Erschienen in: Horbach, Matthias (Hrsg.): Informatik 2013. Informatik angepasst an Mensch, Organisation und Umwelt. - Bonn-Buschdorf: Köllen, 2013. S. 140-142
(Lecture Notes in Informatics 220)

KoGra-DB: Using MapReduce for Language Corpora

Dr. Roman Schneider

Institut für Deutsche Sprache (IDS)
R5 6-13
D-68161 Mannheim
schneider@ids-mannheim.de

Abstract: Linguistic query systems are special purpose IR applications. We present a novel state-of-the-art approach for the efficient exploitation of very large linguistic corpora, combining the advantages of relational database management systems (RDBMS) with the functional MapReduce programming model. Our implementation uses the German DEREKO reference corpus with multi-layer linguistic annotations and several types of text-specific metadata, but the proposed strategy is language-independent and adaptable to large-scale multilingual corpora.

1 Introduction

In recent years, the quantitative examination of natural language (NL) phenomena has become increasingly popular. Both fundamental research on the basic principles of human language, as well as the development of speech and language technology, rely on the empirical verification of assumptions and rules. Besides written (and sometimes spoken) language samples, NL corpora usually contain vast collections of morphosyntactic, phonetic, semantic etc. annotations, plus text- or corpus-specific metadata. The downside of this trend is obvious: Even with specialized applications, our ability to store linguistic data is often bigger than the ability to process all this data. As we go beyond corpus sizes of some terabytes, and at the same time increase the number of annotation systems and search keys, query costs rise disproportionately. This is due to the fact that unlike traditional IR systems, corpus retrieval systems not only have to deal with the "horizontal" representation of textual data, but with heterogeneous metadata on all levels of linguistic description. Given this context, we present a novel approach to scale up to billion-word corpora, using the example of the multi-layer annotated German Reference Corpus DEREKO, that constitutes the largest linguistically motivated collection of contemporary German. It contains fictional, scientific, and newspaper texts and is annotated morphosyntactically with three competing systems (Connexor, Xerox, TreeTagger). All this data is stored within the Corpus Grammar Database (KoGra-DB), using an object-relational DBMS and interfaces for further statistical processing in R.

2 Implementing and Querying KoGra-DB

In order to overcome typical DBMS bottlenecks, we propose a strategy that allows the distribution of processor-intensive computation over several processor cores and facilitates the partition of complex linguistic queries at runtime into independent single

queries that can be executed in parallel. It is based on two presuppositions: (i) Relational DBMS can be used effectively to maintain parsed texts and linguistic metadata. We intensively evaluated different types of tables (heap tables, partitioned tables, index organized tables) as well as different index types (B-tree, bitmap, concatenated, functional) for the storing and retrieval of linguistic data. (ii) The MapReduce programming model supports distributed programming and tackles large-data problems. Though MapReduce is already in use in a wide range of data-intensive applications, its principle of "divide and conquer" has not been employed for corpus retrieval yet.

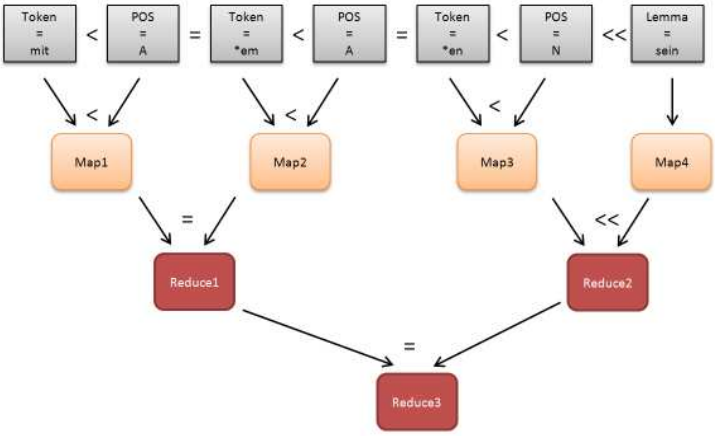

Figure 1: MapReduce processes for a complex grammatical query

Figure 1 illustrates the MapReduce processes for a complex grammatical query, using seven dictinct search keys on different metadata types. The example can be read as: *Find all sentences containing the token "mit" immediately followed by an adjective ending on "em", immediately followed by another adjective ending on "en", immediately followed by a noun, and followed by a word with the lemma (base form) "sein" at any distance*. It finds grammatical phenomena like "*…die mit großartigem persönlichen Einsatz unterwegs gewesen sind*" that are subject for further linguistic analysis.

Within a "map" step, the original query is partitioned into separate key-value pairs. Keys represent linguistic units (position, token, lemma, part-of-speech, etc.), values may be the actual content. The queries can be processed in parallel and pass their results to temporary tables. The subsequent "reduce" processes filter out inappropriate results. Usually, this cannot be executed in parallel, because each reduction produces the basis for the next step. But our framework overcomes this restriction by dividing the process tree into multiple sub-trees. The reduce processes for each sub-tree are scheduled simultaneously, and aggregate their results after they are finished.

In order to prove the feasibility of our approach, we implemented our corpus storage and retrieval framework on a commodity low-end server (quad-core microprocessor with

2.67 GHz clock rate, 16GB RAM). For the reliable measurement of query execution times, and especially to avoid caching effects, we always used a cold-started 64-bit database engine. KoGra-DB stores search results within the database schema, making it easy to reuse them for further processing. Advanced statistical computation is available via a web-based interface to the R software environment (see figure 2).

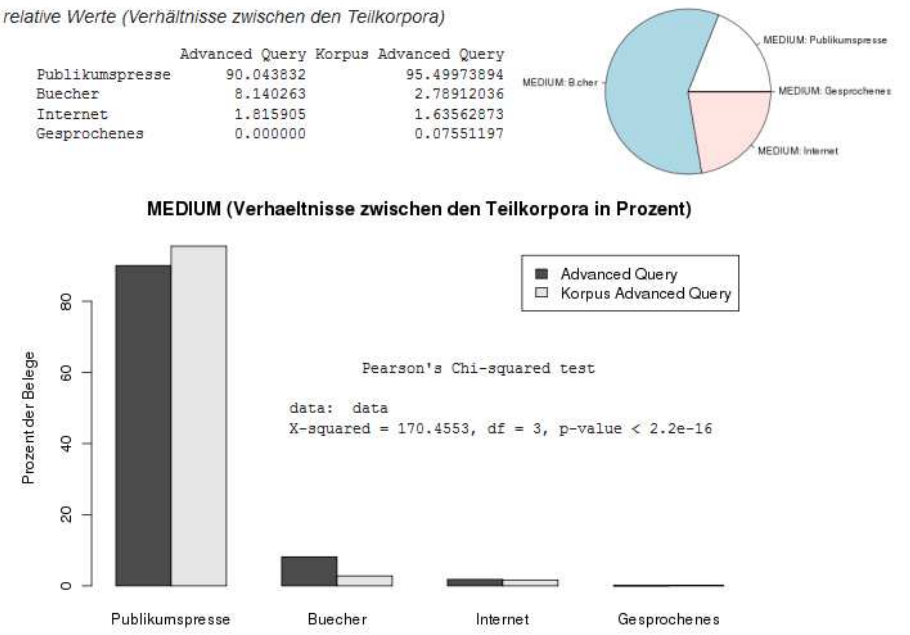

Figure 2: Statistical processing of search results with R via a web-based interface

3 Conclusion and Outlook

We presented a novel computational approach to work with empirical, annotated natural language corpora for the explanation of linguistic phenomena. The results of our study demonstrate that the joining of relational DBMS technology with a functional/parallel computing framework combines the best of both worlds for linguistically motivated large-scale corpus retrieval. On our reference server, it clearly outperforms other existing (e.g., purely relational) approaches. For the future, we plan some scheduling refinements, as well as support for additional levels of linguistic description.

References

[Sc12] Schneider, R.: Evaluating DBMS-Based Access Strategies to Very Large Multi-Layer Annotated Corpora. In: Proc. LREC-2012 Workshop "Challenges in the management of large corpora", Istanbul. *http://www.lrec-conf.org/proceedings/lrec2012/index.html*