Chemical Compounds Knowledge Visualization with Natural Language Processing and Linked Data

Kazunari Tanaka[1,2] Tomoya Iwakura[1,2] Yusuke Koyanagi[1,2] Noriko Ikeda[1] Hiroyuki Shindo[2,3] Yuji Matsumoto[2,3]

FUJITSU LABORATORIES LTD.[1],
RIKEN Center for Advanced Intelligence Project[2],
NARA INSTITUTE of SCIENCE and TECHNOLOGY[3]
{tanaka.kazunari, iwakura.tomoya, koyanagi.yusuke, nona}@jp.fujitsu.com,
{shindo, matsu}@is.naist.jp

Abstract

This paper proposes a visualization system for chemical compounds. New chemical compounds are being produced by every moment and registration of chemical compounds to databases strongly depends on human labor. Our system uses Natural Language Processing technologies for extracting information of chemical compounds from text and for storing the extracted results as Linked Data (LD). By combining the extracted results with LD-based existing chemical compound knowledge, our system provides visualization of chemical compound information such as integrated view of several databases and chemical compounds that have similar structures.

Keywords: Information extraction, chemical compounds

1. INTRODUCTION

Knowledge of chemical compounds has great value for developing new materials, new drugs, and so on. Therefore, databases of chemical compounds are being created. For example, CAS [1], one of the largest databases, includes information on over 100 million chemical compounds. However, the creation of such databases strongly depends on manual labor since chemical compounds are being produced at every moment. In addition, the database creation mainly focuses on English text. Therefore, in other words, chemical compound information other than English is not good enough to be available. For example, although Japan has one of the largest chemical industries and has large chemical compound information written in Japanese text documents, such information is not exploited well so far.

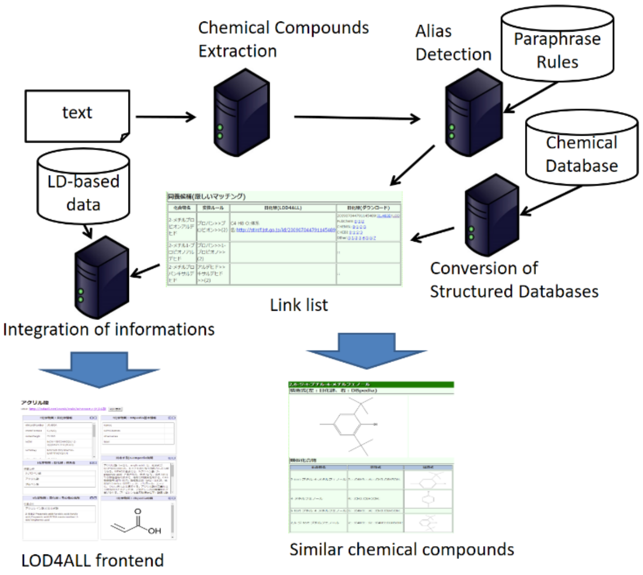
We propose a visualization system based on chemical compound extraction results with Japanese Natural Language Processing and structured databases represented as Linked Data (LD).

Figure 1 shows an overview of our system. First, chemical compound names in text are recognized. Then, aliases of chemical compound names are identified. The extraction results and existing chemical compound databases are represented as LD. By combining these LD-based chemical compound knowledge, our system provides different views of chemical compounds.

2. Chemical Compound Information Extraction

Unknown chemical compounds usually first appear in unstructured text such as scientific papers and patents. In order to extract chemical compound names from text, we use a Japanese Named Entity (NE) recognition method [2]. An NE recognizer is trained with a distant supervision [3] using a dictionary of chemical compound names as a lexical resource. The dictionary is compiled from Nikkaji [4].

The dictionary is used to recognize chemical compound names in text. The chemical compound names in text recognized with the dictionary are used as positive samples for training. Negative samples are documents that would be irreverent with chemical domains such as politics and sports. Then, the automatically created training data was used for training an NE recognizer.



Chemical Compounds Extraction
Alias Detection
Paraphrase Rules
text
Chemical Database
LD-based data
Conversion of Structured Databases
Link list
Integration of informations
LOD4ALL frontend
Similar chemical compounds

Figure 1: An overview of our system.

By using a chemical compound name recognizer based on an NE recognition method, we expect to obtain new chemical compound names from text with contextual information like “boiling point” and “fusing points” or, prominent words consisting of chemical compounds like “acid” and “methylphenol”.

In addition, we extract usage of chemical compounds from text by a rule-base method. Some of the examples are chemical compounds used as a plasticizer and a surfactant. Another example is usage of chemical compounds as replacement candidates. For example, ‘Dioctyl phthalate’ is used as an alternative of ‘Diisononyl phthalate’ for a plasticizer.

3. Identification of Aliases of Chemical Compound Names

Aliases of chemical compound names are possibly used instead of their full spellings [5]. For example, information of chemical compounds is included as different names in different databases and text such as Nikkaji [4], PubChem [6], ChEMBL [7], CPCat [9], DBPedia [11], patents and scientific papers.

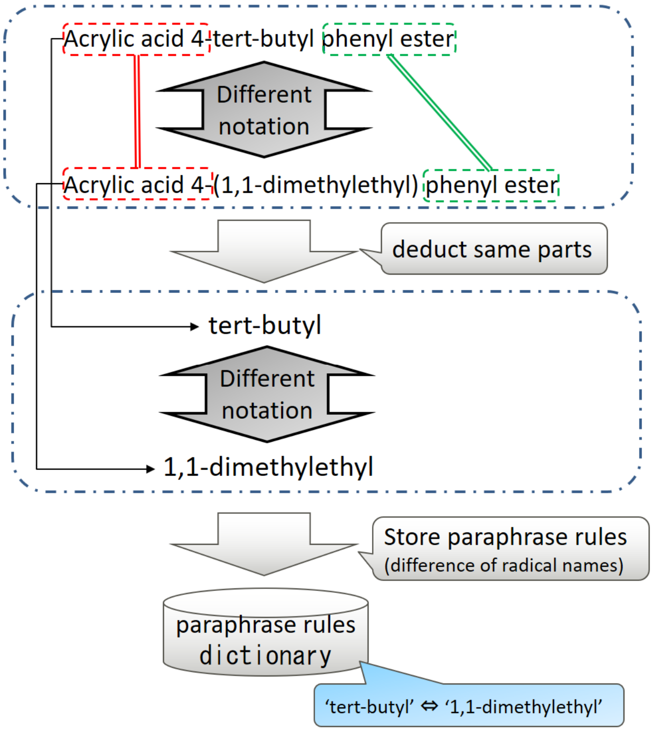

Acrylic acid 4-tert-butyl phenyl ester
Different notation
Acrylic acid 4-(1,1-dimethylethyl) phenyl ester
deduct same parts
tert-butyl
Different notation
1,1-dimethylethyl
Store paraphrase rules (difference of radical names)
paraphrase rules dictionary
‘tert-butyl’ ⇔ ‘1,1-dimethylethyl’

Figure 2: An extraction of a paraphrase rule.

To identify aliases of chemical compounds, we compiled paraphrasing rules from Nikkaji [4]. In order to obtain paraphrasing rules, we split chemical compound names into constituents of structures based on IUPAC nomenclature [5]. Then we extract paraphrase rules by comparing parts of the structure of a chemical compound name with its alias.

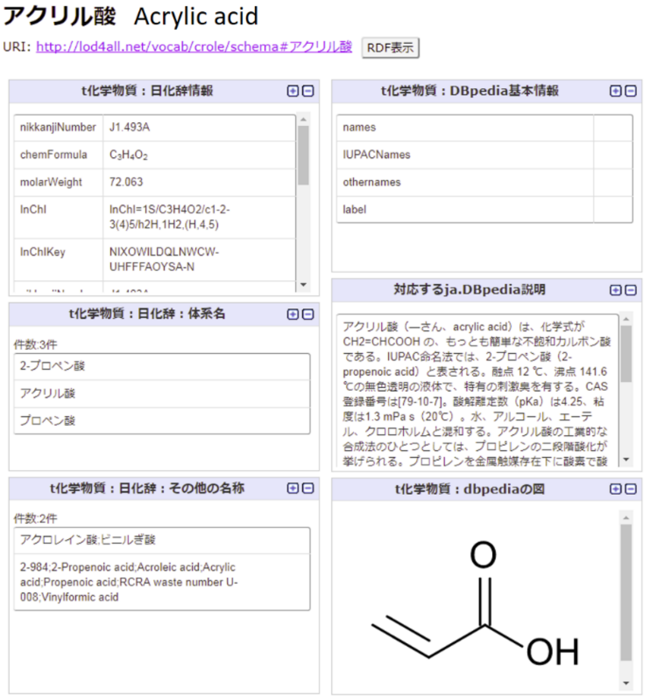
We describe a method to get paraphrase rules by using ‘Acrylic acid 4 - tert - butylphenyl ester’ and its alias ‘Acrylic acid 4 - (1,1 - dimethylethyl)phenyl ester’.

Figure 2 shows an example of an extraction of a paraphrase rule from the above two chemical compound names. At first, we split ‘Acrylic acid 4 - tert - butylphenyl ester’ into ‘Acrylic acid’, ‘tert - butyl’ and ‘phenyl ester’. In a similar way, ‘Acrylic acid 4 - (1,1 - dimethylethyl)phenyl ester’ is split into ‘Acrylic acid’, ‘(1,1 - dimethylethyl)’ and ‘phenyl ester’.

Next, we deduct the common parts of the both chemical compound names. Finally, we extract the pair of ‘tert - butyl’ and ‘1,1 - dimethylethyl’ as a paraphrase rule. We use a dictionary to split chemical compound names into parts of structures.

We expect that a possibility to access useful information would increase by using these rules.

For example, ‘2 - (p - Tolyl)ethanol’ has not been recorded on Nikkaji, However, ‘2 - (4 - Methylphenyl)ethanol’ that is created by paraphrase rule has been recorded.



URI: http://lod4all.net/vocab/crole/schema#アクリル酸

t化学物質：日化辞情報

nikkanjiNumber	J1.493A
chemFormula	$C_3H_4O_2$
molarWeight	72.063
InChI	InChI=1S/C3H4O2/c1-2-3(4)5/h2H,1H2,(H,4,5)
InChIKey	NIXOWILDQLNWCW-UHFFFAOYSA-N

t化学物質：DBpedia基本情報

names	
IUPACNames	
othernames	
label	

t化学物質：日化辞：体系名

件数:3件

- 2-プロペン酸
- アクリル酸
- プロペン酸

対応するja.DBpedia説明

アクリル酸（―さん、acrylic acid）は、化学式がCH2=CHCOOH の、もっとも簡単な不飽和カルボン酸である。IUPAC命名法では、2-プロペン酸（2-propenoic acid）と表される。融点 12 ℃、沸点 141.6 ℃の無色透明の液体で、特有の刺激臭を有する。CAS登録番号は[79-10-7]。酸解離定数（pKa）は4.25、粘度は1.3 mPa s（20℃）。水、アルコール、エーテル、クロロホルムと混和する。アクリル酸の工業的な合成法のひとつとしては、プロピレンの二段階酸化が挙げられる。プロピレンを金属触媒存在下に酸素で酸

t化学物質：日化辞：その他の名称

件数:2件

- アクロレイン酸;ビニルぎ酸
- 2-984;2-Propenoic acid;Acroleic acid;Acrylic acid;Propenoic acid;RCRA waste number U-008;Vinylformic acid

t化学物質：dbpediaの図

Figure 4: Integration different data by LD-based information.

4. Visualization of Chemical Compounds

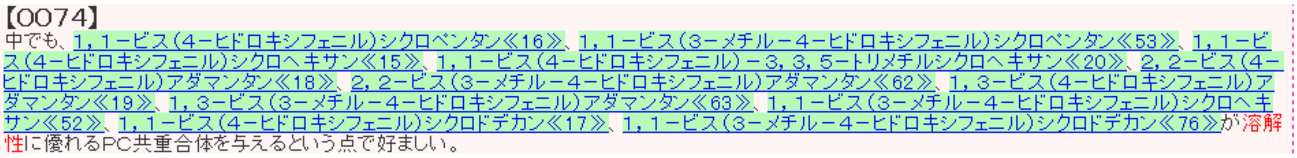
4.1 Highlighting Chemical Compounds

Many chemical patents include sequences of chemical compound names. It is difficult for us to recognize differences between chemical compounds in patent documents. Therefore, a high skill and a large amount of time are required to readers. In order to alleviate such problem, chemical compound names are highlighted on the screen in our system as shown in Figure 3. In addition, each chemical compound has links to the visualization of the following sections. Therefore, users can access additional information by clicking the links in a current text.

4.2 Integrated Data Representation

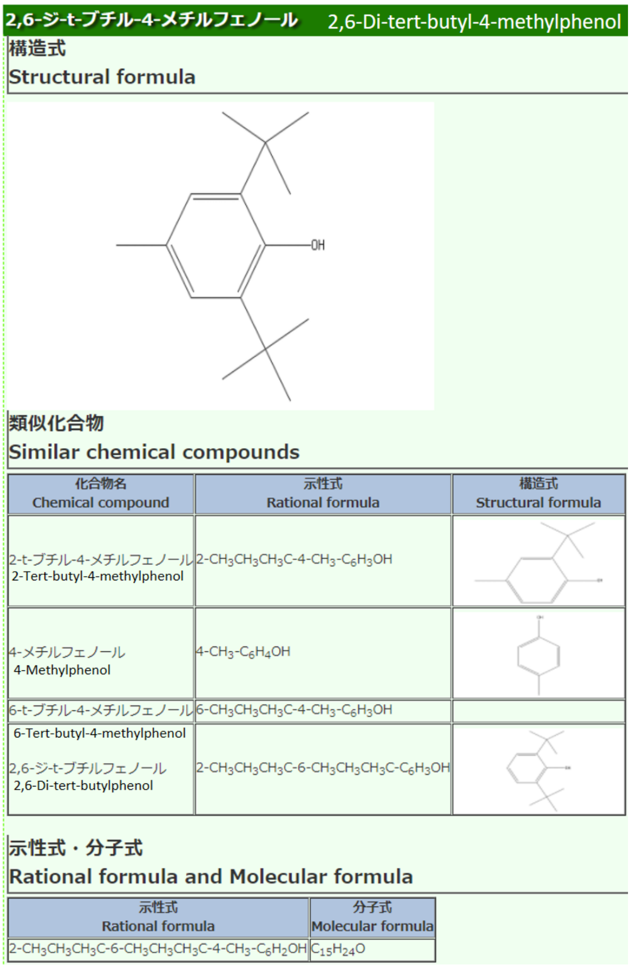
Figure 4 is an example of an integrated visualization of LD-based knowledge of acrylic acid. The visualization includes information of databases such as figures of chemical compounds and information extracted from text as described in Section 2.

The visualization is realized as follows. At first, we extract chemical compound names from documents and create different notations by using paraphrase rules. The processed information is embedded as links in text. Then, if a link is clicked, a search is conducted with extracted names and created notations with paraphrasing. Finally, with the search results, an integrated data representation as shown in Figure 4 is generated.



【0074】
中でも、1，1－ビス（4－ヒドロキシフェニル）シクロペンタン《16》、1，1－ビス（3－メチル－4－ヒドロキシフェニル）シクロペンタン《53》、1，1－ビス（4－ヒドロキシフェニル）シクロヘキサン《15》、1，1－ビス（4－ヒドロキシフェニル）－3，3，5－トリメチルシクロヘキサン《20》、2，2－ビス（4－ヒドロキシフェニル）アダマンタン《18》、2，2－ビス（3－メチル－4－ヒドロキシフェニル）アダマンタン《62》、1，3－ビス（4－ヒドロキシフェニル）アダマンタン《19》、1，3－ビス（3－メチル－4－ヒドロキシフェニル）アダマンタン《63》、1，1－ビス（3－メチル－4－ヒドロキシフェニル）シクロヘキサン《52》、1，1－ビス（4－ヒドロキシフェニル）シクロドデカン《17》、1，1－ビス（3－メチル－4－ヒドロキシフェニル）シクロドデカン《76》が溶解性に優れるPC共重合体を与えるという点で好ましい。

Figure 3: An example of many chemical compounds written in a patent document.



2,6-ジ-t-ブチル-4-メチルフェノール 2,6-Di-tert-butyl-4-methylphenol

構造式
Structural formula

類似化合物
Similar chemical compounds

化合物名 Chemical compound	示性式 Rational formula	構造式 Structural formula
2-t-ブチル-4-メチルフェノール 2-Tert-butyl-4-methylphenol	2-$CH_3CH_3CH_3C$-4-CH_3-C_6H_3OH	
4-メチルフェノール 4-Methylphenol	4-CH_3-C_6H_4OH	
6-t-ブチル-4-メチルフェノール	6-$CH_3CH_3CH_3C$-4-CH_3-C_6H_3OH	
6-Tert-butyl-4-methylphenol 2,6-ジ-t-ブチルフェノール 2,6-Di-tert-butylphenol	2-$CH_3CH_3CH_3C$-6-$CH_3CH_3CH_3C$-C_6H_3OH	

示性式・分子式
Rational formula and Molecular formula

示性式 Rational formula	分子式 Molecular formula
2-$CH_3CH_3CH_3C$-6-$CH_3CH_3CH_3C$-4-CH_3-C_6H_2OH	$C_{15}H_{24}O$

Figure 5: An example of Visualization of Relevant Chemical Compounds.

4.3 Relevant Chemical Compounds

If we can identify the same chemical compounds with different notations by using paraphrasing rules, we can get information from different databases that register the same chemical compounds with different notations. However, the new chemical compounds that have not been registered in databases cannot be found.

In order to help users to understand new chemical compounds, we present additional information about chemical compounds that have similar structures with a given chemical compound.

Figure 5 depicts a set of similar chemical compounds of "2,6-Di-tert-butyl-4-methylphenol". This chemical compounds are recorded in the database and overall structures are displayed. In addition, chemical compounds that have the same parts of structures are displayed as similar chemical compounds.

Table 1: Evaluation of effectiveness of paraprasing

Number of chemical compound names	36
recorded on Nikkaji	12
covered by using paraphrase rule	4
similar chemical compounds	2

By displaying similar chemical compounds, users can infer the structure of a given chemical compound even if it is not registered in databases.

Chemical compounds consist of substituents, which are constituents of structures. Therefore, we can know the structures of chemical compounds by splitting a chemical compound name into substituent names. Furthermore, we can create similar chemical compounds of a given chemical compound by deducting some substituents from the given chemical compound name. Chemical compound names created by deducting substituents of a chemical compound may be recorded in chemical databases. As a result, we can get clues about the structure information of a new chemical compound that have not been registered in databases.

4.4 Analysis of Chemical Compound Names

Figure 6 is an example of a table representation. Eleven chemical compounds extracted from Figure 3 are listed. A combination of some substituents and some cores creates these variations.

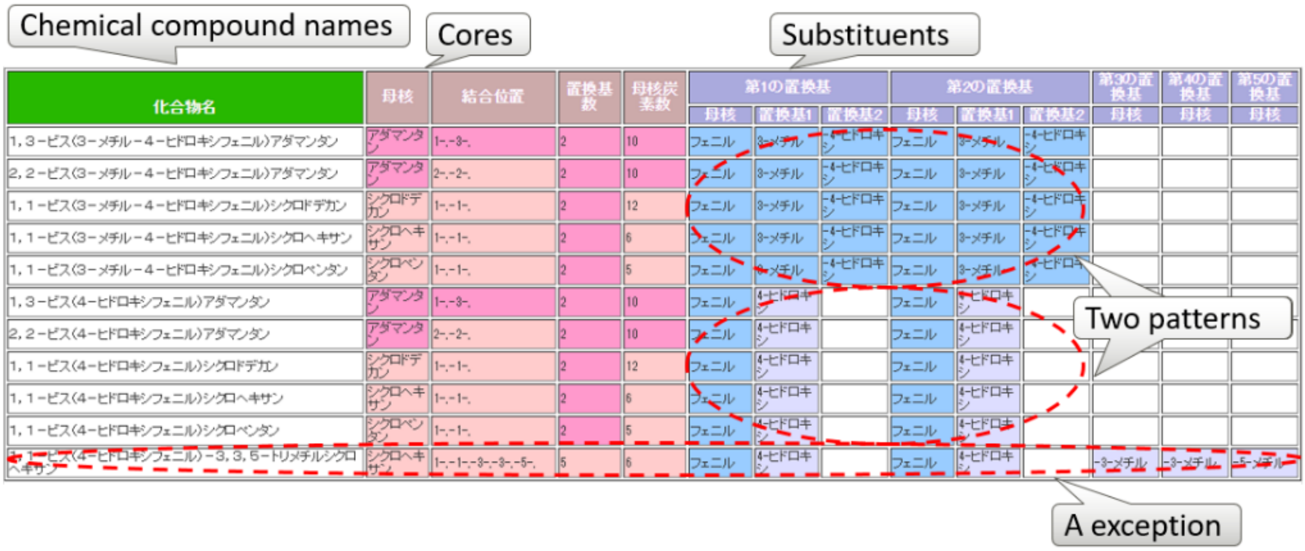
We analyze a hierarchically connection of constituents of structures. In addition, we convert relationships between cores and substituents of chemical compounds into a table representation.

This example shows that there are two patterns in their chemical compounds. And they contain one exception. By displaying difference of chemical compounds, users can realize an overview of them.

5. Evaluation

We evaluated our paraphrasing method, which is one of the key components of our system, with 36 chemical compounds written in the patent document (P2014-263456, paragraph [0017]). The evaluation was done whether the paraphrase contributed to discovery of the same and relevant chemical compounds in databases.

Table 1 shows the experimental result. Twelve chemical compounds out of 36 were recorded in Nikkaji. Four chemical compounds out of 24 were converted into recorded notations by paraphrase rules. Two chemical compounds out of the remaining got similar chemical compounds.



化合物名	母核	結合位置	置換基数	母核炭素数	第1の置換基			第2の置換基			第3の置換基	第4の置換基	第5の置換基
					母核	置換基1	置換基2	母核	置換基1	置換基2	母核	母核	母核
1，3－ビス(3－メチル－4－ヒドロキシフェニル)アダマンタン	アダマンタン	1-,-3-,	2	10	フェニル	3-メチル	-4-ヒドロキシ	フェニル	3-メチル	-4-ヒドロキシ			
2，2－ビス(3－メチル－4－ヒドロキシフェニル)アダマンタン	アダマンタン	2-,-2-,	2	10	フェニル	3-メチル	-4-ヒドロキシ	フェニル	3-メチル	-4-ヒドロキシ			
1，1－ビス(3－メチル－4－ヒドロキシフェニル)シクロドデカン	シクロドデカン	1-,-1-,	2	12	フェニル	3-メチル	-4-ヒドロキシ	フェニル	3-メチル	-4-ヒドロキシ			
1，1－ビス(3－メチル－4－ヒドロキシフェニル)シクロヘキサン	シクロヘキサン	1-,-1-,	2	6	フェニル	3-メチル	-4-ヒドロキシ	フェニル	3-メチル	-4-ヒドロキシ			
1，1－ビス(3－メチル－4－ヒドロキシフェニル)シクロペンタン	シクロペンタン	1-,-1-,	2	5	フェニル	3-メチル	-4-ヒドロキシ	フェニル	3-メチル	-4-ヒドロキシ			
1，3－ビス(4－ヒドロキシフェニル)アダマンタン	アダマンタン	1-,-3-,	2	10	フェニル	4-ヒドロキシ		フェニル	4-ヒドロキシ				
2，2－ビス(4－ヒドロキシフェニル)アダマンタン	アダマンタン	2-,-2-,	2	10	フェニル	4-ヒドロキシ		フェニル	4-ヒドロキシ				
1，1－ビス(4－ヒドロキシフェニル)シクロドデカン	シクロドデカン	1-,-1-,	2	12	フェニル	4-ヒドロキシ		フェニル	4-ヒドロキシ				
1，1－ビス(4－ヒドロキシフェニル)シクロヘキサン	シクロヘキサン	1-,-1-,	2	6	フェニル	4-ヒドロキシ		フェニル	4-ヒドロキシ				
1，1－ビス(4－ヒドロキシフェニル)シクロペンタン	シクロペンタン	1-,-1-,	2	5	フェニル	4-ヒドロキシ		フェニル	4-ヒドロキシ				
1，1－ビス(4－ヒドロキシフェニル)－3，3，5－トリメチルシクロヘキサン	シクロヘキサン	1-,-1-,-3-,-3-,-5-,	5	6	フェニル	4-ヒドロキシ		フェニル	4-ヒドロキシ		-3-メチル	-3-メチル	-5-メチル

Figure 6: An example of table of chemical compounds.

From these results, we see difficulty to cover all chemical compounds by only databases. "4,4'-dihydroxy-3,3'-dimethylphenyl ether" and "4,4'-dihydroxy-3,3'-dimethyldiphenylsulfide" were not covered. However, the constituents of structures, which are 'hydroxy', 'methyl', 'diphenyl ether' and 'diphenylsulfide', had been recorded in chemical databases. In this paper, we created only similar chemical compounds, created by deducting one substituent from the chemical compounds. Therefore, there may be many relevant chemical compounds that we can not find.

On the other hand, even if similar chemical compounds become different from original chemical compounds, users are still difficult to estimate overall structures. Therefore, we provide chemical formulas (Rational formula, Molecular formula) like Figure 5 to estimate overall structure.

6. Conclusion

This paper proposes a visualization system for chemical compound information extracted from Japanese texts and chemical compound databases. This system enables users to get information of chemical compounds not only from existing databases but also from text.

In the future work, we would like to extend our system to languages other than Japanese.

REFERENCES

1. CAS web page: https://www.cas.org/index
2. Tomoya Iwakura. A Named Entity Recognition Method using Rules Acquired from Unlabeled Data. Proc. of RANLP'11. Pp. 170—177.
3. Mintz, Mike and Bills, Steven and Snow, Rion and Jurafsky, Dan. Distant Supervision for Relation Extraction without Labeled Data. Proc. Of ACL'09. pp. 1003—1011. 2009.
4. Nikkaji: http://dbarchive.biosciencedbc.jp/jp/nikkaji/desc.html
5. IUPAC: https://iupac.org/what-we-do/nomenclature/
6. The PubChem Project: https://pubchem.ncbi.nlm.nih.gov/
7. ChEMBL: https://www.ebi.ac.uk/chembl/
8. ChemSpider: http://www.chemspider.com/
9. Chemical and Product Categories (CPCat):

   https://www.epa.gov/chemical-research/chemical-and-product-categories-cpcat
10. LOD4ALL frontend: http://lod4all.net/frontend/
11. DBpedia: http://wiki.dbpedia.org/