Projects and Theses
Information Science Group
www.isg.uni.kn
Content

1. Organizational Information

2. Topics for your Projects and Theses
Concept of the Courses at the Computer Science Department

1. Lecture
   - Spark interest in computer and information science problems
   - Give overview of technologies
   - Teach essential technological skills

2. Seminars
   - Acquire in-depth knowledge of current research
   - Train methodological skills
   - Prepare potential BA/MA project & thesis

3. Project & Thesis
   - Make a small, but meaningful and lasting contribution to the state-of-the-art
Concept of the Courses at the Computer Science Department

1. Seminar (1 Semester)
   – Perform literature review
   – Identify research gaps
   – Develop project ideas

2. BA/MA Project (1 Semester)
   – Develop and implement own approach
   – Implement infrastructure to evaluate existing systems

3. Thesis (1 Semester)
   – Evaluate your or third-party systems
   – Write thesis, explain improvements

In Reality...
General Areas for Projects and Theses

- Information and Web Retrieval
  - Wikipedia
  - Mathematics
  - News
- Natural Language Processing
- Cryptocurrencies (Bitcoin)
- Digital Libraries & Recommender Systems
- Knowledge Management
- Information Visualization
- Web Engineering

If you are interested in doing a project / theses related to these areas, lets talk!

isg.uni.kn/people
(to schedule meetings)

isg.uni.kn/students-corner
(for topic proposals)

isg.uni.kn/wiki
(for guidelines, tutorials, and templates)
Media Coverage of our Research

Click on the respective images to read the full articles.
Possibility to do master project or thesis in Tokyo

- Cooperation between ISG and two research groups at the National Institute of Informatics Tokyo (Japan’s premiere research institute for all areas of CS, approx. 200 professors)

- **Topics** in Natural Language Processing & Information Retrieval (specific project collaboratively decided by supervisors at the Univ. of Konstanz and at the NII)

- **Duration**: 2-6 months

- **Compensation**: approx. 1400€/month (sufficient to cover all living expenses)

More Information: [isg.uni.kn/japan](http://isg.uni.kn/japan)
TOPICS FOR
PROJECTS & THESES
PLAGIARISM DETECTION
Analyzing Patterns of Matching Text to Detect Disguised Plagiarism

Paraphrasing text by replacing content-bearing words with synonyms and altering the order of sentences is a common strategy to disguise plagiarism. Current plagiarism detection systems search for literally matching strings and if the percentage of literal matches surpasses a certain limit, no suspicion is raised. We hypothesize that analyzing patterns of matching text is a better approach to discovering also disguised instances of plagiarism than exclusively considering the mere quantity of matching text.

Goal

Develop a method that analyses patterns of matching text in documents to identify potential instances of plagiarism and integrate them into CitePlag (www.citeplag.org).

Tasks

• Review state-of-the-art text-based PD and pattern analysis methods.
• Develop a pattern analysis and visualization approach for text matches.
• Integrate your analysis approach in CitePlag.
• Evaluate the effectiveness and efficiency of your approach.

Contact

Norman Meuschke
norman.meuschke@uni.kn

Moritz Schubotz
moritz.schubotz@uni.kn

Bela Gipp
bela.gipp@uni.kn
Evaluate and Improve Strategies to Reduce the Retrieval Space for Plagiarism Detection

Realizing a practically feasible plagiarism detection process requires computationally modest methods to identify in a large collection the set of candidate documents that have enough similarity with an input document to justify a closer, computationally more expensive inspection. The key to accomplishing this task is to compare features, e.g., matching text, images, academic citations, mathematics, and combinations of those features that are distinctive enough to effectively limit the collection, yet do not eliminate unacceptably many potential source documents. The later requirement is especially challenging for disguised forms of plagiarism, such as strong paraphrases or translations.

Goal
Evaluate the effectiveness and efficiency of using different feature sets, e.g., matching words, citations, images, and mathematical formulae to reduce a large document collection to a set of potential candidate documents.

Tasks
• Familiarize yourself with the state-of-the-art heuristics retrieval approaches for plagiarism detection.
• Realize the evaluation framework, i.e., setup a test collection, and comparison techniques. (We provide retrieval approaches for text, citations, images, and mathematical expressions)
• Use the framework to conduct a performance evaluation for different feature sets and forms of plagiarism.
• Analyze your results and derive suggestions as to the best performing approach.

Contact
Norman Meuschke
norman.meuschke@uni.kn

Moritz Schubotz
moritz.schubotz@uni.kn

Bela Gipp
bela.gipp@uni.kn
Literature Recommendation
Comparing Semantic Fingerprints

People often work on the same, similar, or overlapping ideas – without knowing it! As a result, time and resources are wasted on redundant work. Teams could collaborate or exchange information, if they knew they are working on similar problems.

In this project, you will develop methods to compute and compare unique ‘fingerprints’ that represent the concepts a user is currently working on from work-in-progress files. The fingerprints shall allow researchers to share the essence of their work, without needing to make their ideas public just to find groups working on similar concepts.

Goal

Analyze features, e.g. keywords, citations, formulas, document structure, bibliography, that should be used to represent the essence of an idea or work-in-progress. Use suitable methods, e.g. NLP, to extract features and devise similarity measures to identify similar ideas.

Tasks

• Develop a method to extract the “essence” of an idea – in form of a fingerprint.
• Develop a system that:
  • Extracts fingerprints locally
  • Timestamps the fingerprints on the blockchain using the OriginStamp API
  • Computes the similarity of different fingerprints

Contact

Corinna Breitinger
corinna.breitinger@uni.kn

Felix Hamborg
felix.hamborg@uni.kn

Bela Gipp
bela.gipp@uni.kn
Developing a Privacy-preserving Recommendation System

In order to receive relevant recommendations, for example, recommendations for literature that will be of interest to a user, users must first share their private data and behavior openly with the recommender system. This requires a lot of trust in the recommendation provider. Users must trust that the provider will not leak or sell confidential data, e.g. user profile data.

Today’s recommendation systems typically do not support user’s privacy, in that they collect and store their users’ data server-side and unencrypted.

In this research-focused project, you will develop algorithmic and organizational approaches that support a privacy-preserving assessment of user’s data for similarity computation.

Goal
Survey existing approaches that enable a privacy-preserving assessment of the similarity between data. Apply suitable approaches to realize a research-in-progress recommendation system.

Tasks
• Review the literature, e.g., from security research, and explain the most common algorithmic approaches, e.g. similarity digests, fuzzy hashing, locality sensitive hashing, etc.
• Determine which approaches are most suitable for which system contexts / needs
• Apply suitable approaches to the task of recommending in-progress-research

Contact
Corinna Breitinger
corinna.breitinger@uni.kn

Norman Meuschke
norman.meuschke@uni.kn

Bela Gipp
bela.gipp@uni.kn
Developing a Hybrid Link-based Similarity Measure for Literature Recommendation

Link-based similarity measures such as PageRank, HITS, SimRank, Co-Citation and Co-Citation Proximity Analysis have been shown to be immensely valuable for numerous information retrieval and data analysis tasks such as ranking web pages, clustering documents by topic or recommending literature.

Individual similarity measures follow different approaches, e.g. PageRank uses a random surfer model while Co-Citation Proximity Analysis considers proximity of links in the linking document. In consequence measures have different strengths and weaknesses. We have some ideas how to combine the strength of individual measures, which we like to explore together with motivated students.

Goal

Analyze the strengths and weaknesses of individual link-based measures, propose and evaluate a novel measure.

Tasks

- Analyze the state-of-the-art with regard to link-based similarity measures.
- Combine the strengths of existing measures to develop your own measure.
- Evaluate the effectiveness and efficiency of your approach.

Contact

Corinna Breitinger
corinna.breitinger@uni.kn

Norman Meuschke
norman.meuschke@uni.kn

Bela Gipp
bela.gipp@uni.kn
Analyzing Co-Citation Sentiment to Improve Literature Recommendation

Academic citations provide valuable information about the relations between documents and the research those documents represent. Considering citation relations and especially co-citation relations, i.e. when sources are cited together in the same document, has proven successful for generating academic literature recommendations.

We hypothesize that analyzing the sentiment expressions within text segments that co-cite sources in close proximity can furthermore improve the results obtainable from a sentiment analysis of individual citations or the analysis of textual proximity between co-cited sources.

Goal

Explore how analyzing co-citation sentiment, i.e. the attitude and intention authors express when co-citing sources in academic documents, can enhance an academic literature recommender system. Devise and evaluate a literature recommendation method that considers co-citation sentiment.

Tasks

• Analyze the state-of-the-art in citation sentiment analysis.
• Develop a sentiment analysis approach for co-cited sources.
• Find or create a suitable test collection.
• Evaluate the effectiveness and efficiency of your approach.

Contact

Corinna Breitinger
corinna.breitinger@uni.kn

Norman Meuschke
norman.meuschke@uni.kn

Bela Gipp
bela.gipp@uni.kn
APPLICATION DEVELOPMENT USING CRYPTOCURRENCIES
Realize a Novel Blockchain-based Application

Cryptocurrencies, such as Bitcoin, are in essence peer-to-peer networks that use permissionless, distributed data structures, the ‘blockchain’ to exchange data. Because digital information can easily be embedded into cryptocurrency transactions, blockchains serve as a novel data structure for a variety of applications aside from currency exchange.

Cryptocurrency networks and their underlying blockchains enable tamper-proof, anonymous, decentralized, and verifiably dateable exchange, as well as persistent storage of information with virtually no chance of interference.

Goal

Use the unique characteristics of the blockchain data structure to develop a novel application which solves a practical problem.

Tasks

- Identify a suitable use case.
- Design and develop a blockchain-based application beneficial to your use case.
- Demonstrate the benefit that your application provides over existing solutions.

Contact

Thomas Hepp
Thomas.hepp@uni.kn

Corinna Breitinger
corinna.breitinger@uni.kn

Bela Gipp
bela.gipp@uni.kn
NEWS ANALYSIS

- Russian troops cross border into Ukraine [NY Times, US]
- Ukraine at the brink of war with Russia [Fakt, PL]
- Kiev, stop making false invasion claims [RT, RUS]
What’s new in news?

*Explicit Semantic Analysis*

Most news articles reuse information from other sources, e.g., from press agencies or news outlets that already reported on the topic. While text matching approaches reliably identify copied text, e.g., if a paragraph was taken literally from a press release, simple text matching approaches fail to recognize more complex forms of text reuse, e.g., if a paragraph summarizes another article.

The goal of this project is to analyze semantic features to identify text passages that originate from other articles either literally or in meaning.

**Goals**

- Enhance and adapt explicit semantic analysis (ESA)
- Develop a visualization to let users explore and understand the findings

**Tasks**

- Review state-of-the-art methods (focus: ESA)
- Develop an ESA method for information origin analysis in news articles (focus: paragraphs, sentences, catch-phrases, …)
- Visualize the results
- Evaluate the effectiveness and efficiency of your approach.

---

**Contact**

Felix Hamborg  
felix.hamborg@uni.kn

Bela Gipp  
bela.gipp@uni.kn
**What’s new in news? Concept Hierarchies**

This project analyzes the hierarchical structure inherent to news and many other text documents: words make up sentences, which make up paragraphs, which make up articles, which—if related—make up the topic or event. The research question is how to extract this tree from a set of related news articles? The motivation of this project is to identify the similarities and differences of two articles, by comparing their trees.

**Goals**
- Develop a method to extract concepts (e.g., words or synsets) from the various levels (sentences, paragraphs, articles, ...)
- Develop a method to compare two or more trees (what are their differences?)
- Visualize the concept hierarchy

**Tasks**
- Review state-of-the-art methods for concepts and tree representations of texts.
- Research how to represent articles in concept hierarchies
- Develop methods to extract concepts from articles (incl. aggregation stats)
- Compare trees to find similarities and differences in news
- Evaluate the effectiveness and efficiency of your approach.

**Tasks**
- Analysis:
- Implementation:
- Literature:

**Contact**
Felix Hamborg
felix.hamborg@uni.kn

Bela Gipp
bela.gipp@uni.kn
The relatedness of places

Geographic, cultural, financial, etc.

Measuring the geographic distance between two places is trivial. But what about the cultural, financial or economic “distance” between two countries or other entities? And how do the distances change over time?

We developed a tool that extracts co-occurrences of entities from news articles. The goal of this project is to devise and implement suitable, normalized distance measures and create a model for a corpus that reflects the extracted relatedness of the entities (similar to WordNet). We will provide you with the necessary tools to create the corpus and to extract the co-occurrences.

Goals

• Devise and implement normalized distance measures for different dimensions
• Evaluate the devised methods

Tasks

• Review state-of-the-art methods and the NewsRelation tool (provided)
• Investigate the different types of relations between entities in news
• Devise & implement normalized distance measures between entities
• Optional: Implement a software library to be used by other researchers (similar to WordNet)
• Evaluate the performance of your approach

Contact

Felix Hamborg
felix.hamborg@uni.kn

Bela Gipp
bela.gipp@uni.kn
Human perception of named entities in news articles

<table>
<thead>
<tr>
<th>Publisher</th>
<th>Excerpt</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York Times</td>
<td>Iraqi fighter jets threatened two American U-2 surveillance planes, forcing them to abort their mission and to return.</td>
</tr>
<tr>
<td>USA Today</td>
<td>U.N. arms inspectors said they had withdrawn two U-2 reconnaissance planes over Iraq for safety reasons.</td>
</tr>
</tbody>
</table>

Journalists may use different words to refer to the same named entity, such as a politician. The goal of this project is to estimate the human perception of each entity in a single sentence of a text using the psychometric lexicon LIWC. LIWC is a dictionary that contains over 60 psychometric dimensions for millions of words. Using LIWC will help to estimate how readers perceive each of the named entities.

Goals

- Devise and implement a method to extract psychometric properties of named entities
- Evaluate the devised methods

Tasks

- Review state-of-the-art methods (only briefly, focus can be simply on LIWC)
- Devise & implement a method to extract psychometric properties of named entities on sentence-level
- Devise & implement a method to aggregate the psychometrics per named entity on paragraphs and documents
- Evaluate the performance of your approach.

Contact

Felix Hamborg
felix.hamborg@uni.kn

Bela Gipp
bela.gipp@uni.kn
WIKIPEDIA-RELATED PROJECTS
Seeding WikiNews with Texts from news-please

WikiNews is a platform where volunteers collaboratively write news articles. The platform aims to provide readers with objective, non-biased news coverage. Unfortunately, not many people contribute and write articles, so that many current events are not covered by WikiNews. The idea of this topic is to support people in producing news content by providing them with suggestions for topics and possible information sources. Therefore, the task is to create a user interface to write WikiNews articles based on information that has been extracted previously by news-please, a crawler and information extractor for news.

Goal
Develop a user interface that integrates news-please into the WikiNews platform.

Tasks
1. State of the art review
3. Evaluation

A successful project contribution has a good chance of going live at Wikipedia and improving the encyclopedia for millions of users.
Mathematical Language Processing for Wikipedia

Mathematical formulae are essential in science, but are highly ambiguous, since a small number of identifiers represents an immense number of concepts. Mathematical Language Processing (MLP) analyzes the natural language text around formulae to semantically enrich the formulae. MLP can for example serve to disambiguate identifiers, e.g. determine whether \( e \) denotes electron charge or the base of the natural logarithm, recognize synonymous identifiers such as \( \sigma \) and SD, and to link identifiers and formulae to natural language keywords and semantic concepts to facilitate retrieval by humans, e.g. recognizing \( F = G \frac{m_1 m_2}{d^2} \) as Newton’s universal law of gravity, hence allow searching using these terms.

Goal
Develop and employ MLP methods to improve the retrieval of and interaction with mathematical formulae in Wikipedia.

Tasks
Depending on your background and interests, your project could address different stages in the MLP process such as improving the collection of semantic information for formulae or improving the storage and retrieval of semantic information about formulae via Wikidata.

A successful project contribution has a good chance of going live at Wikipedia and improving the encyclopedia for millions of users.

Contact
Moritz Schubotz
moritz.schubotz@uni.kn

Norman Meuschke
norman.meuschke@uni.kn

Bela Gipp
bela.gipp@uni.kn

2. Visualizing Author Impact in Wikipedia

In academic publishing, quantitative measures like the Hirsch index for authors, or impact factor for journals are routinely applied to approximate the quality and relative importance of academic contributions. In Wikipedia, these measures cannot reflect the impact, i.e. the quality and relative importance, of author contributions, because Wikipedia articles and their set of authors are variable and edits are constantly revised or reverted.

This topic subsumes two projects.
2. Visualizing author impact in Wikipedia (BA)

Goals
1. Define and compute (BA/MA) [...] 
2. Visualize (BA) [...] 

[...] a measure to quantify the impact, i.e. the extent, quality, and relative importance, of Wikipedia contributions by individual authors.

Tasks
- 1. & 2.: Survey related work on measuring academic contributions.
- 1.: Define an impact measure for Wikipedia authors using the edit history.
- 1.: Calculate the impact measure for all Wikipedia authors using Apache Flink.
- 2.: Develop suitable visualizations of an author’s impact in Wikipedia.
- 2.: Implement the visualizations as a MediaWiki extension and plugins or code snippets usable in popular Content Management Systems and Websites.
- 1. & 2.: Evaluate the newly developed approach.

Contact
Moritz Schubotz
moritz.schubotz@uni.kn

Norman Meuschke
norman.meuschke@uni.kn

Bela Gipp
bela.gipp@uni.kn
Evaluating Linked Based Recommendations for Wikipedia

The Android app for Wikipedia suggests related articles by combining two approaches. Content based filtering and collaborative filtering. For the content based filtering text based methods were used in the past. However, recently a new link based similarity approach has been proposed. Different offline evaluations of this new approach show promising results.

Goal
Analyze differences between text based and link based recommendations and their impact on Wikipedia users.

Task
1. Summarize text based (e.g. Apache MLT) and link based (Citolytics) similarity measures
2. Derive conceptual differences from the literature
3. Derive hypothesis
4. Design user survey
5. Perform survey
6. Confirm or reject hypothesis

Moritz Schubotz
moritz.schubotz@uni.kn

Norman Meuschke
norman.meuschke@uni.kn

Bela Gipp
bela.gipp@uni.kn
User Modelling based on the Wikipedia Edit History

User modelling is an important research area, for Recommender Systems, Expert Systems, Simulations and more applications. However, to test and evaluate user models large datasets of user data is required. While huge internet companies like Google, Facebook and Twitter have access to suitable data sources, researches as typically do not have access to adequate data sets. To alleviate this problem, we propose to use Wikipedia’s edit history, which is open to the public, as dataset to improve user models.

Goal
Predict human behavior based on Wikipedia's edit history and evaluate the prediction quality in a user study

Task
1. Summarize design patterns for user models
2. Derive hypothesis for user behavior based on the edit history
3. Design user survey
4. Perform survey
5. Confirm or reject hypothesis
Runtime Debugging of Distributed Data Processing Programs

Debugging massively parallel data analysis programs is currently a difficult process. Traditional debug cycles involve manual code instrumentations, re-execution and analysis of the resulting data. This is expensive in terms of development time, execution time, amount of data produced, and cognitive overhead. This work proposes a course of research that is meant to alleviate this situation by automating the code instrumentation and by lowering the re-execution time of instrumented code. By using these techniques, we hope to achieve a higher efficiency compared to manual debugging approaches.

Goal

Develop a Java based debugging system for remotely distributed data processing programs.

Task

1. Analyze the state-of-the-art of existing debugging systems.
2. Extend existing backend structure.
3. Design and implement an IntelliJ IDEA plugin to control the system.
4. Evaluate the newly developed system on a large Wikipedia dataset.
INFORMATION EXTRACTION
MathML Conversion of Mathematical Content in PDFs

Adobe PDF is the most widely used document format for scientific publications. While offering excellent visual presentation, the format poses challenges to automated content extraction and processing, which is essential to many information retrieval tasks.

Mathematical formulae contain a high degree of semantic information, which the portable document format does not provide. The XML-based open standard MathML is significantly better suited to store, transmit and access both the presentational and semantic information contained in mathematical content.

Goal
Develop methods and tools to recognize and convert mathematical content in PDFs to MathML.

Tasks
- Review existing methods and tools for extracting and converting mathematical content in PDFs.
- Develop an approach to convert mathematical content in different presentation formats within PDFs to MathML.
- Evaluate the effectiveness and efficiency of your approach.

Contact
Moritz Schubotz
moritz.schubotz@uni.kn

Norman Meuschke
norman.meuschke@uni.kn

Bela Gipp
bela.gipp@uni.kn
Image Credits

2. https://ja.wikipedia.org/wiki/%E3%83%9E%E3%82%AF%E3%82%B9%E3%82%A6%E3%82%A7%E3%83%AB%E3%81%AE%E6%96%B9%E7%A8%8B%E5%BC%8F#.E5.86.85.E5.AE.B9
6. https://commons.wikimedia.org/wiki/File:Social_Network_Analysis_Visualization.png
You can find guidelines, tutorials and templates for your project and thesis in our wiki:

isg.uni.kn/wiki

(feel free to share with others)
You can find these slides and more information at:

isg.uni.kn/students-corner/