1C3-4

ナノ知識探索プロジェクト：実験記録からの知識発見(第３報)

Knowledge Exploratory Project for Nanodevice Design and Manufacturing:
Knowledge Discovery from Experimental Records (3^{rd} Report)

Nanodevice Research Papers Clustering based on Automatic Paper Annotation

Thaer M. Dieb[*1] Masaharu Yoshioka[*1] Shinjiroh Hara[*2]

[*1] Graduate School of IST, Hokkaido University [*2] RCIQE, Hokkaido University

We have been working on the project “Exploring Knowledge for Nanodevice Development” and proposed a framework to annotate useful information (e.g., source material, evaluation parameter, and so on) for analyzing nanodevice development papers. In this paper, we conduct nanodevice research papers clustering experiments, once on annotated papers using automatic annotation results, and once on the non-annotated same papers using bag-of-words approach, and then we compare the results discussing the usefulness of the automatic annotation.

1. Introduction

In order to support nanodevice development process, we are working on a project called “Exploring Knowledge for Nanodevice Development” [吉岡 10]. This project aims at providing insights for nanodevice novice researchers to help them planning their experiments more effectively.

In this project, we have proposed a framework to annotate useful information from research papers related to nanodevice development (e.g., source material, evaluation parameter, and so on) [Dieb 11], and use them for analyzing experiment results. However; since manual annotation of papers might be very time consuming, we have built an automatic annotation framework using machine learning techniques [Dieb 12].

In this paper, we propose to use automatic annotated information for calculating similarity between papers. In order to demonstrate the effectiveness of the similarity measures, we conduct nanodevice research papers clustering experiment based on these measures and compare its results with the one based on ordinary similarity measure based on bag-of-words approach.

2. Automatic Annotation

2.1 Background

Based on discussion with nanodevice development researchers, we have decided to extract characteristic information from research papers to enhance the analysis of nanodevice development experiment results [Dieb 11]. We have decided on 8 types of information as listed below:

- Material (SMaterial) e.g., As, InGaAs
- Characteristic Feature of Material (SMChar) e.g.,(111)B
- Experiment Parameter (ExP) e.g., total pressure
- Value of the Experiment parameter (ExPVal) e.g., 50nm
- Evaluation Parameter (EvP) e.g., peak energy, FWHMs
- Value of the Evaluation Parameter (EvPVal) e.g., 1.22eV
- Manufacturing Method (MMethod) e.g., SA-MOVPE
- Final Product (TArtifact) e.g., semiconductor

The information in the papers is annotated using XML format. Figure 1 shows an example of an annotated paper.

```
We demonstrate the successful formation of <TArtifact>
<SMChar>ferromagnetic</SMChar>
<SMaterial>MnAs</SMaterial> nanoclusters </TArtifact> self-
assembled on <SMaterial> GaInAs</SMaterial><SMChar>(1 1 1)
B </SMChar> surfaces by <MMethod>metalorganic vapor phase
epitaxy </MMethod><MMethod>MOVPE</MMethod>).
The <TArtifact><SMChar>hexagonal</SMChar>
<SMaterial>MnAs</SMaterial>nanoclusters</TArtifact> show
<EvPVal>strong</EvPVal> <EvP>ferromagnetic
coupling</EvP><ExPVal>at room temperature </ExPVal> when
the <ExP>external magnetic fields </ExP> are applied <ExPVal>in
a direction parallel to the
<SMaterial>InP</SMaterial><SMChar>(1 1 1) B</SMChar>
wafer planes</ExPVal>.
```

Fig.1 Example of annotated paper

2.2 Automatic annotation framework

Since manual annotation of research papers is a very slow process, and requires domain experts, we have decided to build an automatic annotation framework using the machine learning techniques, and the already manually annotated papers as training data [Dieb 12].

There are 3 main issues to be discussed in the automatic annotation framework:

Contact: Thaer M. Dieb, Graduate School of IST, Hokkaido University, diebt@kb.ist.hokudai.ac.jp

(1) Overlapped tag structure:

In nanodevice development domain, tags are not simple structures as might be in other domains, i.e. tags can overlap within each other. Figure.2 shows an example of tag overlapping.

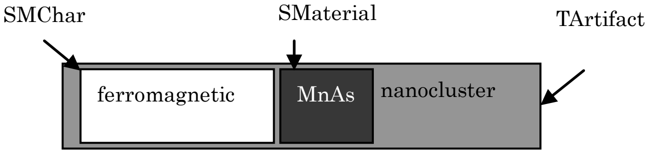


Fig.2 Example of overlapped tag structure

It is not easy for the machine to learn to set the correct tags information within small window size all at once. It is necessary to separate overlapped tags in the process of training the machine. We separated tag types into 4 groups where tags don't overlap within the same group. Based on these 4 groups, we divided the learning process into 4 cascading levels i.e. cascading named entity recognition [Kano 11]. In each level, the machine will try to use tag information estimated from previous levels to estimate tags information in the current level.

(2) Chemical entity recognition:

Most of the Source Material items in the research papers are chemical compounds. If we have large training data, the machine might be able to identify source material based on the training data only. However; since training data size is still very small (so far only 5 full papers manually annotated), identifying chemical entities will help the machine to recognize Source Material tags.

We have developed a new chemical entity recognizer called CNER. CNER is a rule-based chemical entity recognizer that uses a modified version of the rules in OSCAR3 [OSCAR3]. It uses regular expressions to identify chemical compounds. In addition to that, it uses syntactic rules to eliminate some mismatches that might occur between chemical entities and general text. We have proven the efficiency of using CNER for better automatic annotation quality.

(3) Parameter identification:

In order to further improve the quality of automatic annotation, we have added another component to our automatic annotation framework, which is the parameter identification component.

As we have discussed in the chemical entity recognition, and since the training data size is still small, identifying common parameters will help the learning process in recognizing experiment and evaluation parameters. However; we can't separate experiment and evaluation parameter without context, because some parameters such as size can be used as both experiment and evaluation parameter depending on the context. Because of that, we only identify parameters in general before the automatic annotation process, and then the machine will separate experiment parameter from evaluation parameter based on the training data.

We have made a list of most common parameters in nanocrystal development. This list was extracted from review papers discussing parameters for nanocrystal development, and also from revising representative research papers.

Units might be a good clue for the machine to annotate parameters. We enhanced the parameter identification process by identifying units also.

In this framework, YamCha [YamCha] (a text chunk annotator based on SVM) is used as a machine-learning system for estimating the chunk annotation using given features (POS, orthogonal, and so on). For POS tagging, we use GPoSTTL [GPoSTTL], which is a modified version of Brill tagger.

3. Paper Clustering

3.1 Introduction

Usually researchers have to go through a trial and error process, modifying parameter settings for experiment several times before they can reach the most convenient settings that can yield to the desired final product. This trial and error process is time and money consuming.

Research papers contain summarization for several nanocrystal device development experiments and evaluation about experiment results. Research papers can be used by novices to find similar experiments done before to help them plan their new experiment more effectively.

Different similarity measures can be used to study the similarity between papers. Clustering research papers based on the similarity of their content can allow us to study similarity measures and their effect on papers' similarity.

We propose to use the automatic annotation framework discussed in section 2 in the process of calculating similarity between papers. Automatic annotation can provide additional similarity measures, hence deeper analysis on the study of similarity, especially when categories of information are playing different roles in determining the similarity between research papers. Automatic annotation can tell which information category is more important in finding similarity.

In order to discuss the effectiveness of automatic annotation on the similarity measure analysis, we are conducting two clustering experiments on the same set of papers, one without automatic annotation, and the other with automatic annotation using our automatic annotation framework.

There are different ways to cluster research papers based on similarity; However, in this paper, we study one way using bag-of-words approach. Other ways of clustering might also be studied and analyzed in the future.

3.2 Papers similarity

There are many ways to find similarity between 2 papers. However, in any way, we need to transform the paper into representative model to be able to calculate similarity between it and another paper. We use bag-of-words approach as a model for papers to find similarity. Bag-of-words approach is a simplifying representation of text documents as unordered vector of words and their frequencies.

As we have mentioned in section 3.1, we have conducted two experiments, the first one is on non-annotated papers (base system), and in that case, we transformed each research paper (non-annotated) into a bag-of-words model. The other one is on annotated papers, and in this case, we transformed each paper (annotated) into an array of vectors; each vector contains a bag-of-words representation of all chunks annotated under certain information category. We have 8 categories of information that has been annotated. In addition to that, we added extra category called "Other", that is non-annotated chunks of text within the annotated paper. In total, we have 9 categories of information. Each annotated paper is transformed into an array of 9 vectors; each one is a bag-of-words representation of tags under certain information category.

In this paper, we use weighted cosine similarity metric. Cosine similarity metric is given by the equation (1)

$$\text{Similarity} = \frac{\mathbf{a} \bullet \mathbf{b}}{\|\mathbf{a}\| . \|\mathbf{b}\|} = \frac{\sum_{i=1}^{n} a_i x b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} x \sqrt{\sum_{i=1}^{n} (b_i)^2}} \quad (1)$$

Where **a**, **b** are document vectors.

In the case of non-annotated paper, we use whole bag-of-words vector as a document vector for calculating similarity. However; in case of annotated paper, where paper is segmented into 9 vectors of bag-of-words, each one has different weight, there are different ways to calculate weighted cosine similarity. In this paper, we have encoded 2 ways to calculate weighted cosine similarity:

- Long vector encoding: we construct a long vector that concatenates 9 vectors with weight that represents importance of each information category

  $a = (\alpha_1 \mathbf{a_1}, \alpha_2 \mathbf{a_2}, \cdots, \alpha_9 \mathbf{a_9})$

  Where $\mathbf{a_i}$ and α_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).

  Similarity is calculated using equation (1).
- All sum encoding: we calculate cosine similarity for bag-of-words vectors for every category including "Other" and summed all similarity with weight.

$$\text{Similarity} = \sum_{i=1}^{n} \alpha_i \frac{\mathbf{a}_i \bullet \mathbf{b}_i}{\|\mathbf{a}_i\| . \|\mathbf{b}_i\|} \quad (2)$$

Weighted cosine similarity can allow us to neglect or emphasize certain category of information based on the importance it plays in determining similarity between papers.

4. Experiments

4.1 Experiment setup

We conducted nanodevice paper clustering experiments by using conference proceedings [SSDM 08], and session categories for these papers are used for the correct category labels (classes). We pick up 5 sessions (A-E) and select 32 papers from each session (total 160 papers). All papers are annotated using our automatic annotation framework. We have used hierarchal clustering technique [Gothai 10] with R language [R] to perform both experiments. Hierarchal clustering seeks to build a hierarchy of clusters, each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. There are different methods to merge the observations on the way up depending on the distance between them. We have tested them, and "complete linkage" method, that uses the maximum distance performed best. We cut the clustering result tree into 5 levels representing 5 clusters, and then we compare the 5 resulting clusters with the original 5 classes to determine the quality of clustering. We evaluate the clustering quality using the entropy and purity measures. Entropy measures how the conceptual classes are distributed within each cluster. The smaller the entropy the better the clustering is. Purity is the fraction of the cluster size that the largest class of chunks assigned to that cluster represents. The larger the purity, the better the clustering is.

4.2 Base system (non-annotated paper clustering)

We performed the first experiment using non-annotated papers. Figure 3 shows the resulting clusters. Entropy and purity were 0.28 and 0.38 respectively.

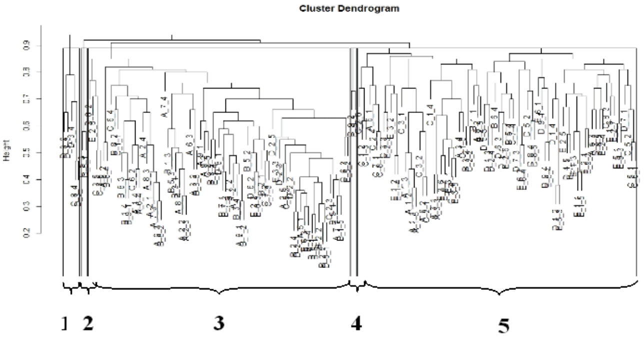


Fig.3 hierarchal clustering result for non-annotated papers

Analyzing figure 3, we find that, it is unbalanced structure of 5 clusters. There are 2 big clusters and 3 very small ones. However; deep look onto the distribution of papers within clusters, we found that A and B session papers are mostly clustered in cluster 3, and session E papers are mostly clustered in cluster 5. C and D session papers are distributed mostly between cluster 3 and 5. Because of the varieties of papers in both cluster 3 and 5, entropy and purity are not good enough.

4.3 Annotated paper clustering

As we discussed in section 3.2, we have 2 ways to encode the paper vector for calculating similarity, which are long vector and all sum. In each way of encoding, we can apply different weights for different information categories, and observe the quality of the clustering to find the best clustering strategy.

Table 1 shows the entropy and purity for different weighting strategies in both long vector and all sum encoding.

Table 1 Clustering performance for annotated papers

[SM, SMC, MM, TA, EP, Ev, EPV, EvV, O]	**Entropy**		**Purity**	
	Long vector	All sum	Long vector	All sum
Base system (non-annotated papers)	**0.28**		**0.38**	
[1,1,1,1,1,1,1,1,1]	0.27	0.28	0.39	0.31
[1,1,1,1,1,1,1,1,0]	0.3	0.29	0.27	0.32
[1,1,1,1,1,1,0,0,1]	0.27	0.30	0.37	0.26
[10,10,10,10,10,10,10,10,1]	0.28	0.29	0.34	0.31
[1,1,1,1,10,10,0,0,1]	0.27	0.29	0.4	0.33
[1,10,1,1,10,10,0,0,1]	**0.26**	0.29	**0.4**	0.29
[1,20,1,1,20,20,0,0,1]	0.28	0.3	0.34	0.28

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

4.4 Results analysis

We have conducted various weight settings to find out effective categories to calculate similarity. Since the clustering result of [10,10,10,10,10,10,10,10,1] is worse than one of the base system, we assume it is better to select useful categories for similarity calculation. We conducted several experiments and found following three categories are more effective to calculate similarity; Characteristic Feature of Material (SMChar), Experiment Parameter (ExP), and Evaluation Parameter (EvP). In addition, it is better to ignore values of parameters (ExPVal and EvPVal). Figure 4 shows the hierarchal clustering of best performed weight setting.

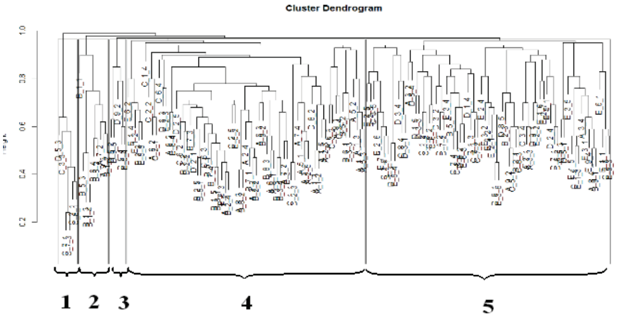


Fig.4 hierarchal clustering results for [1,10,1,1,10,10,0,0,1]

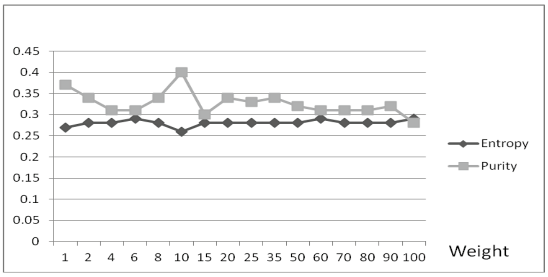
Analyzing best performance hierarchal clustering after annotation, we found it has almost the same structure as non-annotated papers clusters, keeping also 2 big clusters: 4, and 5, where cluster 4 has the bulk of sessions A and B papers, and cluster 5 has the bulk of E session papers. However there are some documents moved to smaller clusters making 2 small clusters bigger in size and almost pure. Even though the clustering structure and clustering performance were not much better than the basic system, but this means the automatic annotation might have some good effect on the quality of clustering. If the automatic annotation quality increases, it might increase the quality of the clustering; However, we cannot confirm that based only on these results. It is necessary to do more experiments with larger data size and more balanced clustering structure.

Considering the entropy and purity values from table 1, we can find the following notes:

- Encoding method considerably affect the clustering quality, noticing that the best performed weight setting in long vector encoding did not do well in the all sum encoding. 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. Figure 5 shows the relation between clustering quality and weight in long vector encoding best performance weights array.

5. Conclusion

In this paper, we have conducted nanodevice research paper clustering experiments based on similarity of the content. We have conducted 2 experiments, one before annotating the papers using automatic information annotation framework, and the other one after the annotation. We discussed the effect of automatic annotation on similarity measures analysis.



Weight for SMChar, ExP, EvP in [1,x,1,1,x,x,0,0,1] weight array

Fig.5 weight vs. performance in long vector encoding

In the future, we plan to use clustering to discover relations between different parameters, specially, experiment and evaluation parameters. Such relations can be very useful in determining the effect of a certain parameter change on the quality of the final product. In addition to that, we plan to discuss different clustering approaches.

ACKNOWLEDGEMENT

This research was partially supported by a grant for Hokkaido University Global COE program, Next-Generation Information Technology Based on Knowledge Discovery and Knowledge Federation", from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[Dieb 11] Dieb, T. M., Yoshioka, M., and Hara, S.: Construction of Tagged Corpus for Nanodevices Development Papers, GrC, 2011 Proceedings of the 2011 IEEE International Conference on Granular Computing, pp. 167–170, 2011.

[Dieb 12] Dieb, T. M., Yoshioka, M., and Hara, S.: Automatic Information Extraction of Experiments from Nanodevices Development Papers, IIAIAAI 2012 Proceedings of 2012 IIAI International Conference on Advanced Applied Informatics,pp.42-47,2012

[Gothai 10] Gothai, E. Performance evaluation of hierarchical clustering algorithms, (INCOCCI), 2010, Proceedings of 2010 International Conference on Communication and Computational Intelligence, pp.457-460, 2010

[GPoSTTL] GPoSTTL http://gposttl.sourceforge.net/.

[Kano 11] Kano, Y., Miwa, M., Cohen, K., Hunter, L., Ananiadou, S., and Tsujii, T., U-Compare: a modular NLP workflow construction and evaluation system. In IBM Journal of Research and Development, vol. 55, no. 3, pp. 11:1-11:10, 2011.

[OSCAR3] http://apidoc.ch.cam.ac.uk/oscar3/

[R] http://www.r-project.org/

[SSDM 08] Extended Abstract of the 2008 International Conference on Solid State Devices and Materials, Tsukuba, 2008

[YamCha] http://chasen.org/~taku/software/yamcha/.

[吉岡 10] 吉岡 真治, 冨岡 克広, 原 真二郎, 福井 孝志: ナノ知識探索プロジェクト:実験記録からの知識発見.2010 年度人工知能学会全国大会(第 24 回)論文集, CD-ROM 1B3-3, 2010.