

# Challenges in Document Mining

Edited by

Hamish Cunningham<sup>1</sup>, Norbert Fuhr<sup>2</sup>, and Benno Stein<sup>3</sup>

<sup>1</sup> University of Sheffield, UK, [gate.ac.uk/hamish](mailto:gate.ac.uk/hamish)

<sup>2</sup> Universität Duisburg-Essen, Germany, [norbert.fuhr@uni-due.de](mailto:norbert.fuhr@uni-due.de)

<sup>3</sup> Bauhaus-Universität Weimar, Germany, [benno.stein@uni-weimar.de](mailto:benno.stein@uni-weimar.de)

---

## Abstract

This report documents the programme and outcomes of the Dagstuhl Seminar 11171 *Challenges in Document Mining*. Our starting point was the observation that document mining techniques are often applied in an isolated manner, with the consequence that their potential is still to be fully realised. The goal of the seminar was to analyze this untapped potential. To this end researchers from the main areas of document mining were invited to present their views, to synthesise an understanding of where and how the latest disciplinary achievements can be combined, and to develop a more integrative view on the state of the art and the prospects for future progress.

**Seminar** 25.–29. May, 2011 – [www.dagstuhl.de/11171](http://www.dagstuhl.de/11171)

**1998 ACM Subject Classification** H.1.2 [Information Systems: User/Machine Systems]; H.3.1 [Information Systems: Content Analysis and Indexing]; H.3.3 [Information Systems: Information Search and Retrieval]; I.2.7 [Computing Methodologies: Learning]; I.2.7 [Computing Methodologies: Natural Language Processing]; I.5.3 [Computing Methodologies: Clustering]

**Keywords and phrases** Cluster analysis, HCI, Retrieval models, Social mining and search, Semi-supervised learning

**Digital Object Identifier** 10.4230/DagRep.1.4.65


**Edited in cooperation with** Melikka Khosh Niat

## 1 Executive Summary

*Hamish Cunningham*

*Norbert Fuhr*

*Benno Stein*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Hamish Cunningham, Norbert Fuhr, Benno Stein

## About Document Mining

Document mining is the process of deriving high-quality information from large collections of documents like news feeds, databases, or the Web. Document mining tasks include cluster analysis, classification, generation of taxonomies, information extraction, trend identification, sentiment analysis, and the like. Although some of these tasks have a long research history, it is clear that the potential of document mining is still to be fully realised.

Part of the problem is that relevant document mining techniques are often applied in an isolated manner, addressing – from a user perspective – only a part of a task. For example, an intelligent cluster analysis requires adequate document models (from information retrieval) that are combined with sensible merging algorithms (from unsupervised learning), complemented by an intuitive labelling (from information extraction, natural language processing).



Except where otherwise noted, content of this report is licensed under a Creative Commons BY-NC-ND 3.0 Unported license  
Challenges in Document Mining, *Dagstuhl Reports*, Vol. 1, Issue 4, pp. 65–99  
Editors: Hamish Cunningham, Oren Etzioni, Norbert Fuhr, and Benno Stein



DAGSTUHL REPORTS  
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The deficit that we observe may also be understood as a lack of application and user orientation in research. For example, given a result set clustering task, users expect:

1. as many clusters as they identify topics in the result set,
2. that the documents within each cluster are semantically similar to each other, and
3. that each cluster is labeled intuitively.

In order to achieve such a satisfying solution, the state-of-art of concepts and algorithms from information retrieval, unsupervised learning, information extraction, and natural language processing have to be combined in a user-focussed manner.

### Goals of the Seminar

The general idea was to take an overview the state of the art in document mining research and to define a research agenda for further work. Since document mining tasks are not tackled by a single technology, we wanted to bring a sample of the leading teams together and look at the area from a multidisciplinary point of view. In particular, the seminar should focus on the following questions:

- What are the relevant document mining tasks? The expectations and the potential for document mining changed significantly over time. Influential in this connection is the discovery of the enormous contributions of users to the Web, among others in the form of blogs, comments and reviews, as highly valuable information source.
- What are the options and limitations of cluster analysis? A major deal of cluster analysis research has been spent to merging principles and algorithms; today, and especially in document mining, the focus is on tailored document models, user integration, topic identification and cluster labelling, on the combination with retrieval technology (e.g. as result set clustering). Especially non-topical classification tasks attracted interest in this connection, such as genre classification, sentiment analysis, or authorship grouping. Moreover, theoretical foundations of cluster analysis performance in document mining as well as commonly accepted optimality measures are open questions.
- What are the document mining challenges from a machine learning perspective? A crucial constraint is the lack of sufficient amounts of labelled data. This situation will become even more unbalanced in the future, and current research—to mention domain transfer learning and transductive learning—aim at the development of technology to exploit the huge amount of unlabelled data to improve supervised classification.
- How will NLP and IE affect the development of the field? The use of NLP and IE in document mining is a success factor of increasing importance for document mining. NLP contributes technology for document modelling, style quantification, document segmentation, topic identification, and various information extraction and semantic annotation tasks. In this regard authorship and writing style modelling is still coming of age; this area forms the heart for high-level document mining tasks such as plagiarism analysis, authorship attribution, and information quality assessment.
- Are new interaction paradigms on the rise? Interface design and visualization are very important for effective user access to the output of the document mining process. Moreover, interactive document mining approaches like e.g. scatter-gather clustering pose new challenges for both the interface and the backend.
- How to evaluate and compare the different research efforts? Evaluation is essential for developing any kind of data mining method. So far, mainly system-oriented evaluation approaches have been used, where the data mining output is compared to some “gold

## Program of the Dagstuhl Seminar 11171, 25.-29. May, 2011

	Tuesday	Wednesday	Thursday	Friday
9:00 - 10:00	Welcome, Talks	Talks	Talks	Working group presentation
10:30 - 11:00	Coffee break			
11:00 - 12:30	Talks	Working group topics	Talks	Wrapup
12:30 - 14:00	Lunch			
14:00 - 15:30	Talks	Working groups	Social event: Excursion to Trier	
15:30 - 16:00	Coffee break			
16:00 - 18:00	Talks	Working groups		
18:00 - 19:00	Dinner			
19:00 - 22:00	Demos	Working groups	Open work	

standard”. There is a lack of user-oriented evaluations (e.g. observing users browsing a cluster hierarchy), that also take into account the tasks the users want to perform—e.g. using Borlund’s concept of simulated work tasks.

### Seminar Organization

To stimulate debate and cross-pollination we scheduled a mixture of of talks, working groups and demos. Following Dagstuhl tradition, the talks were characterized by interactive discussions and provided a platform for presenting and discussing new ideas. The working group topics were arrived at by a brainstorming session. Due to Easter Monday we had only four days for our seminar and shifted parts of the program to the evening. The table shows the schedule of our seminar at a glance.

### Selected Results

This week showed that there is a number of recurring themes that are addressed by different researchers:

1. The processing hierarchy: Classic methods in document mining deal with document clustering and classification, thus regarding documents as a whole (or an “atomic” unit). Recently, researchers have become interested in deeper analyses of texts, such as sentiment analysis and the extraction of entities and relations.
2. Unsupervised vs. supervised methods: The former can be applied easily, but often lead only to modest results. Supervised methods produce more valuable results, but require large training sets for generating high-quality output. However the two approaches are not real alternatives: there are various attempts for their combination, like e.g. using prior knowledge for improving clustering, or using unclassified data with clustering for classification.
3. Whereas most supervised methods are strongly domain-dependent, there are now attempts for developing more domain-independent or cross-domain methods that can be applied more universally.
4. User feedback and user interaction has become an important component of document mining: There are many approaches aiming at better visualizations of the mining results. More recently, visual analytics methods have become popular, which aim at supporting

the user during the mining process itself, thus incorporating the user's knowledge in the actual analysis (and not only during the training stage or for result presentation).

The week also showed (or confirmed) deficiencies in document mining and pointed to future research directions:

1. The multiplicity of retrieval questions, mining solutions, test corpora and evaluation measures (to mention only a few determinants) emerges naturally when satisfying individual information needs in our information-flooded society. However, this multiplicity hinders the comparison of solutions and hence the consequent improvement of the most promising technology. What is the ideal research infrastructure to exploit synergies?
2. That we suffer from an information overload is a commonplace. We ask: A single person has no chance to process a substantial part of the information at our disposal—but a machine can. How can we benefit from this fact?
3. Current retrieval and mining technology is text-centered. The question is if and how the respective machinery can be applied to complex objects and artificially generated data: Which elements of the state-of-the-art retrieval technology is of generic type, and, can we develop a retrieval theory for complex structures?

First answers and arguments related to these questions can be found in the working groups section of this report.

## 2 Table of Contents

### Executive Summary

<i>Hamish Cunningham, Norbert Fuhr, Benno Stein</i> . . . . .	65
---	----

### Overview of Talks


Knowledge Discovery in the Web: Potential, Automation and Limits <i>Stefan M. Rüger</i> . . . . .	71
Full Lifecycle Information Extraction and Multiparadigm Indexing with GATE <i>Hamish Cunningham</i> . . . . .	71
Unsupervised and Semi-Supervised Approaches to Cross-Domain Sentiment Classification <i>John A. Carroll</i> . . . . .	73
Challenges in Mining Social Media: Sparsity and Quality <i>Thomas Gottron</i> . . . . .	73
Simulation Data Mining in Artificially Generated Data <i>Steven Burrows</i> . . . . .	74
Unsupervised entailment detection for IE query expansion <i>Ted Briscoe</i> . . . . .	74
Piggyback: Using Search Engines for Robust Cross-Domain Named Entity Recognition <i>Hinrich Schütze, Massimiliano Ciaramita</i> . . . . .	74
Feedback Methods in Large Scale Visual Document Analysis <i>Michael Granitzer</i> . . . . .	75
The Optimum Clustering Framework <i>Norbert Fuhr</i> . . . . .	75
TIRA – Research Assistant for Empirical Evaluations <i>Tim Gollub</i> . . . . .	76
LFRP-Search: Multi-Layer Faceted Search with Ranking and Parallel Coordinates – Interactive Retrieval for Complex Documents and Individual Information Needs <i>Andreas Henrich</i> . . . . .	76
Beyond Search: Interactive Web Analytics <i>Alexander Loeser</i> . . . . .	77
The Retrievability of Documents <i>Leif Azzopardi</i> . . . . .	77
Looking at Document Collections from a Bird’s Eye View: Exploratory Search as Based on the Context Volatility of Terms <i>Gerhard Heyer</i> . . . . .	78
Facet-Streams and Search-Tokens – Tangible User Interfaces for Information Seeking <i>Harald Reiterer</i> . . . . .	78
What do you mean? – Determining the Intent of Keyword Queries on Structured Data <i>Wolf Siberski</i> . . . . .	79

Exploiting Unlabeled Data in Information Retrieval Tasks <i>Benno Stein</i> . . . . .	79
Cross-lingual adaptation using structural correspondence learning <i>Peter Prettenhofer</i> . . . . .	80
Search by Strategy <i>Arjen P. de Vries</i> . . . . .	80
Challenges in Patent Retrieval and Mining <i>Dennis Hoppe</i> . . . . .	81
Cancelled Talks . . . . .	81
<b>The Working Groups</b>	
Apparatus, Reproducibility, and Evaluation . . . . .	82
What can computers learn from reading one million books? . . . . .	87
Retrieval of Complex Structures . . . . .	90
Sentiment and Opinion . . . . .	93
<b>A Theme: Towards More Open Search In Europe</b>	
Discussion . . . . .	96
<b>Recap of the Proposal</b> . . . . .	97
<b>Acknowledgements</b> . . . . .	98
<b>Participants</b> . . . . .	99

### 3 Overview of Talks

#### 3.1 Knowledge Discovery in the Web: Potential, Automation and Limits

*Stefan M. Rüger (The Open University, GB)*


License  Creative Commons BY-NC-ND 3.0 Unported license  
© Stefan M. Rüger

What is wrong with a picture of Tony Blair hunting and pecking the keyboard in the midst of school children staring at their respective screens? What works on the web, and what doesn't? Can we use Linked Open Data to persist and share experimental setups? Stefan's talk reflects on the value and potential of interlinked data, semantic web, social networks and data-mining. At the same time he elicits new research directions, which are only enabled by the sheer mass of data, sensors, facts, reports, opinions and inter-linkage of people.

There are a number of information requests that traditional web services can cover well: shortest distance from A to B, opening times of theatre plays, book reviews given a snapshot of a book cover, extensive and competent wikipedia articles on virtually every aspect of our lives. We can hardly imagine a world that does not offer these web services, which have complemented traditional methods of resource discovery, say in libraries. It is quite possible that automated processing of excessive amounts of data paired with new methods of semantic web, information retrieval, human information interaction, social networks, and cloud computing opens up possible new areas of knowledge discovery: Will we have wikiversities and do we want them? Will activities such as TrueKnowledge and Wolfram Alpha be in a position to answer all complex questions that have factual answers? Will the tradition of linear argumentation in printed scientific articles give way to graphical, linked argumentation landscapes? What would be possible or necessary instruments to answer questions such as "What led to the most recent war in Iraq?" Can we create digital research labs such as virtual microscopes and telescopes? What can we learn through computers being able to learn 1 million books? Or watch the news of thousands of television channels in 100 countries?

#### 3.2 Full Lifecycle Information Extraction and Multiparadigm Indexing with GATE

*Hamish Cunningham (University of Sheffield, GB)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Hamish Cunningham

Hamish summarised the last few years work with GATE (<http://gate.ac.uk/>), including developments around collaborative manual annotation, process support tools, cloud computing and a mixed-mode index server: "How I was sentenced to 20 years hard labour and the potential implications for over-priced IR systems":

We talk, we write, we listen or read, and we have such a miraculous facility in all these skills that we rarely remember how hard they are. It is natural, therefore, that a large proportion of what we know of the world is externalised exclusively in textual form. That fraction of our science, technology and art that is codified in databases,

taxonomies, ontologies and the like (let's call this *structured data*) is relatively small. Structured data is, of course, machine tractable in ways that text can never be (at least in advance of a true artificial intelligence, something that recedes as fast as ever over the long-term horizon). Unfortunately structure can also be inflexible and expensive to produce in ways that text is not.

Language is the quintessential product of human cooperation, and this cooperation may well have shaped our capabilities and our culture more than any other single factor. Text is a beautiful gift that projects the language of particular moments in time and place into our collective futures. It is also infernally difficult to process by computer (a measure, perhaps, of the extent of our ignorance regarding human intelligence).

When scientific results are delivered exclusively via textual publication, the process of replicating these results is often inefficient as a consequence. Although advances in computational platforms raise exciting possibilities for increased sharing and reuse of experimental setups and research results, still there is little sign that scientific publication will cease its relentless growth in the near future.

This talk summarises a research programme (now 20 years old) that has resulted in GATE, a General Architecture for Text Engineering (<http://tinyurl.com/gatebook>). In recent years GATE has grown from its roots as a specialist development tool for text processing to become a rather comprehensive ecosystem bringing together software developers, language engineers and research staff from diverse fields. GATE now has a strong claim to cover a uniquely wide range of the lifecycle of text-related systems. It forms a focal point for the integration and reuse of advances that have been made by many people (the majority outside of the authors' own group) who work in text processing for biomedicine and other areas. The talk seeks to draw together a number of strands in this work that are normally kept separate, and in so doing demonstrate that text analysis has matured to become a predictable and robust engineering process. The benefits of deriving structured data from textual sources are now much easier to obtain as a result.

In line with the trends towards openness in life sciences R&D and in publishing, GATE is 100% open source. This brings the usual benefits that have been frequently recognised (vendor independence; security; longevity; flexibility; minimisation of costs; etc.). Less often remarked upon but nonetheless significant in many contexts are traceability and transparency. Findings that are explicable and fully open are often worth much more than results that appear magically (but mysteriously) from black boxes.

[Hamish Cunningham, *The Infernal Beauty of Text*, 2011.]



### 3.3 Unsupervised and Semi-Supervised Approaches to Cross-Domain Sentiment Classification

*John A. Carroll (University of Sussex – Brighton, GB)*

**License** © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license  
© John A. Carroll

**Joint work of** Carroll, John A.; Read, Jonathon; Zagibalov, Taras; Bollegala, Danushka; Weir, David  
**Main reference** Danushka Bollegala, David Weir and John Carroll, “Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification,” Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.  
**URL** <http://www.informatics.sussex.ac.uk/research/groups/nlp/carroll/papers/acl11.pdf>

In recent work we have been addressing the problem of developing accurate sentiment classifiers for domains and languages for which there is little or no appropriate training data.

In one strand of this work, we have developed a sentiment classification method that is applicable when we do not have any labeled data for a target domain but have some labeled data for multiple other domains. The approach is based on automatically creating a sentiment sensitive thesaurus in order to find the association between words that express similar sentiments in different domains. Unlike previous cross-domain sentiment classification methods, our method can efficiently learn from multiple source domains.

In a second strand, we have developed a novel unsupervised sentiment classification technique, based on a ‘seed’ vocabulary of positive/negative words, and iterative retraining to bootstrap a substantial high quality training corpus from unlabelled documents. The technique has been applied successfully to text in Chinese, Japanese, Russian and English.

### 3.4 Challenges in Mining Social Media: Sparsity and Quality

*Thomas Gottron (Universität Koblenz-Landau, DE)*

**License** © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license  
© Thomas Gottron

**Joint work of** Naveed, Nasir; Gottron, Thomas; Kunegis, Jérôme; Che Alhadi, Arifah  
**Main reference** Nasir Naveed, Thomas Gottron, Jérôme Kunegis and Arifah Che Alhadi, “Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter,” Proc. 3rd ACM International Conference on Web Science (WebSci’11).  
**URL** [http://www.websci11.org/fileadmin/websci/Papers/50\\_paper.pdf](http://www.websci11.org/fileadmin/websci/Papers/50_paper.pdf)


Online communities generate large amounts of text based contents. These contents, however, are very different under several aspects. They cover a wide variety of topics and languages, differ in style and length and are of very different quality. Especially the short texts in microblogs and their sparsity of features pose challenges for many applications in the fields of text retrieval, classification or clustering.

We enrich the representation of microblog entries by annotating them with second level features, such as topics or sentiments. This richer representation can then be used to derive a notion of content quality for social media. Already by itself, this allows for interesting insights into the dynamics of social media.

Preliminary results further suggest that this notion of content quality can be used as a static quality measure to improve retrieval on microblogs.

### 3.5 Simulation Data Mining in Artificially Generated Data

*Steven Burrows (Bauhaus-Universität Weimar, DE)*


License  Creative Commons BY-NC-ND 3.0 Unported license  
© Steven Burrows

Systems simulation and data mining can complement one another to form simulation data mining and harness intelligence to automatically suggest improvements to simulation models. One challenge in simulation data mining is to develop recommendations for competing variables. In aviation, for example, simulation data mining has been applied to develop cost-effective compromises between flight time and maintenance efforts over aircraft lifetimes.

This talk will introduce a simulation data mining project called Matilda (Mining Artificial Data). In this project, a broader view of document mining is taken to consider artificially generated non-text documents containing data for bridge specifications. It will be shown how simulation data mining can be applied to support the interactive design of bridge structures by removing some of the work required to manually simulate design iterations in turn. Another area of potential is to apply clustering algorithms to automatically identify related models with the development of appropriate distance measures.

### 3.6 Unsupervised entailment detection for IE query expansion

*Ted Briscoe (University of Cambridge, GB)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Ted Briscoe

Main reference BioNLP 2011, A workshop of ACL/HLT 2011, Portland, Oregon, USA  
URL <http://compbio.ucdenver.edu/BioNLP2011/program.shtml>

Query expansion for IE systems hasn't been explored much, but would be useful in contexts where, say, the user identified a prototypical predicate denoting a relation of interest. The first step in such an approach would be to detect entailed predicates given the user-defined predicate.

Entailment detection systems are generally designed to work either on single words, relations or full sentences. We develop a new approach – detecting entailment between dependency graph fragments of any type – which relaxes these restrictions and leads to much wider entailment discovery. An unsupervised framework is described that uses intrinsic similarity, multi-level extrinsic similarity and the detection of negation and hedged language to assign a confidence score to entailment relations between two fragments.

### 3.7 Piggyback: Using Search Engines for Robust Cross-Domain Named Entity Recognition

*Hinrich Schütze (Universität Stuttgart, DE), Massimiliano Ciaramita (Google Zurich)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Hinrich Schütze, Massimiliano Ciaramita


Joint work of Schütze, Hinrich; Ciaramita, Massimiliano  
URL <http://ifnlp.org/schuetze/piggyback11/piggyback11.pdf>

We use search engine results to address a particularly difficult cross-domain language processing task, the adaptation of named entity recognition (NER) from news text to web

queries. The key novelty of the method is that we submit a token with context to a search engine and use similar contexts in the search results as additional information for correctly classifying the token. We achieve strong gains in NER performance on news, in-domain and out-of-domain, and on web queries.

### 3.8 Feedback Methods in Large Scale Visual Document Analysis


*Michael Granitzer (Know-Center Graz, AT)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Michael Granitzer

Michael discusses feedback methods utilizing visual representations of large document sets in order to adapt algorithmic parameters and metrics; in this regard, he presents a system for visualizing and analyzing topical, temporal and metadata-based correlations in large document sets. The methods combine well-known visualization techniques for large document sets (i.e. multi-dimensional scaling, self-organized maps) with recent techniques to learn high-dimensional metrics. His talk focuses on two methods: (1) a user can directly manipulate the parameters of the algorithm generating the visualization, and (2) a user can interactive label visualized elements. For getting an impression of the content of a document, Michael shows that the effectiveness of key phrases are as good as text summaries, but much more efficient.

### 3.9 The Optimum Clustering Framework

*Norbert Fuhr (Universität Duisberg-Essen, DE)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Norbert Fuhr


Starting point are the following questions: Is there a principle similar to the Probability Ranking Principle for document clustering? Can we define a cluster metric in which quality is relative to shared relevance to a set of queries?

To answer these (and related questions) Norbert introduces a theoretic foundation for optimum document clustering. Key idea is to base cluster analysis and evaluation on a set of queries, by defining documents as being similar if they are relevant to the same queries. Three components are essential within this optimum clustering framework (OCF): (1) a set of queries, (2) a probabilistic retrieval method, and (3) a document similarity metric.

Based on these components appropriate validity measure can be introduced, and optimum clustering can be defined with respect to the estimates of the relevance probability for the query-document pairs under consideration. Norbert shows that well-known clustering methods are implicitly based on the three components, but that they use heuristic design decisions for some of them. Altogether, the OCF can help to make targeted research for developing better document clustering methods possible.

### 3.10 TIRA – Research Assistant for Empirical Evaluations

*Tim Gollub (Bauhaus-Universität Weimar, DE)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Tim Gollub

Joint work of Gollub, Tim; Stein, Benno


We present and demonstrate TIRA, software that supports the data mining community in conducting replicable and comparable experimental analysis.

Although desired, the comparison of empirical evaluations in scientific publications is often impossible due to differing datasets, baselines, or parameter settings. Furthermore, details concerning the implementation of algorithms are not given in many cases. With TIRA, we want to contribute to a more efficient and cooperative research community that benefits from sharing data, algorithms, experiments, and CPUs.

TIRA provides an open framework for designing, running, and publishing data mining experiments. We try to acquire a comprehensive list of datasets, algorithms, and evaluation metrics. TIRA manages parametrization and distribution of experiments, as well as storing results for further performance studies and visualizations.

### 3.11 LFRP-Search: Multi-Layer Faceted Search with Ranking and Parallel Coordinates – Interactive Retrieval for Complex Documents and Individual Information Needs

*Andreas Henrich (Universität Bamberg, DE)*





License  Creative Commons BY-NC-ND 3.0 Unported license  
© Andreas Henrich

In enterprise search scenarios information needs and documents are quite diverse. In addition, information needs often are vague and unclear, which means that users cannot explicitly define the search criteria that specify their search request. In these cases, exploratory search approaches are necessary to support users in interactively refining their search queries.

To address such settings, Andreas presents a retrieval system for complex search situations that is based on four constituent parts. The approach deals with the heterogeneity of potential target objects when performing a search considering multiple artefact layers (e.g., projects, products, persons, and documents). The overall system is designed as a kind of data warehouse which supports relations between artifact types and considers them in ranked retrieval. The biggest effort for building such a system is the specification of the ETL (extract / transform / load) processes for the different sources. To cope with result sets of different granularity, ranking facilities based on facet values as well as Query-by-Example functionalities are included. Parallel coordinates are used to visualize the characteristics and dependencies of (intermediate) results in order to provide users with a deeper understanding of the data under investigation. In his talk, Andreas discussed the conflicting priorities of ease of use and expressive power.

### 3.12 Beyond Search: Interactive Web Analytics

*Alexander Loeser (TU Berlin, DE)*





License     Creative Commons BY-NC-ND 3.0 Unported license  
© Alexander Loeser

Today, the Web is one of the world's largest databases. However, due to its textual nature, aggregating and analyzing textual data from the Web analogue to a data warehouse is a difficult problem. For instance, users may start from huge amounts of textual data and drill down into tiny sets of specific factual data, may manipulate or share atomic facts, and may repeat this process in an iterative fashion.

Andreas presents the GooLap System (<http://www.goolap.info/>) for interactive fact retrieval from the Web supporting several dozens of named entity types and the corresponding relationships. The keyword-based query interface focuses on both, simple query intentions, such as, “display everything about Airbus” or even complex aggregation intentions, such as “List and compare mergers, acquisitions, competitors and products of airplane technology vendors.” For retrieval, he describes a new query language which also allows for joins on the extracted fact relations. Andreas discusses the fundamental problems in the iterative fact retrieval process: What are common analysis operations of “business users” on natural language Web text? What is the typical iterative process for generating, verifying and sharing factual information from plain Web text? Can we integrate both, the “cloud”, a cluster of massively parallel working machines, and the “crowd”, such as users of GoLAP.info, for training 10.000s of fact extractors, for verifying billions of atomic facts or for even generating analytical reports from the Web?

### 3.13 The Retrievability of Documents

*Leif Azzopardi (University of Glasgow, GB)*


License     Creative Commons BY-NC-ND 3.0 Unported license  
© Leif Azzopardi

How do individuals interact with information? Can we exploit models from transportation and planning to derive new suite of measures for information retrieval?

Leif introduces the concept of accessibility from the field of transportation planning and explains how it can be adopted within the context of information retrieval. By drawing analogy between the two fields we are able to develop a new suite of measures for information retrieval that considers how easily a document can be retrieved using a particular retrieval system. This intuitive measure provides the basis for examining different aspects of the influence of a retrieval system has upon the documents in the document collection. He shows that distribution of retrievability values over a document collection depends on the employed retrieval model and discusses possible ways in which retrievability measures might be used for document mining tasks.

### 3.14 Looking at Document Collections from a Bird’s Eye View: Exploratory Search as Based on the Context Volatility of Terms

*Gerhard Heyer (Universität Leipzig, DE)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Gerhard Heyer

Was is exploratory search? In many text retrieval applications the user is not primarily interested in facts, but more so in the way facts are reported on, and how the conceptualizations and judgments expressed in the reporting texts are changing over time. Investigations in journalism, technology mining, or eHumanities are examples that can be considered as instances of exploratory search. The classical paradigm of query-and-retrieve is not suitable here because a user who wants to retrieve those documents that best satisfy his information needs first needs support (a) to make himself familiar with a search domain, (b) to identify terms that are of potential interest to the topic he is researching, and (c) to follow variant paths to explore the domain of interest.

Gerhard presents the notion of context volatility of terms as a measure for the interest-iness of terms. Basically, this measure computes the variance of a term’s co-occurrences during some period of time. Given a time-stamped collection of documents, terms with a high degree of context volatility for some period of time usually represent “hot topics” as they have been controversially discussed during that period. Using this measure of context volatility, he presents a system for exploratory search that for a time-stamped collection of documents, such as the New York Times corpus, (1) generates a list of “hot topics”, and (2) supports the user to identify those periods of time where the discussion of these “hot topics” erupts. The whole search process is highly interactive, and an instructive instance of visual analytics.

### 3.15 Facet-Streams and Search-Tokens – Tangible User Interfaces for Information Seeking

*Harald Reiterer (Universität Konstanz, DE)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Harald Reiterer

Social activities such as collaborative work and group negotiation can be an essential part of information seeking processes. However, they are not sufficiently supported by today’s information systems as they focus on individual users working with PCs. Reality-based UIs with their increased emphasis on social, tangible, and surface computing have the potential to tackle this problem. By blending characteristics of real-world interaction and social qualities with the advantages of virtual computer systems, they inherently change the possibilities for collaboration, but until now this phenomenon has not been explored sufficiently. Harald presents two examples of Tangible User Interfaces (TUIs) and analyzes their power and expressiveness for information seeking activities.


The first example is “Facet-Streams”, a hybrid interactive surface for co-located collaborative product search on a tabletop. Facet-Streams combine techniques of information visualization with tangible and multi-touch interaction to materialize collaborative search on a tabletop. It harnesses the expressive power of facets and Boolean logic without exposing users to complex formal notations. The second example is “Search-Tokens”, also a hybrid

interactive surface for co-located collaborative information exploration on a tabletop. The physical appearance of the Search-Tokens provides a higher visual and tangible affordance than a GUI that is solely based on digital sliders, text fields, buttons, etc. By placing a Search-Token on the tabletop, it is augmented by visualizations, when a search criterion is entered, rotating the Search-Token allows users to define the criterion's weight.

But, how useful are these new user interaction paradigms? Can search be made more like interacting with the non-digital world? Harald reports on user studies that reveal how visual and tangible expressivity are unified with simplicity in interaction, or how different strategies and collaboration styles are supported.

### 3.16 What do you mean? – Determining the Intent of Keyword Queries on Structured Data

*Wolf Siberski (Universität Hannover, DE)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Wolf Siberski


More and more factual information is mined from the Web and provided as structured data, for example in the Linked Data initiative. While this opens up the potential for fulfilling information needs much better than just Web page content, it also requires new approaches to retrieval, because relevance functions for text don't work well on structured data, and complex query languages are the wrong tool for most users.

As one such novel approach, we have developed QUICK, which combines the convenience of keyword search with the expressiveness of structured queries.

Users start with a keyword query and then are guided through a process of incremental refinement steps to specify the intention of their query. We show how QUICK identifies possible intentions and computes a construction wizard for query intent specification with a few mouse clicks.

### 3.17 Exploiting Unlabeled Data in Information Retrieval Tasks

*Benno Stein (Bauhaus-Universität Weimar, DE)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Benno Stein


Joint work of Stein, Benno; Nedim, Lipka; Meyer zu Eissen, Sven; Prettenhofer, Peter

Data that is labeled with class information is a valuable and expensive resource, and there are different efforts to combine labeled data with unlabeled data to improve classifier effectiveness during machine learning. We discuss three approaches, namely constrained clustering, co-training, and structural correspondence learning, since they pursue quite different data exploitation paradigms. All approaches can be called semi-supervised, whereas constrained clustering exploits additional labeled data in order to inform an - usually - unsupervised analysis. By contrast, co-training as well as structural correspondence learning exploit additional unlabeled data. With respect to the former, we explain why it is difficult to scale-up the co-training idea, i.e., the extension of the training set. With respect to the latter we introduce a cross-language sentiment classification approach where the use of unlabeled

data scales extremely well, leading to a significant improvement over state-of-the-art machine translation baselines.

### 3.18 Cross-lingual adaptation using structural correspondence learning

*Peter Prettenhofer (TU Graz, AT)*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Peter Prettenhofer

**Joint work of** Prettenhofer, Peter; Stein, Benno

**Main reference** Peter Prettenhofer and Benno Stein, “Cross-lingual adaptation using structural correspondence learning,” *ACM Transactions on Intelligent Systems and Technology* (to appear), ACM, 2011

Cross-lingual adaptation is a special case of domain adaptation and refers to the transfer of classification knowledge between two languages. We describe an extension of Structural Correspondence Learning (SCL), a recently proposed algorithm for domain adaptation, for cross-lingual adaptation in the context of text classification. The proposed method uses unlabeled documents from both languages, along with a word translation oracle, to induce a cross-lingual representation that enables the transfer of classification knowledge from the source to the target language. The main advantages of this method over existing methods are resource efficiency and task specificity.


We report on experiments in the area of cross-language topic and sentiment classification involving English as source language and German, French, and Japanese as target languages. The results show a significant improvement of the proposed method over a machine translation baseline, reducing the relative error due to cross-lingual adaptation by an average of 30% (topic classification) and 59% (sentiment classification). We further report on empirical analyses that reveal insights into the use of unlabeled data, the sensitivity with respect to important hyperparameters, and the nature of the induced cross-lingual word correspondences.

#### References

- 1 Peter Prettenhofer and Benno Stein. *Cross-lingual adaptation using structural correspondence learning*. *ACM Transactions on Intelligent Systems and Technology* (to appear), 2011
- 2 Peter Prettenhofer and Benno Stein. *Cross-language text classification using structural correspondence learning*. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics*, Uppsala, Sweden, 2010

### 3.19 Search by Strategy

*Arjen P. de Vries (CWI Amsterdam, NL)*

**License**  Creative Commons BY-NC-ND 3.0 Unported license  
© Arjen P. de Vries

Today Dagstuhl, tomorrow the world. Making IR and DB engineers redundant with a dynamically reconfigurable DB-based IR system (<http://devel.spinque.com/BilthovenDemo/>). For interactive information access Arjen and his group has been developing a system that allows for visually constructing a search strategy by connecting building blocks. The backend is a combined IR-DB system based on probabilistic relational algebra.


The topic of his talk is focused on the process after text mining has taken place – assuming that the mining process would lead to semantic annotations of the original document space.



The key idea is to let users decide for themselves how to navigate this “semantically” enriched document space, but then do support them in exploration of this annotated data. Hereto, a new interaction paradigm is followed, called “search by strategy”: an iterative two-stage search process that separates search strategy definition (the “how”) from the actual searching and browsing of the collection (the “what”). A visual query environment (the “search strategy builder”) allows professional searchers to visually express their search strategy for a particular need, and then execute this strategy on a probabilistic relational database system. The idea is that by bringing together the visual search strategy definition and faceted browsing of result sets will allow the user to discover during query formulation which semantic annotations are useful for their information need, and exploit their value for finding better search results in less time.

### 3.20 Challenges in Patent Retrieval and Mining

*Dennis Hoppe (Bauhaus-Universität Weimar, DE)*

License  Creative Commons BY-NC-ND 3.0 Unported license  
© Dennis Hoppe

Over 50 million multi-language patents no longer permit that we “search as we did 28 years ago,” answered Henk Tomas in 2007 at the IRF Symposium as he was asked about the greatest problem of patent retrieval. Patent retrieval concerns not only companies and inventors, but also occasional users and researchers.

Current challenges in patent retrieval include the need for multi-language support, robustness against errors in patent text caused by error-prone OCR recognition, handling inconsistencies in metadata, and yielding a high recall while retrieving patents. However, various issues make the retrievability of patents difficult: (1) domain-specific vocabulary, (2) complex syntactic structure, (3) vague translations of companies and persons from far-east, and (4) companies change their names over time, merge or demerge.

The emphasis of this talk is on methods to improve patent retrievability such as automated patent classification, and image-based patent retrieval and classification.

### 3.21 Cancelled Talks

A few participants were unable to present their talks due to unforeseen circumstances:

- C. J. Keith van Rijsbergen (University of Cambridge) Title: Document Clustering Revisited
- Ingo Frommholz (University of Bedfordshire) Title: A Quantum-inspired Polyrepresentation Framework
- Oliver Niggemann (Hochschule Ostwestfalen-Lippe) Title: A Probabilistic MajorClust Variant
- Marc Lechtenfeld (Universität Duisburg-Essen) Title: Determining the Polarity of Postings for Discussion Search
- Oren Etzioni (University of Washington)

## 4 The Working Groups

The following working groups were proposed by the participants:

- Out-takes: negative results in information retrieval.
- Conference reviewing mechanisms.
- Apparatus, setups, infrastructure for reproducibility.
- Human-in-the-loop scenarios, evaluation.
- Automated readers: knowledge extraction from on-line million-book libraries.
- Retrieval of complex structures.
- Sentiment, opinion, and social media.
- Optimum clustering.

From these suggestions we selected the following:

1. Apparatus, setups, infrastructure for reproducibility, plus Human-in-the-loop scenarios, evaluation.
2. Reading 10 million books. Automated readers: knowledge extraction from on-line million-book libraries.
3. Retrieval of complex structures. Cross-over of data-oriented methods and retrieval methods: models of structured data retrieval (e.g. entities, entity graphs).
4. Sentiment, opinion, and social media.

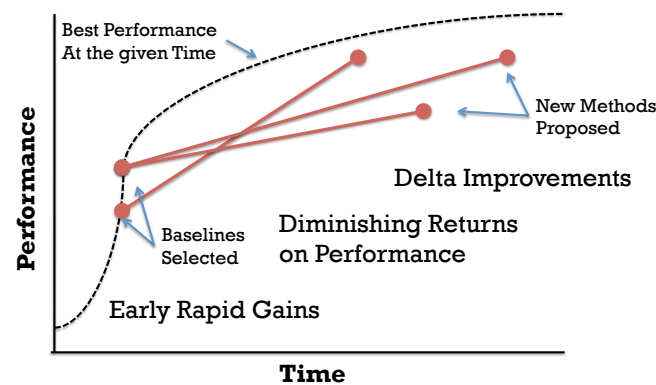
### 4.1 Apparatus, Reproducibility, and Evaluation

Rapporteur: Leif Azzopardi

#### 4.1.1 Building Research Infrastructure

This working group considered the design and development of common frameworks, applications and resources for experimentation in Information Retrieval (IR) and Document Mining (DM). The group considered a number of aspects on this topic: (1) the current problems with evaluations and comparisons of methods, models and systems, (2) the components and building blocks of experimental research, the research process and the infrastructure, and (3) the ideal research infrastructure, its benefits, drawbacks, barriers and challenges.

Initial discussions acknowledged that many of the experiments and empirical work that is currently undertaken is performed independently and often under very different conditions. Each research group and potentially each researcher designs and develops experimental scripts and processes to conduct experimental research to determine the effectiveness and efficiency of methods, models and systems. All the possible variables and factors in the experimental process (such as the algorithm implementations, the choice of the algorithms/methods used, the toolkits used, the data processing, the parameter tuning, etc) vary across the body of research reported in the field. In conjunction with the pressure to publish results that show "significant" differences, this can lead to overly optimistic results being reported. Subsequently, published work often contains comparisons where poor baselines have been used, newly proposed methods are over trained, or some other manipulation to help bolster the evidence for the method (to try and maximise the chance of publication). The consequence of this is that many of the experiments and the results reports and methods described in the literature are difficult to replicate and validate. Also, it is difficult to contextualize any performance improvements made by different retrieval and mining methods across the research area.

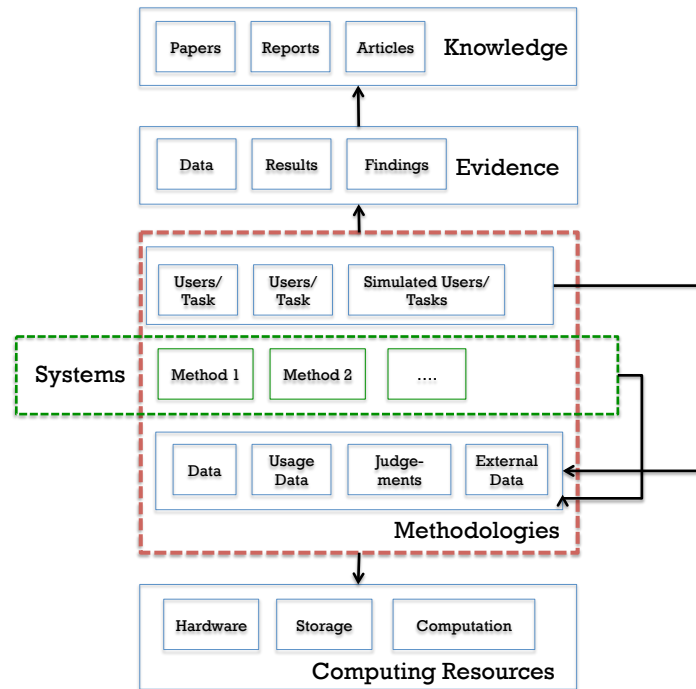


■ **Figure 1** The Armstrong Effect: The current best performance is rarely used when proposing a new method. Instead, an older baseline is preferred and used. The dotted line represented a smoothed version of the overall progress in the field in terms of performance for a given task.

For example, in Amstrong et al [?], they showed that evaluations conducted within IR rarely used the current state of the art to perform comparisons. Instead, older and weaker baselines were used in order to compare the proposed methods. As a result, newly proposed models would significantly outperform the weak baselines, yet it would rarely outperform the current state of the art. In Figure 1, we have illustrated this effect, where the dotted black line denotes the best performance at the time (smoothed across innovations), and the lines joined by balls denote a baseline vs. a proposed method. The shape of the dotted black line shows that initially there is usually early rapid gains in performance when a task is introduced. Then diminishing returns in performance begins to kick in, before the work becomes delta (i.e. only marginal improvements are being achieved by the best newly proposed methods. If the actual performance across the field was known then the progress being made on a particular tasks could then be readily contextualized and compared accurately.

Otherwise, without a centralized repository of results and access to standardized configurations it is difficult to determine whether significant progress is actually being made in an area or not. i.e. have we reached the top of an innovation curve/cycle with the current methods and technology? Also, currently, reviewers must invest a lot of trust in the work presented to them, and they have to assume that the reported results hold (and that the experiments and findings are repeatable and reproducible by others. For those replicating existing methods, they face a gambit in that, if they discover that they can not reproduce findings similar to past work, or worse they obtain contradictory finding, then they run the risk of trying to publish negative findings, which was perceived to be almost impossible. This means many researchers waste a lot of time discovering certain methods do not perform as well as purported, leaving them in a difficult position. Essentially, there was a consensus among participants in the working group that the current way we perform research leaves a lot to be desired. And a potential solution would be to provide some sort of common evaluation infrastructure, which extends and enhances current evaluation forums like TREC, CLEF, etc. The infrastructure would provide a standardized environment where:

1. experiments are reproducible and repeatable,
2. verification of results is performed externally and independently,
3. the cost of experimentation is significantly reduced, and
4. data from previous experiments and methods would be available for comparison purposes.



■ **Figure 2** An abstraction of the research process and supporting research infrastructure.

It was hypothesized that if a common framework for experimentation existed which would enable highly controlled evaluation, then the above problems may be mitigated. While, it would provide a number of other benefits, it was also acknowledged that there is likely to be a number of potential problems in designing and developing such infrastructure.

To discuss this hypothesized framework, we took a very abstracted view of what such infrastructure might be, and assumed task independence. i.e. we imagined that the framework could be applied to various retrieval and mining tasks, like ad-hoc retrieval, classification, clustering, etc. The purpose of such a framework would be to:

- enable the evaluation of systems, methods and models,
- ensure replication and repeatability of experiments,
- facilitate the dissemination and reporting of research results,
- improve experimenter efficiency by lowering the start up costs of running and replicating experiments,
- focus research on a particular well defined task, and
- provide standardization, automation and controlled experimentation.

#### 4.1.2 Towards an Ideal Research Infrastructure

Figure 2 provides a high-level schematic of our perspective on the research process and how the research infrastructure differs from existing evaluation forums and tools. The end goal of the research process is to generate new knowledge, to question and verify existing knowledge and to revise knowledge. These conclusions and implications are supported by the experimental data collected, the results obtained and the findings made. To create this experimental data a scientific methodology is employed, and in IR and DM, usually consists of a several key components:

- Different sets of users (even simulated users) and corresponding tasks, potentially along with their interactions.
- A system (or set of sets), such as Lemur or Terrier, to perform the retrieval or mining with, and a series of different methods/models that are to be tested.
- Data to form the test collections for evaluation, the judgments, external data and the measurements that are to be used.

Put together a set of these components creates a methodology. Evaluation forums like TREC and CLEF often define the process and the components, and let the systems and methods be varied. The research infrastructure would provide a system or an interface for systems to enable the testing of methods (or combinations of methods), house a repository of data and results. The ideal infrastructure would be able to cater for various users, including researchers, academics, businesses, living labs and organizations that facilitate research. It was acknowledged that different users will have their own different needs, concerns, constraints and problems. But they should also be able to benefit from the infrastructure as well.

*Collaboration, Cooperation and Competition in Research:* Having an abstraction of the ideal infrastructure, then promoted discussion in the working group on how collaboration and cooperation between researchers and research groups might improve. It was mooted that, the research infrastructure would be able to promote sharing and trust between researchers. And this would create a cooperative environment where science of IR is driven forward in a transparent manner. The research infrastructure would also help to keep researchers honest, because all the results and findings are independently verified and recorded by the infrastructure. Though the collaboration and cooperative it would be possible to perform reproducible and repeatable experiments. So while, *Competition drives Innovation* it can potentially obfuscate and diffuse the progress of the research within the field. Through cooperation the field may be able to identify very quickly areas of research where increases in performance are marginal and becoming delta.

However, the decision by researchers to cooperate or to compete presents the *The Researcher's Dilemma*. If researchers choose not to cooperate, then we continue with the current state of evaluation within IR and DM, i.e. piecemeal and ad-hoc testing and reporting of results. As mentioned above, the downside is that progress is obfuscated and misleading. However, if researchers cooperate by building an open research infrastructure, then there is the potential to advance progress in the area rapidly, and to identify when progress in a particular topic is converging (leading to only marginal or delta improvements). Of course, there may be researchers that cooperate, and others that stay in competition. Here, the cooperative researchers, run the risk of losing out to the competitors. The cooperative researchers are likely to forgo obtaining short term gains through publications, because they are investing in the development of the infrastructure. If the infrastructure fails to be adopted or does not obtain critical mass then the cooperative researchers will incur even greater losses. The competing researchers would be able to gain the upper hand, by continuing to publish without having to conform to any standard. For a competing researcher to switch and become a cooperative researcher is also a difficult choice, as they may already have custom built infrastructure which is giving them a strategic advantage over other researchers. Without some external regulation (if reviewers demanding independent verifications, for instance, or funding agencies requiring publicly accessible results), then the cost of switching is likely to be too high. Given the publish or perish environment, this will potentially provide the competing researchers to stay in the game, while unsuccessful cooperative researchers will need to find work elsewhere (in particular for early researchers). An alternative is if the

cooperative forms a closed piece of research infrastructure, where only those participating have access. This may provide enough protection for cooperative researchers in the short term to mitigate the risks. Nonetheless, researchers face a decision in how to act and work. At times, we compete and other times we cooperate. In terms of building research infrastructure, a clear case for the different positions and what can be gained through cooperation needs to be put together.

### 4.1.3 Challenges and Barriers

*Barriers to Adoption:* While the idea of having research infrastructure in place was generally viewed as a positive direction for the field, we also recognized that there are a number of barriers to its adoption. These included:

- the framework/infrastructure may impose too many constraints
- if the infrastructure is difficult to use or requires a steep learning curve
- the cost of setting up experiments in the new infrastructure might too high, and not worth the switch, and
- researchers may be unwilling to give up any strategic advantage that they currently have with their own custom-built infrastructure.
- researchers may not want to participate because their methods may not achieve the same performances that they have either reported and/or produced under their own conditions.

*Success Factors:* For the success of a common research infrastructure, the group identified a number of key factors which would help its adoption, and provide benefit to its users. These were:

- Getting a critical mass of participants, in particular the major research groups in the field
- Providing a flexible and general interface that can support the various kinds of research methods to be tested
- Providing a reasonable good level of support, documentation and ensuring that the infrastructure is easy to use
- Ensuring engagement and usage through forums and workshops (such as TREC, CLEF, etc)
- Providing transparency in the research process
- Lowering the cost of research, and lowering the entry/start up costs to research
- Valuing of the results and findings from the research infrastructure by reviewers over those from out-with such infrastructure, since the research infrastructure will be providing independent verification of the results.

*Further Challenges:* The group mainly discussed the issues associated with the design and development of research infrastructure such that it would support experimental research in IR and Data Mining. A major challenge is to operationalize this vision and to produce a concrete design of the infrastructure, and then develop it. Then, further challenges will emerge in terms of sustainability and maintainability. To address these resources will be needed to create and then sustain the initiative. However, with the development of such infrastructure likely to be quite costly, requiring time, money and manpower, it is likely that such an initiative would require a long-term project to be funded.

## 4.2 What can computers learn from reading one million books?

This working group took up Greg Crane’s question “What Do You Do with a Million Books?” (D-Lib Magazine, 12(3), March 2006) and contextualised it to the Dagstuhl workshop theme of document mining.

An active reading life of 70 years enables us to read around 25,000 books in a lifetime assuming one can finish one book per day on average. Ongoing digitisation efforts have, of course, created many more books in computer-readable form. So, what could we mere humans learn from computers that read one million books for them? And what does it mean that a computer reads a book, for that matter?

To give a simple example, one hope is that algorithms are able to assign one or more topics to passages of text. These topics are embedded in the context of other topics and sentiment (see also the workgroup on sentiment and opinion in Section 4.4) as well as being spread over place and time. This alone should provide for interesting automated analysis how the topic of, for example, slavery, has shifted over time, with different speed in the UK and the US, where slavery was abolished much later.

Our working group revolved around developing a set of challenges and then exploring possible methods and approaches to tackle them.

### 4.2.1 Challenges v1.0

We discussed a range of challenges and activities for computer algorithms; here is our summary of key challenges roughly in ascending order of difficulty:

- Identify key phrases — already happening in any self-respecting natural language processing (NLP) package
- Detect topics — a more difficult task, as it involves document understanding, but already currently tackled more or less successfully by the NLP community
- Discover memes or ideas — involves yet another level of abstraction
- Visualise peculiarities and outliers given a topic, for example, by creating a bird’s eye view of a topic
- Establish relations, trends and geographic patterns of key phrases, topics, and ideas
- Follow the provenance, propagation and dispersion of ideas
- Compute sentiment associations with topics, ideas and text passages
- Classify documents as novels, scientific papers, recipes, advertisement, travel reports, poems, newspaper articles, school books, essays, manuals, magazine articles, or similarly for images, classify these as landscape scenes, portraits, chemical structures, electrical circuits, mathematical formulae, architectural drawings etc
- Extract summaries of individual documents by assigning importance to sentences and keeping the highest rated ones
- Bias summaries on queries reflecting the information need of the user
- Elicit change of meaning of words, topics or ideas over time (historic semantics), and as corollary, answer the question which words and topics are invariant
- Recommend related material
- Spot prototypical ideas and creativity
- Recognise important documents (original or influential)
- Create a database of facts to be used for question answering tasks
- Enable focussed retrieval for facets and references — with particular focus on the long tail
- Delineate variable quality of writing and writing style

- Generate multi-document summaries
- Compare topics across different languages
- Assign authorship
- Uncover contradictions, lies or polemic
- Describe the knowledge and world view of a period — lending support to the discipline of history of science
- Explore limitations of the world view of a time (including those of the currently held world view!)

These challenges are all connected to the overriding aim to understand documents through machine processing, and some of these assume that there are sufficiently many documents such that a meaningful analysis can be carried out in time windows and that the results inform about (continuous) changes in time.

#### 4.2.2 Challenges v2.0

The next level of challenges would be to attempt an automated *understanding of language* rather than a particular document, broadly assuming that the first level of challenges will have been mastered in the foreseeable future to sufficient quality. Automated language understanding will have been achieved if, for example, computers can

- Generate novels or poetry automatically
- Establish and deploy models of creativity
- Tell apart irony from parody and facts from fiction
- Read aloud and correctly intonate books
- Interpret graphs, figures and diagrams — for example electrical-engineering circuits or chemical structures
- Predict differences in cultures
- Analyse patents as to prior art

#### 4.2.3 Methods

The discussions revolved around the current best practice for attempting some of the challenges that we identified, with a less expressed speculation on what could be used to tackle the more difficult or frankly at this point utopian challenges.

*Statistics.* A simple counting exercise can yield amazing insights, particularly in very large datasets. For example, a count of the references to locations on the globe in the English-language wikipedia can reveal the areas of the world that are of particular interest to the English-speaking population, ie, yield their “view of the world geography”.

Co-occurrence analysis of words can give key insights, too. For example, terms that occur together and have a similar profile of frequency and volatility may be good candidates for synonyms.

Statistical methods are not confined to counting. They subsume probabilistic language models, the use of ontologies, factor analysis and latent semantic indexing, to name but a few. The latter have shown to be able to crystallise memes, ideas and topics in documents, which in turn are the basis of semantic profiling and many a similarity function between document parts.

*Citation analysis and bibliometrics.* The overwhelming majority of books and text documents do not feature the explicit references that are common in scientific papers. However, the hypothesis in our discussion was that there is a weak variant of citations by



creating links between books and, at higher granularity, between chapters and passages in the collection through common references to events, situations, ideas or through explicitly copied or cited passages. This gives rise to a mathematical graph from these links that is akin to a citation network. The analysis of such a graph might provide insights into which texts have been original or influential or both (tackling questions such as “is Magna Carta really the cornerstone of the idea of democracy?”)

*Methods from Natural Language Processing and Linguistics.* NLP and linguistics are the obvious areas from which methods can be derived that might help addressing above challenges. In particular, sentiment analysis is a branch that holds the promise of being able to classify documents as to their perception by readers, while, for example, the branch of forensic linguistics contains some of the ingredients for assigning authorship of documents through internal features or for extracting “fingerprints” of documents — it may even turn out that the analysis of stop word usage alone is a significant contributor.

*Visualisation.* As with all data mining problems and exploratory statistics, bespoke visualisation delivers important tools for understanding and detecting structures. It was recognised that time plays a vital role in many of the research challenges. We discussed several paradigms, such as a geological metaphor making time the depth of sediments in the earth crust, bespoke animations, a series of snapshots and timeline visualisations. Other ideas focussed on the notion of relevance and information content of document parts that led to the paradigm of islands, ie, relevant text components, emerging from a sea of irrelevance in reaction to a relevance slider that sets a desired threshold. Theme rivers were yet another paradigm touched upon.

*Methods from Machine Learning and Data Mining.* As with NLP and linguistics, machine learning and data mining are obvious areas that harbour methods to derive aspects of automated language and document understanding:

Any supervised methods such as classification of text passages need ground truth, which may be difficult to obtain and — once obtained — still poses barriers such as poor assessor agreements caused by genuine disparate opinions of the people involved. Crowd sourcing via “games for a purpose” may be a useful tool, in particular with document collections that are of general interest and are likely attract many users. Clustering as an unsupervised method lends itself to visualisation, for example, through Kohonen’s self-organising maps that aim to keep topological neighbourhood relations. Shopping basket analysis and association rules might provide some insight into relations of ideas and topics together with powerful visualisation methods that were developed for this.

#### 4.2.4 Summary

Concluding our discussion, after having reflected on it, there was a real sense that there is a great deal of knowledge that can be discovered from automated document processing. The extent and scale of many and large repositories is likely to support knowledge discovery through redundancy and many examples rather than posing an unwelcome challenge. Knowledge discovery in large text repositories (or on the web) is likely to become a thriving and intellectually challenging research direction.

And what to do when all computational methods have been exhausted for a particular task? As one of the participants remarked: “Well, there is always the option to read for oneself!”

### 4.3 Retrieval of Complex Structures

Rapporteur: Wolf Siberski

Text document retrieval has reached a high quality standard, and it has become part in many areas of our everyday life (e.g. Web search, library catalogues). In comparison, retrieval of complex structures (RCS) is still in its infancy. The topic has gained interest for several reasons:

- The Semantic Web or Web of Data gets momentum, with hundreds of structured data sources available already as part of the LinkedOpenData initiative
- With advances of entity recognition, entities and their relation can be extracted from documents and provided as graph structure
- A huge amount of structured data is created automatically (e.g. by sensor networks), especially as result of scientific experiments. For all these cases, facilities to search for specific information and/or to explore the complex information spaces are needed.

The goal of the working group was to discuss the nature of retrieval of complex structures (in particular its differences to traditional document retrieval), to discuss the state of research in the field, and to identify promising research directions for this challenge.

#### 4.3.1 Towards a Classification of Complex Structures in IR

The complexity of an information space can have different sources:

- Inherent complexity of the underlying structure (e.g., the database schema).
- Complexity arising from the content type (e.g., CAD models, movies).
- Complexity due to contextual dimensions which are part of the information (e.g., temporal or spatial context).
- Complexity due to the use of different language levels (object language, meta language) or due to the relation between information on different abstraction levels (e.g., the relation between extracted facts and to the documents on which these facts are based).
- Complexity arising from matching different levels of semantic to each other (e.g., IR has to match ambiguous user queries to a structured information space like databases or graphs).
- Complexity arising from integration of different information sources.

The working group identified different “complexity classes” for these complex information spaces, inspired by Chomsky’s formal language hierarchy. In the following, these classes are characterized by the expected or standard means that is necessary to satisfy an information need. The classes are organized from

1. The information need can be satisfied by a bag of words model or one of its variants. One can imagine the information space defined by a set of “flat” documents.
2. The information need can be satisfied by a graph-based data model. In particular, graph structures may be used to specify structured textual entities and their relations (usually represented in a relational schema or an ontology).
3. The information need deals with multimedia aspects of arbitrary kind, which includes also technical drawings.
4. The information need is not restricted, which means this class pertains to information spaces of unlimited structural complexity.

Notice that, in contrast to Chomsky’s model, these classes do not form a hierarchy. Among others, this is rooted in the fact that even highly complex information needs can be formulated as natural language documents, which in turn can be represented as bag of

words models; hence, encodings or transformations are conceivable that losslessly transform information needs across classes. Instead, the complexity classes here are oriented at data models, intended to measure the increasing modeling effort. The proposed class boundaries reflect “natural” modeling leaps as they were experienced by the group members.

### 4.3.2 Towards a Quantification of Complex Structures in IR

While the introduced “complexity classes” form a qualitative estimate for complex structures in IR, a quantitative estimate is yet missing. As shown by related fields like machine learning, quantifying complexity can yield new theoretical insights on algorithmic properties and achievable accuracy bounds. For example, the Vapnik-Chervonenkis dimension  $VC$  [3]—a complexity measure for the richness and expressiveness of hypothesis classes in Machine Learning—induced the development of generalization bounds for different types of classifiers. Similarly, topologies of metric spaces can be measured using different estimates, like for example by the Hausdorff-Besicovitch dimension.

According to the introduced “complexity classes”, example quantifications of complexity could be given as follows:

- Flat documents and their retrieval based on the Vector Space Model could be measured by means of the  $VC$  dimension for linear classifiers, with  $VC(H) = d + 1$  and  $d$  being the number of dimensions.
- Similarly, different  $VC$  bounds for graph structures exist. Such kind of bounds allow for deriving improved time bounds on shortest path queries [1], which points out the importance of quantifying information space complexity.
- Quantifying multimedia content and unlimited structural complexity remains an open issue for us. However, by considering retrieval as a topological problem, that is selecting a topological coherent region as answer for a query, topological measures like the Hausdorff-Besicovitch dimension may give possible estimates on information space complexity.

Obviously unsatisfactory is that the above classes and, consequently, their quantification, lack of a common scale or common measurement paradigm: given the short time available the working group could not identify a kind of language that underlies all forms of information needs. Perhaps such common basis does not exist at all (cf. the next subsection). Moreover, given accepted quantification measures for complex structures in IR, the question remains open as whether such a quantification yields general theoretical bounds on retrieval effectiveness and efficiency in practical applications.

### 4.3.3 Research Directions for Retrieval of Complex Structures

When humans use documents to communicate information, they use natural languages to express this information. Thus, while there is a variety of document flavors and domain-specific terminology, retrieval can still rely on the fundamental characteristics and regularities of human language. This explains why fairly generic techniques could be developed which can be applied successfully to a wide variety of document collections.

However, for complex structures this does not hold anymore: domain-specific multimedia content is frequently encoded in very specific ways which do not follow a generic language pattern (e.g., CAD models). But even textual information fragments in structured information don’t follow natural language patterns anymore, because the “information atoms” (such as a surname) are not connected in sentences, but live in isolated (database) fields.

Therefore, in the view of the working group participants, trying to create general-purpose retrieval systems for complex structures—in the same fashion in which we created general-

purpose document retrieval systems—seems to be a futile endeavor. But, although the complex structures itself are too heterogeneous to attempt a generalized retrieval approach, we observe that the information needs in various domains follow generic patterns. I.e., we can foster advances in this area by focusing on the process level instead of the system level. In this light, research can advance the field with three types of contributions:

1. by establishing well-defined processes for principled design of domain-specific complex structure retrieval systems,
2. by identifying and/or inventing reusable components (algorithms, patterns, frameworks) for the realization of such systems, and
3. by developing theoretical retrieval effectiveness and efficiency measures based on empirical complexity estimates.

A similar trend to adaptive retrieval strategies and customizable retrieval systems can be observed in the field of document retrieval (cf. Arjen’s talk).

#### 4.3.4 Principled Design of Retrieval Processes on Complex Structures

A methodology guiding the design of retrieval processes on complex structures should cover several aspects. A core aspect is finding suitable representations for the complex structures at hand. If the representation is too elaborated, the search process will become cumbersome for users. On the other hand, if the model chosen is too compressed, users not finding what they need might not be able to refine their search appropriately. This applies not only to the representation of the information in the collection, but also to the representation of the information need (the query). In addition to supporting the model design, the methodology should also guide retrieval system developers in splitting the process into subtasks and selecting the right components for these subtasks. The working group discussed the following already existing components.

*Faceted Search.* Faceted search is successfully used to enable users expressing constraints on their search. It was noted that while facets are good as filter criteria, their usage for influencing the ranking of results is limited. One reason might be that the mechanics of score aggregation is not transparent for the average user.

*Time/Space Representation.* Time and space are fundamental concepts in understanding and structuring our experience and can be well expressed and presented to users. In case of the temporal dimension, a time line is the most natural visualization, while for spatial information the obvious visualization is a map. On the query side, constraints are expressed as time span and location-distance pair respectively.

*Top-k and Skyline Queries.* Top- $k$  queries are suitable for structured data for which good relevance functions have been identified. The usage of Skyline queries is appropriate for low (2-5) dimensional data, when scoring functions for the individual data dimensions are available. However, they cannot be combined into a single relevance function.

*Keyword queries on Structured Data.* A lot of work has been done in the last years to extend the keyword search paradigm to structured data. Summarized, all these approaches view the structured data as graph and try to find small subgraphs (i.e., join paths) containing the query terms. Then they assess their relevance, and present the most relevant results to the user. While a lot of progress has been made, the retrieval quality level is still not very impressive [2]. Another direction in this area is the idea to infer the query intention in a multi-step retrieval process before assessing the query results (cf. Wolf’s talk).

*User-defined Search Interfaces.* In accordance with the core idea to support retrieval process design rather than to build generic retrieval systems for complex structures, frameworks for

retrieval have been built which allow the expert users (or system administrators) to specify their customized retrieval process (cf. Andreas' and Arjen's talks). These frameworks glue together individual retrieval components, and hence form a foundation for the transition of a retrieval process definition to a working retrieval system.

## 4.4 Sentiment and Opinion

Rapporteur: John A. Carroll

### 4.4.1 Tasks

Opinion mining can be viewed as 'natural' language-based task, as being something that a person might do in everyday life: finding out about what others think about a product, service, brand, organisation, or the views and actions of others. The information may come from a number of different types of document, such as online product reviews, newspaper editorials, blog posts, etc.

There are a number of opinion mining tasks, including:

- Subjectivity classification – whether a sentence, paragraph or whole document is expressing some sort of opinion
- Polarity (or sentiment) classification – whether an opinion being expressed is broadly positive or negative, or positive / neutral (i.e. non-opinionated) / negative; other more fine-grained classification schemes have also been used
- Relevance – whether a particular document is relevant to a given topic of interest
- Authority – the degree of influence of the opinions expressed in a document on its readers
- Opinion holder and target identification – who is expressing an opinion, and about what
- Sentiment towards features of a product or aspects of a topic – for example whether the picture quality, price, and battery life of a camera are good or bad; or whether government policy on taxation is good or bad
- Detecting key messages related to a brand or marketing campaign – reflecting the impact that it has had
- Classification of documents along other affective dimensions – such as anger, frustration, or satisfaction
- Other opinion-related tasks – for example rating the perceived 'green-ness' of politicians, or the proportion of people in favour of building new nuclear plants

In the past many of these tasks were carried out by human analysts, originally working from printed media; with the increasing amounts of online text and capabilities of language processing systems, automatic opinion mining has become a highly active area of basic and applied research.

### 4.4.2 Barriers to Progress

*Data Quality.* Most research in sentiment classification has used data consisting of reviews of products or services scraped from review websites, for example IMDb (the Internet Movie Database), Amazon, or epinions. When review authors give 'star' ratings to accompany their reviews, this can result in large amounts of data annotated at the document level.

In contrast, there is very much less text available that is annotated for sentiment at more detailed levels (sentence or phrase), or annotated to support other opinion mining tasks.

Notable exceptions include the MPQA Opinion Corpus and the datasets produced for the NTCIR MOAT workshops. However, it is difficult to satisfactorily circumscribe some opinion mining tasks: for example in opinion holder and target identification, should holder and target phrases include post-modifiers, and if so what types of post-modification? Without well-defined annotation guidelines, inter-annotator agreement will be poor leading to a low upper bound on system performance.

Another approach to obtaining annotated data is crowd-sourcing, for example via the Amazon Mechanical Turk. The quality of annotations obtained this way can be questionable, but may be assured by framing the task carefully, for instance requiring that a quote from the document be pasted into a text box to justify the sentiment annotation being entered.

*Evaluation.* Funding is hard to obtain for long-term evaluation campaigns or shared tasks. There are only two of these worldwide which contain an opinion mining component: TREC (blog track) and NTCIR (multilingual opinion analysis task), organised by NIST and NII respectively. Other evaluation exercise series that have had opinion-related components include CLEF, SemEval, and i2b2. These exercises often rely for annotation on ad-hoc groups of volunteers – usually the participants themselves. Short timescales and lack of funding make it difficult to generate significant amounts of high quality data, limiting the advances that can be made.

An important methodological issue (not restricted to this research area) is reproducibility of experiments. Datasets and systems are not always publicly available, and algorithms and parameter settings are often not described in sufficient detail to be able to reproduce results. This may make it difficult or even impossible for subsequent researchers to conduct appropriate comparative evaluations, hindering progress.

#### 4.4.3 Research Priorities

*Cross-domain.* Accuracy of sentiment classification can be severely degraded if a system trained on data from one domain is applied to a distinct domain. However, there is often insufficient annotated data in a new domain to adequately train a supervised machine learning algorithm. Domain adaptation techniques, such as structural correspondence learning (e.g. work by Blitzer) and related approaches are a promising way of tackling this important issue. These techniques use annotated data in a source domain in conjunction with unannotated data in the target domain (and possibly also a further small amount of annotated data in the target domain).

*Cross-language.* Compared with most languages, there is a large amount of data in English that could be used for developing opinion mining systems. Cross-language systems take advantage of this by mapping information from a source language (e.g. English) to the target language. The mapping can be done with techniques similar to those used for domain adaptation; however this is a difficult problem since different cultures may express opinions using different types of language.

*Novel sources of data.* It is relatively easy to obtain suitable annotated data for sentiment classification of product reviews (e.g. from review sites' star ratings), but it is in general difficult for other opinion mining tasks. Finding good sources of annotations avoids the need for costly annotation efforts. Examples of such sources are: metadata for authority (for example '10 out of 15 people found this useful'), and summaries of key points abstracted from longer reviews in special interest magazines (such as about cars or yachts).

*Emergent features.* Given some knowledge of a type of product or service it is possible to come up with a list of features that reviewers might have an opinion on, and develop a system

to find opinions on these; however, it is also important to be able to capture ‘emergent’ features that people have an opinion on but which hadn’t been anticipated.

*Presentation of results.* Traditionally, human analysts in media monitoring companies write reports for clients which contain a range of quantitative and qualitative information presented as text, tables and graphics. An automatic system could not currently produce such reports. The question then is how to present the results of opinion mining. Some possibilities are: overall aggregate sentiment, a selection of typical documents, a set of features and the sentiment ascribed to each, opinion associated with sub-topics in the domain, or opinion-driven summarisation. A further question is to do with explanatory level: why does an individual perceive something as being good or bad? This relates to sentiment-specific aspects of entailment.

*More detailed analysis.* Media monitoring companies have traditionally used a 5-category sentiment classification scheme (strongly / weakly negative, neutral, and strongly / weakly positive), rather than the 2- or 3-category scheme usually adopted in computational work. More fine-grained analysis requires deeper representations than the typical ‘bag of words’ and deeper processing than supervised classification over these representations. Such analysis might involve (at least) parsing, integration of contextual valence, and principled treatments of negation and hedging.

#### 4.4.4 Summary

Opinion mining has many different facets and presents numerous difficulties, as outlined above. A lot of progress has been made, but as in most areas of document mining, there are still many outstanding research challenges.

#### References

- 1 Ittai Abraham, Daniel Delling, Amos Fiat, Andrew V. Goldberg, and Renato F. Werneck, *VC-Dimension and Shortest Path Algorithms*, In Proceedings of ICALP 2011.
- 2 J. Coffman and A. C. Weaver, *A framework for evaluating database keyword search strategies*, In Proceedings of CIKM 2010.
- 3 V. Vapnik and A. Chervonenkis. *On the uniform convergence of relative frequencies of events to their probabilities*. Theory of Probability and its Applications, 16(2):264-280, 1971.

## 5 A Theme: Towards More Open Search In Europe

The European Union is investing more than 5 billion euros in Galileo, a global navigation system similar to the US-owned GPS. One of the motivations for this investment is the potential impact of a GPS shutdown (perhaps in time of crisis or a period of intense trade disputes or similar). Far-fetched, perhaps, but then the disappearance of hundreds of billions of euros into a black hole of financial chaos appeared far-fetched until recently.

An obvious corollary question is this:

What would be the economic cost to the EU of a shutdown of the search engines provided by Google, Yahoo and Microsoft?

As the World Wide Web has become a key repository of human knowledge, interaction and commerce, so the ability to discover relevant pages has become a core function of our economies. It is now routine to describe 21st century industry as a *knowledge economy*, but what price knowledge if it is impossible to find? How many of the web addresses that you use (those ‘uniform resource locators’ that start with `http://`) can you remember?

Let's imagine a sudden restriction of service from the big three search providers (effectively now a 'big two' after the deal between Yahoo and Microsoft). Technologically speaking removing service availability based on region can be done trivially – witness the current mechanisms to control access to BBC programmes abroad, for example.

The economic consequences for the EU would be immediate and wide-reaching. Our elected representatives and state bureaucrats would be tasked with finding a replacement – but their first port of call would be the search box in their web browser...

Secondly, there is increasing concern that detailed individual and private data is becoming more and more centralised and open to abuse. Let's imagine that we suffer a number of scandalous cases at some future point, and the legal framework changes to the degree that the current models no longer work. At present the EU would be powerless to implement the requisite technological changes.

We are not the first to recognise this achilles heal in European independence. The Quaero initiative announced by Jacques Chirac and Gerhard Schröder in 2005 aimed to fill the gap, but by 2007 *The Economist* reported that the project had effectively been scrapped (for various reasons – but the huge cost of the compute infrastructure and bandwidth consumption required must be a large factor). Several results have survived: multimedia retrieval projects in Germany and the Exalead engine in France, but the dream of triggering construction of a pan-European search engine has remained a dream.

What to do?

Resurrecting the vision of Quaero and a full-blown European competitor for Google and Microsoft/Yahoo is a matter for forces more powerful than those we represent or can hope to influence. However, all is not lost, partly because of the changing landscape of the web, of utility computing, and of social interaction in the networked age. If Europe can find the resources to fund a set of incremental improvements in our search R&D then we can take a major step towards greater openness and independence from the dominant corporate US infrastructures.

We propose these measures:

1. Support the use of cloud computing to develop national search service platforms within existing information retrieval research computing centers. Long-term commitment is more important than a dramatic expansion of funding; the current situation is short-termist and discourages robust search provision.
2. Create new marketplaces for value-added services on top of the cloud infrastructure. In particular the need for analysis of social media is becoming very widespread (for example: customer relations used to be about sitting next to the phone or the cash till; now its about reading 250,000 Twitter posts per week).
3. Develop aggregation and syndication models of distributed clustering and indexing to allow the composition of international and cross-domain search services, drawing on the above provisions. (For example recent work on peer-to-peer indexing and clustering in web servers like Apache.)

The cost of an alternative to Google is probably too large a mouthful to swallow in one go, but a smaller set of chunks can bring us a solution in the short to medium term.

## 5.1 Discussion

- issues with 3. (P2P): spam; speed (necessitates regional/national aggregation)
- grey literature, intranets: organisations can create custom search views that incorporate their own private data



- richer model of privacy, the ability to buy privacy (currently we sell privacy – we take free services in return for giving up some privacy – but have no alternative if we want to buy privacy)
- what about the energy efficiency of running the big 4 web companies (Google, M\$/Yahoo, Facebook, Twitter) in a single country?

## 6 Recap of the Proposal

Document mining is the process of deriving high-quality information from large collections of documents like news feeds, databases, or the Web. Document mining tasks include cluster analysis, classification, generation of taxonomies, information extraction, trend identification, sentiment analysis, and the like. Although some of these tasks have a long research history, it is clear that the potential of document mining is far from full realisation.

Part of the problem is that relevant document mining techniques are often applied in an isolated manner, addressing – from a user perspective – only a part of a task. E.g., an intelligent cluster analysis needs adequate document models (information retrieval) that are combined with sensible merging algorithms (unsupervised learning), complemented by an intuitive labelling (information extraction, natural language processing). The deficit that we observe may also be understood as a lack of application- and user-orientation in research.

In this seminar we want analyze the untapped potential. To this end we bring together researchers from the main areas of document mining to present their view, to understand where and how latest disciplinary achievements can be combined, and to develop a more integrative view on document mining.

The seminar is especially timely as in recent years the field has grown by a large amount, and this has lead to increasing specialisation of the research community. Now is the time to bring a sample of the leading teams back together and look at the problem from a multidisciplinary point of view

The seminar will focus on the following aspects of document mining:

1. What are the relevant document mining tasks? The expectations and the potential for document mining changed significantly over time. Influential in this connection is the discovery of the enormous contributions of users to the Web (among others in the form of blogs, comments, reviews) as highly valuable information source.
2. Cluster analysis is a key technology in document mining and involves several issues on its own, which are detailed below. A major deal of cluster analysis research has been spent to merging principles and algorithms; today, and especially in document mining, the research focus is on tailored document models, user integration, topic identification and cluster labelling, on the combination with retrieval technology (e.g. as result set clustering), or on support technology for supervised classification. Moreover, theoretical foundations of cluster analysis performance in document mining as well as commonly accepted optimality measures are open research questions.
3. While cluster analysis can be mainly considered as unsupervised, several advanced document mining tasks combine unsupervised with supervised text classification. Especially non-topical classification tasks attracted interest in this connection, such as genre classification, sentiment analysis, or authorship grouping.
4. Document mining poses various challenges from a machine learning perspective. An important constraint is the lack of sufficient amounts of labelled data. This situation will become even more unbalanced in the future, and current research (domain transfer

- learning, transductive learning) aims at the development of technology to exploit the huge amount of unlabelled data to improve supervised classification.
5. Robustness and efficiency of document mining technology is a key issue from the user perspective and for future applications. Both may be achieved by the combination of algorithms (ensemble clustering) or the combination of different retrieval models; the respective research is still at its infancy.
  6. The use of NLP in document mining is a success factor of increasing importance for document mining. Among others this field contributes technology for document modelling, style quantification, document segmentation, topic identification, and various information extraction and semantic annotation tasks.
  7. The assessment of information quality and credibility will have a large impact on future document mining solutions. It can tackle information need issues and performance problems in our information-flooded society at the same time. However, there are various open research question related to the measurability of information quality.
  8. Authorship and writing style modelling is still coming of age; this area forms the heart for high-level document mining tasks such as plagiarism analysis, authorship attribution, and information quality assessment.
  9. Future technology for entity resolution and fact or relation extraction. Current approaches are limited to closed environments where the target entities and relations are known in advance. In practice, however, this assumption is often violated or the number of entities and relations is so large that automatic methods are needed. The Open Information Extraction paradigm, coined by Banko and Etzioni, addresses these issues. Its goal is to extract a diverse set of relational tuples from text without any relation-specific input such as hand-labeled examples or hand-crafted lexico-syntactic patterns.
  10. Interface design and visualization are very important for effective user access to the output of the document mining process. Moreover, interactive document mining approaches like e.g. scatter-gather clustering pose new challenges for both the interface and the backend.
  11. Finally, evaluation is essential for developing any kind of data mining method. So far, mainly system-oriented evaluation approaches have been used, where the data mining output is compared to some ‘gold standard’. There is a lack of user-oriented evaluations (e.g. observing users browsing a cluster hierarchy), that also take into account the tasks the users want to perform – e.g. using Borlund’s concept of simulated work tasks.

The general idea is to collect the state of the art in document mining research, and to define a research agenda for further work in this area. For this purpose, we want to bring together experts from the different research areas listed above. Each of the participants first should present her/his view on particular document mining challenges by highlighting latest results and naming crucial research issues. Based on these contributions, we will aim at developing an integrated view on the problem of document mining, identify open research problems and then point out steps towards resolving these problems. Altogether we would like to come up with a framework that associates document mining tasks with the scientific and technological elements of their adequate solution.

## **7** Acknowledgements

Our thanks to the staff of Schloss Dagstuhl for excellent organisation and facilities.

## Participants

- Leif Azzopardi  
University of Glasgow, GB
- Ted Briscoe  
University of Cambridge, GB
- Steven Burrows  
Bauhaus-Universität Weimar, DE
- John A. Carroll  
Univ. of Sussex – Brighton, GB
- Massimiliano Ciaramita  
Google Switzerland – Zürich, CH
- Hamish Cunningham  
Sheffield University, GB
- Arjen P. de Vries  
CWI – Amsterdam, NL
- Norbert Fuhr  
Universität Duisburg-Essen, DE
- Tim Gollub  
Bauhaus-Universität Weimar, DE
- Thomas Gottron  
Universität Koblenz-Landau, DE
- Michael Granitzer  
Know-Center Graz, AT
- Andreas Henrich  
Universität Bamberg, DE
- Gerhard Heyer  
Universität Leipzig, DE
- Dennis Hoppe  
Bauhaus-Universität Weimar, DE
- Melikka Khosh Niat  
Universität Duisburg-Essen, DE
- Marc Lechtenfeld  
Universität Duisburg-Essen, DE
- Alexander Löser  
TU Berlin, DE
- Peter Prettenhofer  
TU Graz, AT
- Andreas Rauber  
TU Wien, AT
- Harald Reiterer  
Universität Konstanz, DE
- Stefan M. Rieger  
The Open University – Milton Keynes, GB
- Hinrich Schütze  
Universität Stuttgart, DE
- Wolf Siberski  
Leibniz Univ. Hannover, DE
- Benno Stein  
Bauhaus-Universität Weimar, DE

