the Web, GoOLAP interprets user interactions as input to identify missing facts. These user interactions trigger a fact retrieval process with the goal to populate the GoOLAP fact base from Web-scale indices of existing search engines in an ad-hoc fashion. This process is powered by the FactCrawl engine that leverages sophisticated keyword generation techniques [3, 5] and page classification techniques [6] to retrieve only pages that likely contain missing and rare factual information [7]. (3) GoOLAP combines these approaches to drastically avoid crawling, indexing and extracting potentially billions of irrelevant pages. GoOLAP’s human-machine machine generated fact base nearly reaches 6 million facts and objects; a dimension that is still an order of magnitude smaller than community-generated fact bases, such as DBpedia [8] or CrunchBase [9]. However, our prototype extracts facts in a hybrid approach: It leverages the crowd, the power of GoOLAP users, to discover new objects in queries, extracts and augments with the help of parallel running machines new factual information and again utilizes the power of GoOLAP users to verify and re-trigger this process, for instance in the case of incomplete information.

### 4.2 Related Work

We discuss relevant related work in the areas of focused fact retrieval from full text indexes, Open Information Extraction and fact search engines.

**Keyword query generation with QXtract.** The authors of QXtract [10] pioneered work on automatically generating keyword phrases for fact retrieval from full-text indices. The system executes the following stages in a one time learning process: sample seed documents, observe/score phrases in these