Embedded Distributed Text Mining and Semantic Web Technology

Daniel Sonntag

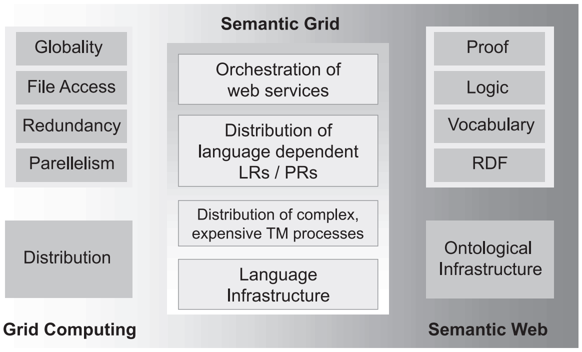
German Research Center for Artificial Intelligence
66123 Saarbrücken, Germany
sonntag@dfki.de

Abstract. We first position Text Mining (TM) components and challenges in a Grid-based distributed TM architecture. On the basis of this infrastructure we declare an embedded TM workflow.

1 Common View

No one would contradict the statement that the former small market of information retrieval systems will shift to a huge commercial natural language text processing (NLP) market. Decision support and security are important fields with similar requirements. Grid technology [1, 2] is intended to allow for seamless access and use of distributed computing or processing resources (PRs), and information or language resources (LRs). In distributed TM efficiency of subtasks will be guaranteed by specialised distributed classification methods. Geographically dispersed systems and heterogeneous data meet the requirements of advanced NLP and the nature of text as data in different languages. The second goal is to draw direct analogies between the Grid, distributed TM, and the Semantic Web for NLP applications (cf. [3–7] for example).

A common view on the Grid, distributed TM as Semantic Grid application, and the Semantic Web in the context of NLP helps to identify the influential forces. Grid Computing can be related to the Semantic Web via Semantic Grid structures based on a language infrastructure.


Semantic Grid
Globality
File Access
Redundancy
Parellelism
Distribution
Grid Computing
Orchestration of web services
Distribution of language dependent LRs / PRs
Distribution of complex, expensive TM processes
Language Infrastructure
Proof
Logic
Vocabulary
RDF
Ontological Infrastructure
Semantic Web

2 Embedded Text Mining Workflow

What keeps all embedded TM applications together is the need for sophisticated TM algorithms, and the need for qualitative data input. We hypothesise that the key to successful embedded TM application depends fore and foremost on the quality of the textual input, which otherwise would result to what is called the *Semantic Gap*. Unfortunately, unstructured text data on the Web or in large repositories are not suitable for automatic text mining, since the semantics of individual data items is not clear. Hence the data are not suitable for more sophisticated embedded text mining affairs. We give an account of ways to overcome the poor quality of textual data—which results in a research agenda for text annotations on the basis of database technology, ontologies, and statistical NLP [8–11]. Correct data entries, no duplicates, and referential integrity are examples of quality assets in relational databases, and quality factors like correct process models examples in data warehouse management, which introduced data cleaning as integral part of data maintenance processes. Ontology-based Semantic Web annotations, such as linguistic named entity classes, build the groundwork for representational data quality, which in turn helps to implement the vision of embedded distributed TM applications. In this connection texts are annotated by Semantic Web data structures, building up the ontological infrastructure. Finally, *Ontology Learning* provides the instrument for adapting to different application domains.

References

1. Foster, I., Kesselman, C.: The GRID. Morgan Kaufmann Publishers, Inc., San Francisco (1999)
2. Goble, C., Roure, D.D.: The Semantic Grid: Myth Busting and Bridge Building. Proceedings of ECAI (2004)
3. Zhang, D., Lee, W.S.: Question classification using support vector machines. In: Proceedings of the 26th Annual International ACM SIGIR, ACM Press (2003)
4. Sonntag, D.: Distributed NLP and Machine Learning for Question Answering Grid. Proceedings of the workshop on Semantic Intelligent Middleware for the Web and the Grid at ECAI (2004)
5. Abney, S., Collins, M., Singhal, A.: Answer Extraction. In: Applied Natural Language Processing (ANLP). (2000)
6. Cancedda, N., Gaussier, E., Goutte, C., Renders, J.M.: Word-Sequence Kernels. Journal of Machine Learning Research (February 2003) 1059–1082
7. Buitelaar, P., Declerck, T., Calzolari, N., Lenci, A.: Towards a Language Infrastructure for the SW. In: Proceedings of the ISWC workshop on HLT. (2003)
8. Sonntag, D.: Assessing the Quality of Natural Language Text Data. In: Proceedings of GI Jahrestagung (1) 2004: 259-263
9. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB Journal: Very Large Data Bases **10**(4) (2001) 334–350
10. Charniak, E.: Statistical Parsing with a Context-Free Grammar and Word Statistics. In: Proceedings of the 14th AAAI Conference. (1997)
11. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The M.I.T. Press, Cambridge (Mass.) and London (1999)