Who are we now?
An attempt to understand the ICDAR community!

Andreas Dengel
## What makes a conference memorable?

1. Find a conference place with a ‘wow factor’
2. Provide a excellent service to the audience
3. Incorporate workshops that require involvement, i.e. encouraging senior and junior people to contribute in an open atmosphere
4. Makes sure that the reception and the banquet are exceptional and top the expectations of the participants
5. Invite keynote speakers who are engaged, convincing, entertaining, and talk about topics which are beyond traditional approaches
From a historical point of view, we considered a document image as a subject of study and interpretation.

A document is a ...

- a piece of paper, booklet, etc., providing information, especially of an official or legal nature
- a piece of text or text and graphics stored on a computer as a file for manipulation by document processing software
- evidence or a proof
- a written or drawn representation of thoughts
- a textual file along with its structure and design (fonts, colors, and additional images)
- a written proof used as evidence

Document Analysis and Recognition* includes contributions dealing with computer recognition of characters, symbols, text, lines, graphics, images, handwriting, signatures, as well as automatic analyses of the overall physical and logical structures of documents, with the ultimate objective of a high-level understanding of their semantic content.

* Sources: Wikipedia; Dictionary; International Journal on Document Analysis and Recognition (IJDAR)
The analysis task may not be restricted to a single document only but to an entire set of documents capturing important information in their combination.

A documentation is a set of documents provided on paper, or online, or on digital or analog media:
- It is mainly required to transfer information through time.
- It is becoming less common to see paper (hard-copy) documentation.
- User manuals, quick-reference guides, or conference proceedings are typical examples.

Professionals educated in this field are termed documentalists. This field changed its name to Information Science in 1968, but some uses of the term documentation still exists and there have been efforts to reintroduce the term documentation as a field of study.
Document Analysis and Recognition has a long tradition culminating in the foundation of the ICDAR conference series documented within the set of proceedings.

- The proceedings corpus provides is a valuable source for a whole bunch of quantitative and qualitative analyses.
- We may apply OCR and layout analysis on the entire corpus for reediting, reformatting, rearranging, etc., for, e.g., generate a book on specific topic or a historical overview.
- We may analyze the contents to classify the various papers, to get an idea about the employment of specific approaches, etc.
The analysis may be also expanded one step more to get insights into the social network behind the proceedings corpus.

In general, a social network is a social structure made of nodes (which are generally individuals or organizations) that are tied by one or more specific types of interdependency, such as values, visions, ideas, financial exchange, friends, kinship, dislike, conflict, trade, web links, sexual relations, disease transmission, or airline routes. The resulting structures are often very complex.

Corpus captures a large-scale but hidden “social (semantic) structure with

- directed weighted relations,
- network properties,
- performances of subsets, and
- location of actors,

based on counting the incoming and outgoing links and applying mathematical models.
Generally three basic types of networks can be created from a set of scientific publications

- **Co-authorship networks** is used as an indicator for the collaboration of authors (and their affiliated institutions)
  - Nodes represent the author
  - Node size conveys the number of publications by the author
  - Edge label denotes the number of times two authors have co-authored a paper

- **Publication citation networks** show relationships among scientific articles based on their citations
  - Direct citation network
  - A bibliographic coupling network where an edge is drawn between two publications if both cite the same previous publication(s)
  - Co-citation network where an edge between two publications exist if they both cite each others publication

- **Semantic networks** are formed on the basis of occurrence of a keyword in a set of publications (nodes represent words and edges represent the co-occurrence of these words in one article)
This way, we approached this problem by a holistic document analysis approach to identify relations and to measure impact factors.

The goal is to reveal

- historical contributors, their relationship, topics, and technologies addressed by community members,
- scientific trends and opinion leaders,
- citation behavior as well as collaborating cliques
Syntactic Document Analysis addresses the physical composition of a document

PDF

1. Identify format (single/multiple columns)
2. Ignore tables, images
3. Transform text (image) into Unicode
4. Normalize the format of the references

Meta Data Extraction

1. Extract Meta Data (title, header, keywords, abstract and references)
2. Extract sentences capturing citations/references using Named Entity Recognition

Structured Data

- Citation Data (from 1993 – 2015)
- Collaboration Data (joined publication)
- Keyword Information (original pdf or IEEE Bibtex)
Semantic Document Analysis deals with the interpretation of the document structure and content

XML

Network Generation

1. Build a co-authorship network for each document
2. Build a author citation network for each document
3. Aggregate the individual data and weight both graphs, i.e. nodes wrt (1) and edges wrt (2)

Data Cleaning and Polarity Measure

1. Delete “non-community” author names, i.e. authors who never published at ICDAR conferences
2. Name Resolution
3. Normalize the format of the author names and adapt network
4. Citation Sentiment Analysis

Structured Data

- B-PER, I-PER, E-PER tags for describing unique author names
- Aggregated collaboration and publication data

Enriched Networked Community Structure
Pragmatic Document Analysis tries to get insights into what is meant by the various propositions, i.e. its intent and effect

**Enriched Networked Community Structure**

- **Data Clustering**
  - Clustering of co-authorship networks into groups of authors who frequently publish together

- **Performance Indicator Detection**
  - Statistical summarization
  - Co-authorship evolvement analysis
  - Region-based co-authorship analysis
  - Measuring author centrality

- **Visualize Graph Data**
  - Intuitive representations of a large amount of data realizing multiscale navigations
  - Interactive community data regarding authors, their collaboration and joint publishing
  - Various statistics regarding the different roles of individuals, stratification, and cliques

**Academic Community Explorer**
In the period between 1993 and 2015, more than 3,500 authors from 55 countries published papers in the ICDAR proceedings.

Macro Indicators (not taken from statistics but from document analysis results)

- The average community member authored 6.01 papers
- The average number of co-authors per paper continuously increased from 3.23 (in 1997) to 4.75 (in 2015)
- The community clustering coefficient is quite high compared to other communities, i.e. authors collaborate more frequently compared to other communities we investigated
- The largest connected component of the ICDAR community co-citation network emphasizes increased from 17% in 1993 to 46% in 2003 to 60% in 2009 and 70% in 2015
- The total number of co-citations between 2005 and 2015 has increased by 150% although the number of publications remain almost the same

Note: in 2017 we have a total of 212 (52 oral & 160 poster) papers
Regarding the countries contributing to ICDAR conferences there are obvious long term trends.

Trend Line of Contribution from Exemplary Countries

Note: From the early beginning of ICDAR, France has continuously contributed a high number of papers!
A co-authorship network is a special kind of a social network where co-authors are represented as nodes and co-authorship is represented by edges between two nodes.

As micro indicators in co-authorship and co-citation networks, we employ centrality measures helping to find nodes reflecting different dimensions of importance, e.g. influence, leadership, and position.
A high degree centrality denotes the existence of authors who collaborate very frequently with other authors.

**Micro Indicators in co-authorship**

- **Degree centrality** $DC :=$ number of ties to directly neighbored nodes

  $$C_D(v) = deg(v)$$

- **In our case** we may distinguish between indegree and outdegree centrality

- **Koichi**, in our case, is considered as the most central actor
A high closeness centrality means a short path lengths to all other network participants

- **Closeness centrality** \( CC := \text{sum of distances (geodesics) to all other nodes} \)

\[
C(u) = \frac{n - 1}{\sum_{v=1}^{n-1} d(v, u)}
\]

- George and Seiichi, in our case, have the highest closeness centrality value

\[ CC(\text{George}) = \frac{9}{14} \]
Nodes having a high betweenness centrality may control the flow of information among subsets and have therefore a central and powerful position.

**Betweenness centrality**

\[ BC(v) = \frac{\sum_{s \neq v \neq t} \sigma_{st}(v)}{\sigma_{st}} \]

- Cheng, in our case, has the highest betweenness centrality value.

- For Cheng this value is 14 (7 geodesics each for Horst and Josep).
There are some interpretations for the individual centrality measures

**Degree Centrality**
Denotes authors who are open and collaborate very frequently with other authors

**Closeness Centrality**
Denotes the prominence an authors as an opinion leader since she/he is well connected

**Betweenness Centrality**
Denotes authorities who act as bridges between small subgroups in a community (critical ties)
Authors citation networks are networks where authors are nodes and direct edges exist between A and B if “A cites B” in one of his papers.

**Micro Indicators**

- Having a high **in-degree centrality** implies that an author is a dominating person in a community based on her/his scientific contribution (most cited in the case of highest in-degree).

- A high **betweenness centrality** in authors citation networks is given if an author is diversely publishing with different cliques in the community.

  In most cases such authors show a almost balanced score in citing others and being cited by others.

- An author who has a high **Eigenvector centrality** is considered as being highly influential, e.g. authors who are most likely to produce new research ideas.

  An author’s Eigenvector centrality is proportional to the sum of the Eigenvector centralities of all nodes directly connected to her/him.
In order to find important patterns of co-citation, we cluster author instances into various cliques citing each other's work more often

Clustering of a social structure is motivated by the principle of 'homophily', which is the tendency of individuals to associate and bond with similar others.

Individuals in homophilic relationships share common characteristics (in our case: topics, approaches, applications, etc.) that make communication and relationship formation easier.

Instead of vertex betweenness, we in this case extend this measure to the case of edges, employing the "edge betweenness", i.e. number of shortest paths between pairs of nodes that run along it.

The Girvan-Newman Clustering* algorithm detects communities by progressively removing edges of high betweenness centrality from the original network.

1. Calculate betweenness centrality for all edges existing in the network
2. Remove edge(s) with highest betweenness centrality
3. Recalculate betweenness centrality of all remaining edges
4. Repeat Steps 2 and 3 until no edges are left

Prominent researchers in ICDAR reveal different roles when considering the normalized centrality values resulting from the analysis.

<table>
<thead>
<tr>
<th>Author</th>
<th>Coauthorship networks</th>
<th>Author citation networks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Closeness</td>
<td>Betweenness</td>
</tr>
<tr>
<td>U. Pal</td>
<td>0.82</td>
<td>0.72</td>
</tr>
<tr>
<td>C. Liu</td>
<td>0.93</td>
<td>0.56</td>
</tr>
<tr>
<td>M. Liwicki</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>C. Tan</td>
<td>0.87</td>
<td>0.74</td>
</tr>
<tr>
<td>S. Uchida</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>C. Suen</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>D. Karatzas</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>K. Kise</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>A. Dengel</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>J. Llados</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>S. Srihari</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>M. Nakagawa</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>D. Doermann</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>C. Jawahar</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>D. Lopresti</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>M. Blumstein</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>V. Govindaraj</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>M. Cheriet</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>M. Iwamura</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>R. Ingold</td>
<td>0.92</td>
<td>0.74</td>
</tr>
</tbody>
</table>

- **“Opinion Leaders”**
- **“Networkers”**
- **“Community Experts”**
- **“Connectors”**
- **“Collaborators”**
- **“Scientific Leaders”**
Anyway, the top-influential authors of the ICDAR community show different characteristics, i.e., they play slightly different roles.

### Results of Quantitative Analysis

<table>
<thead>
<tr>
<th>Author</th>
<th>Betweenness Centrality</th>
<th>Degree Centrality</th>
<th>Indegree Centrality</th>
<th>Eigenvector Centrality</th>
<th>Outdegree Centrality</th>
<th>Closeness Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. LIU</td>
<td>0.645475144281</td>
<td>1</td>
<td>0.800762631054</td>
<td>1</td>
<td>1</td>
<td>0.996097761509</td>
</tr>
<tr>
<td>H. BUNKE</td>
<td>1</td>
<td>0.962396430875</td>
<td>1</td>
<td>0.132500818835</td>
<td>0.632625085683</td>
<td>0.96363997615</td>
</tr>
<tr>
<td>B. GATOS</td>
<td>0.340646448672</td>
<td>0.622370936879</td>
<td>0.541468064756</td>
<td>0.134087160471</td>
<td>0.560657964943</td>
<td>0.957930080984</td>
</tr>
</tbody>
</table>
For visualizing hidden structures of cliques in the network, we may set different thresholds for the minimum number of co-citations.

The examples show a cluster or clique from ICDAR community where each edge represents a citation strength of at least $n$ between all of its members.

- **The ICDAR Social Network with edges with a minimum citation count of 6!**
  - 746 authors with 19,239 links
- **The ICDAR Social Network with edges with a minimum citation count of 11!**
- **The ICDAR Social Network with edges with a minimum citation count of 21!**
  - 81 authors with 2,875 links
Moreover, zooming into the community networks shows the author connections in more detail where the center captures the most influential authors.
Having a closer look, we receive cliques of authors who collaborate and cite each other frequently.

The examples show clusters or cliques from ICDAR community where each edge represents a citation strength of at least 21 between all of its members.

Note: Same color means a high number of co-authored papers, i.e. intensive collaboration.
Clicking on a node shows individual citation networks of authors, i.e. most influenced community members.
Different authors show different data characteristics reflected by the Overlap Index Graph aggregating and showing the diversity of an author in the community.
The graph may be investigated via navigation options

The ICDAR Social Network with edges with a minimum citation count of 11!
When clicking on one node in the network, we get a detailed overview of all relevant data dimensions.
When clicking on one node in the network, we get a detailed overview of all relevant data dimensions.
Furthermore, topic clouds offer insights into an author’s domain of research at a glance.

Topic Cloud of T. Breuel (Minimum Count of 11)

Topic Cloud of S. Srihari (Minimum Count of 11)

Topic Cloud of D. Karatzas (Minimum Count of 11)
In addition, we try to address another very challenging problem which is of high relevance but very hard to solve (work in progress)

Motivation:

- There are behavioral patterns of researchers' literature review practice in each community capturing the author's opinion toward the cited work.
- Citation Sentiment Analysis (CSA) is a newly emerged research topic inspired by traditional citation context analysis in scientometrics and applied linguistics.
- In a so-called Senti-Index, we intend to consider citation sentiment as the polarity of a citing author's opinion using the categories positive, negative and neutral (or many more).

Just to show (again) the target of other communities:

- Document Analysis is a form of qualitative research in which documents such as public records of an organization (community) are interpreted by the researcher to give voice and meaning around an assessment topic, e.g. to provide a confluence of evidence that breeds credibility.
While for the SNA we focused on authors, abstract, keywords and list of references, in CSA we analyzed the text of the paper including the captured references.

Text

1. Tokenizer
2. Regular Expression Sentence Splitter

Tagged Structured Text

1. Reference Transducer
2. Reference Sentence Transducer
3. TokenML Transducer

Filtering

1. POS Tagger
2. WorNet Suggester
3. Stemmer

Tagged Struct. Ref. Sentences

1. Sentiment Analysis
2. Nature of References

Information Extraction

Features

Reference Polarity Score


D. Mercier, A. Bhardwaj, S. Ahmed, and A. Dengel, SentiCite: An Approach for Publication Sentiment Analysis, ICAART-18, 10th Int’l Conference on Agents and Artificial Intelligence, Madeira, Portugal (Jan. 2018),

©andreas.dengel@dfki.de - 2017 - P. 33
CiTO*, the Citation Typing Ontology, is an ontology to enable characterization of the nature or type of citations, both factually and rhetorically.

The citations characterized may be:
- direct and explicit (as in the reference list of a journal article),
- indirect (e.g. a citation to a more recent paper by the same research group on the same topic)
- implicit (e.g. as in artistic quotations or parodies, or in cases of plagiarism)

Properties: agreesWith, cites, citesAsAuthority, citesAsDataSource, citesAsEvidence, citesAsMetadataDocument, citesAsRelated, citesAsSourceDocument, citesForInformation, confirms, containsAssertionFrom, corrects, credits, critiques, disagreesWith, discusses, disputes, documents, extends, givesBackgroundTo, givesSupportTo, includesExcerptFrom, includesQuotationFrom, isAgreedWithBy, isCitedAsAuthorityBy, isCitedAsDataSourceBy, isCitedAsEvidenceBy, isCitedAsMetadataDocumentBy, isCitedAsRelatedBy, isCitedAsSourceDocumentBy, isCitedBy, isCitedForInformationBy, isConfirmedBy, isCorrectedBy, isCreditedBy, isCritiquedBy, isDisagreedWithBy, isDiscussedBy, isDisputedBy, isDocumentedBy, isExtendedBy, isParodiedBy, isPlagiarizedBy, isQualifiedBy, isRefutedBy, isReviewedBy, isRidiculedBy, isSupportedBy, isUpdatedBy, obtainsBackgroundFrom, obtainsSupportFrom, parodies, plagiarizes, providesAssertionFor, providesDataFor, providesExcerptFor, providesMethodFor, providesQuotationFor, qualifies, refutes, reviews, ridicules, sharesAuthorsWith, supports, updates, usesDataFrom, usesMethodIn.

* D. Shotton and S. Peroni, CiTo: The Citation Typing Ontology, http://www.sparontologies.net/ontologies/cito/source.html
The Senti-Index allows to hover over a node or edge and get an idea about an authors (author2author) citation sentiment.
Take-Aways!

- I presented an approach for a holistic analysis of an entire document corpus which in combination of all containing documents represents a rich source of information.
- All results shown are available via the Academic Community Explorer framework (cf. [http://www.dfki.uni-kl.de/ace/](http://www.dfki.uni-kl.de/ace/)) analyzing scholarly document metadata to study a scientific community.
- The topic clouds offer an overview of an author’s domain of research and may be used to search for appropriate reviewers for ICDAR and IJDAR.
- The employment of the co-authorship network may avoid “conflict of interest” cases when assigning reviewers because it implicitly shows frequent co-authorship.
- The Senti-Index is a first but promising approach to disclose sentiment in citations but has to be considered with great caution.
- Anyway, finally this work will continue and should be used as a recommending and exploring system for community members.
Finally, I hope I was able to stimulate discussion regarding both, document analysis research questions/options and the hidden patterns of our community.

Special thanks to Akansha Bhardwaj, Sheraz Ahmed, Dominique Mercier, Hisham Hashmi.

Questions?

Prof. Andreas Dengel
DFKI GmbH
P.O. Box 2080
D-67608 Kaiserslautern
e-mail: andreas.dengel@dfki.de