KNOWLEDGE EXPLORATORY PROJECT FOR NANODEVICE DESIGN AND MANUFACTURING: KNOWLEDGE DISCOVERY FROM EXPERIMENTAL RECORDS

NANODEVICE RESEARCH PAPERS
CLUSTERING BASED ON AUTOMATIC PAPER ANNICATION

Thaer M. Dieb   Masaharu Yoshioka   Shinjiro Hara

Graduate School of Information Science and Technology
Hokkaido University
In order to support this process, a Knowledge Exploratory Project for Nanodevice Design and Manufacturing was conducted. Using MOVPE had been proven a very effective way to produce nanocrystals. Manual Annotation was used to annotate research papers.

Automatic information extraction using machine learning and named entity recognition were applied to non-annotated research papers to construct an annotated corpus.

- **Nanocrystal with MOVPE**
  - Using MOVPE had been proven a very effective way to produce nanocrystals.

- **Manual Annotation**

- **Nanocrystal with MOVPE**
  - Using MOVPE had been proven a very effective way to produce nanocrystals.
OBJECTIVE

- Utilize the extracted information
  - Document clustering
CORPUS CONSTRUCTION

INFORMATION CATEGORIES

Source material

Material characteristics

Product

Hole opening

Temperature

Surface conductivity

Evaluation parameters

Control parameters

Manufacturing method

Control parameters

Values

Values
We investigated hexagonal symmetric-shaped NiAs-type MnAs NC array, whose period was 1 µm in a triangular lattice arrangement, position-controlled on the GaAs (111)B substrates with the SiO2 mask opening size of 300 nm by SA-MOVPE. The $T_g$ and V/Mn ratio were 850 °C and 375, respectively. Figure 1(a) shows a SEM top view of the NC array. Typical NCs measured 180 nm high and 480 nm wide. MFM for the “as-grown” NCs revealed that magnetic responses were taken only for the MnAs NCs grown in the SiO2 mask openings, and that no significant signal was detected for the SiO2 masks covered on GaAs wafers, as shown in Figs. 1(b)–1(e).

A) Material Information InGaAs
B) Material features (111)B
C) Experimental parameters growth temperatures
D) Value of experimental parameters 600–700°C
E) Evaluation parameters Photoluminescence (PL) spectra
F) Value of the evaluation parameter 0.95 to 1.3 eV
G) Manufacturing method catalyst-free selective-area metal organic vapor phase epitaxy
H) Artifact nanowires

Manually annotated document
AUTOMATIC ANNOTATION FRAMEWORK

- A combination of
  - Machine learning techniques
  - Named entity recognition
    - Chemical compound supports source material annotation
    - Parameter supports experiment and evaluation parameter annotation
We demonstrate the successful formation of ferromagnetic MnAs nanoclusters self-assembled on GaInAs (1 1 1) B surfaces by metalorganic vapor phase epitaxy (MOVPE). The hexagonal MnAs nanoclusters show strong ferromagnetic coupling.
Utilization of Corpus Information

- Finding similar documents
  - Researchers can find discussion about similar experiment settings.
  - Use clustering techniques to group similar documents
    - Different similarity metrics can be used.
      - Basic approaches like bag-of-words leads to one type of clustering.
**Annotated Document Clustering**

- Biased clustering
  - different types of clustering based on certain kind(s) of information category(s).

- We do not know which information categories are more important in document clustering.
  - Experiment on classified documents is necessary
Clustering Experiment
Conference Sessions Classification

- Finding the effective information categories in separating conference sessions
  - 160 papers from conference proceeding * classified into 5 classes (A-E) based on session
- Two clustering experiments: with and without automatic annotation
- Hierarchical clustering using R language

Paper Similarity

Paper Representation

- Bag-of-words approach
  - Non-annotated paper: plain Bag-of-words
  - Annotated paper: array of vectors, each vector contains a bag-of-words representation of all chunks annotated under certain information category.
Weighted cosine similarity

\[
\text{Similarity} = \frac{\mathbf{a} \cdot \mathbf{b}}{\| \mathbf{a} \| \cdot \| \mathbf{b} \|} = \frac{\sum_{i=1}^{n} a_i x b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (b_i)^2}}
\]

Where \( \mathbf{a}, \mathbf{b} \) are document vectors.
Paper Similarity

2 different encoding

- Long vector
  \[ a = (\alpha_1 a_1, \alpha_2 a_2, \ldots, \alpha_9 a_9) \]
- All sum

\[ \text{Similarity} = \sum_{i=1}^{n} \alpha_i \frac{a_i \cdot b_i}{\|a_i\| \cdot \|b_i\|} \]

Where \( a_i \) and \( \alpha_i \) represent bag-of-words vector of a document and weight for \( i \)-th category (SMaterial, SMChar, MMethod, TArtifact, ExP, EvP, ExPVal, EvPVal, and Other).
Experiments

Experiment Results

- Annotated papers
  - 2 ways of encoding
  - Different weight vectors
    - Testing 0 for unimportant, 1 for normal, 10 for important

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Base system (non-annotated)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[1,1,1,1,1,1,1,1]</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>[1,1,1,1,1,1,1,1]</td>
<td>0.27</td>
<td>0.29</td>
</tr>
<tr>
<td>[1,1,1,1,1,1,0,0]</td>
<td>0.27</td>
<td>0.3</td>
</tr>
<tr>
<td>[1,1,1,1,1,1,0,0]</td>
<td>0.27</td>
<td>0.29</td>
</tr>
<tr>
<td>[10,10,10,10,10,10,10,11]</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>[1,1,1,1,1,10,10,0,1]</td>
<td>0.27</td>
<td>0.29</td>
</tr>
<tr>
<td>[1,10,1,1,10,10,0,0,1]</td>
<td><strong>0.26</strong></td>
<td><strong>0.29</strong></td>
</tr>
</tbody>
</table>

Effective information categories: **SMChar** (source material characteristics), **ExP** (experiment parameter), **EvP** (evaluation parameter)

Unimportant information categories: **ExPVal**, **EvPVal** (values)

SM=SMaterial, SMC=SMChar, MM=MMethod, TA=TArtifact, EP=ExP, Ev=EvP, EPV=ExPVal, EvV=EvPVal, O=Other
EXPERIMENTS
RESULTS ANALYSIS

- Different information category plays different roles in similarity
- Encoding method considerably affect the clustering quality. Long vector encoding generally performed better.
- The “Other” category seems to play significant role in similarity, and that is because automatic annotation quality is still not good enough.
Increasing weights of effective information categories does not always increase the quality of the clustering.
Experiments

Base system

- Non-annotated papers

Hierarchical clustering result for non-annotated papers

Entropy: 0.28  Purity: 0.38
Experiments
Best Performance

- Annotated papers

Hierarchical clustering result for annotated papers

Entropy: 0.26  Purity: 0.40
CONCLUSION

- Proposed a framework to cluster research documents based on similarity using annotated information
  - Annotation of documents might have improved the clustering quality, but not confirmed
    - Clusters are biased
    - Effective information categories are valid at least for the experimental conference proceeding
      - Different proceedings might have different session classification criteria.

- Future work
  - Plan to test the framework on more balanced collection of papers
御静聴有難う御座いました