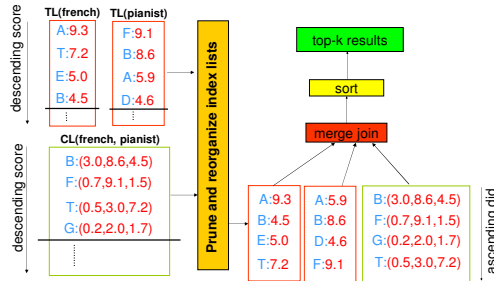


INTRODUCTION

An increasing amount of today's data is available in semistructured form: the data provides more structure than plain text documents, but does not have the regular structure of a traditional, relational database. Important examples for such semistructured information include texts that have been annotated with semantic markup to denote grammatical structure or named entities, social networks with complex relationships of people and their data, often including items like images or bookmarks that have been collaboratively tagged, and graph-structured knowledge networks such as ontologies. Our research focuses on managing, querying and analyzing large collections of such data. Our work integrates aspects of data management and information retrieval, requiring both effective retrieval models to find relevant results and efficient data structures and algorithms to quickly retrieve these results.

Efficient Text Retrieval

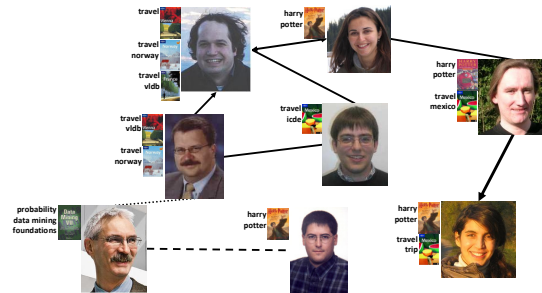
- Term pair lists for real-time query processing over massive text collections with dependable response time [3] (on average <50ms for ~5TB Web documents, with good quality)



- Approximate top-k queries under budget constraints [8]
 - Adapt scheduling of existing top-k algorithms (NRA, IO-Top-k)
 - Two phases:
 - Prefer lists with **high-scoring entries** to find good candidates
 - Prefer lists with **sharp drop in scores** to eliminate bad candidates
 - Random accesses only for huge budgets
 - Result quality **close to optimal** with reasonable budget (70%-90%)
- Upcoming project: **Highly Interactive Information Retrieval**, with Norbert Fuhr, U Duisburg-Essen (funded by DFG, starting 2011)

SENSE – Search in Social Networks

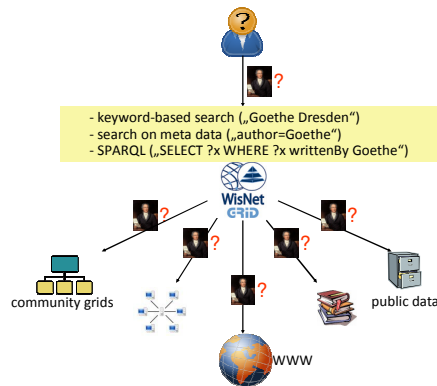
- Tag-based search in social networks [7]
 - Similarity measures** for friends based on distance and content
 - Context-aware score** for weighted combination of results from friends
 - Efficient top-k algorithm for **dynamic aggregation** in the network, including self-tuning tag expansion
 - Large-scale **experimental evaluation** of **quality** and **efficiency** with three crawls from real networks (LibraryThing, Flickr, del.icio.us)



- Efficient precomputation of shortest distances in very huge graphs
- Efficient maintenance of precomputed nearest neighbors
 - under insertions and deletions of **edges**
 - under updates of **content**
- Efficient maintenance of tag similarities under content updates

Knowledge Extraction and Knowledge-Based Search

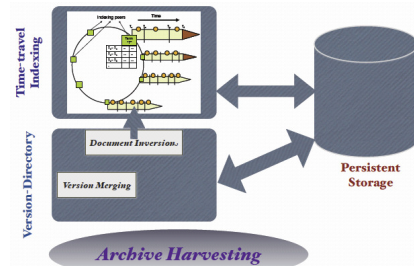
- WisNetGrid – A Knowledge Management Infrastructure for D-Grid (with TU Dresden, KIT, Göttingen U, and others; BMBF, 2009-2012)
 - Distributed **harvesting of information** from Grid sources
 - Federated search** over heterogeneous knowledge



- Efficient and effective ranking for **SPARQL+FullText** [5]
- Explanation** of search results through extraction witnesses [4]
- Diversity-aware **summarization** of entities

Large-Scale and Long-Term Web Archives

- EverLast: P2P-based Web archiving [2]
 - Human-assisted harvesting** through Browser plugins & proxies to cope with increasingly high change rates of Web sites
 - Two-level distributed index, partitioned by **term** and **time**
 - Reliable storage through **block-based replication** of index lists



- Efficient query processing in centralized & distributed archives through **index list partitioning**
 - Time-based **vertical partitions**
 - Cost-based **horizontal partitions**
 - Partition selection for **approximate query processing** [1]
- Document version identification** through temporal shingling [6]

REFERENCES

- A. Anand, S. Bedathur, K. Berberich, R. Schenkel: **Efficient Temporal Keyword Queries over Versioned Texts**. Int. Conf. on Information and Knowledge Management (CIKM), Toronto, Canada, October 2010
- A. Anand, S. Bedathur, K. Berberich, R. Schenkel, C. Tryfonopoulos: **EverLast: A Distributed Architecture for Preserving the Web**. Joint Conf. on Digital Libraries (JCDL), Austin, USA, June 2009
- A. Broschart, R. Schenkel: **MMCI at the TREC 2010 Web Track**. Text Retrieval Conference, Gaithersburg, USA, November 2010
- S. Elbassuoni, K. Hose, S. Metzger, R. Schenkel: **ROXXI: Reviving witness dOCuments to eXplore eXtracted Information** (demo). Int. Conf. on Very Large Data Bases (VLDB), Singapore, September 2010
- S. Elbassuoni, M. Ramanath, R. Schenkel, M. Sydow, G. Weikum: **Language-model-based Ranking for Queries on RDF-Graphs**. Int. Conf. on Information and Knowledge Management (CIKM), Hongkong, China, October 2009
- R. Schenkel: **Temporal Shingling for Version Identification in Web Archives**. European Conf. on Information Retrieval (ECIR), Milton Keynes, UK, March 2010
- R. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. Parreira, G. Weikum: **Efficient Top-k Querying over Social-Tagging Networks**. ACM SIGIR, Singapore, July 2008
- M. Shmueli-Scheuer, C. Li, Y. Mass, H. Roitman, R. Schenkel, G. Weikum: **Best Effort Top-K Query Processing Under Budgetary Constraints**. Int. Conf. on Data Engineering (ICDE), Shanghai, China, March 2009