2009 10th International Conference on Document Analysis and Recognition

The Understanding and Structure Analyzing for Online Handwritten Chemical Formulas

Xin Wang, Guangshun Shi, Jufeng Yang
Institute of Machine Intelligence, Nankai University, Tianjin, China
{ xinwang, gsshi, jfyang }@imi.nankai.edu.cn

Abstract

In this paper, we propose a novel approach for understanding and analyzing the online handwritten chemical formulas. With the structural characteristics, semantic rules, and more importantly grammatical rules, the analyzing process is divided into 3 levels: formula level, molecule level, and text level. A formal description of the chemical formula based-on the grammatical rules is summed up and applied to the analyzing process which generates grammar spanning graphs from the analyzed result step-by-step, and that is used for the further structure representation and data retrieval. Our work, as an important component of applying mobile computing research in education, is proved effective and promising.

1. Introduction

Chemical formula has been used widely. However, the speciality of casual writing determines that precise comprehension of those formulas become a challenge in applying Mobile Computing technology to chemistry education [1]. Because it may contain complicated molecule structure which expresses specific meaning in 2D plane, analyzing and recording chemical formula structures especially the molecule structure and combination relationship of the symbols become more important compared to recognizing isolated symbols.

There are several storage methods of representing the chemical structure, Brookhaven Protein Data Bank format, the Molecular Design Limited (MDL) MOLFILE format, the Standard Molecular Data (SMD) format [2], where the mathematical methods such as adjacency matrices and connection tables are reflected. A Chemical Markup Language (CML) based-on XML has been developed recently. However, all the chemical formats cannot easily, directly, and conveniently be used for representing recognition apparatus of a molecule, because none of them records the layout information and abbreviated information. Moreover, these formats can't be used directly for the whole formula structure analyzing process. On the other hand, there are related research on recognition and structure analyzing for chemical formulas. Boyer and others [3,4] proposed an overall solution of structure storage and recognition process on chemical formula documents. Ramel [5] presented a method which could recognize the graphic entities in handwritten chemical expressions. They reconstructed the chemical structures from the graphics and recognized using the features extracted from these structures. Ouyang [6] presented a sketch recognition system designed to interpret hand drawn chemical diagrams. However, these works only focused on off-line processing, and the research on recognizing online handwritten chemical formula(OHCF) is still lack.

Based on the characteristics of OHCF, we conclude a useful knowledge-base, which is composed of layout information, timing information, and vital grammatical rules. Moreover, a novel approach for understanding and analyzing OHCFs has been raised. Meanwhile grammatical rules has been formatted and described into 3 levels. After the analyzing process on an OHCF, the result of structure and content combined with online information and independent users writing habits have been perfectly preserved, and that has a significant effect for improving user experiences.

2. Characteristic of OHCFs

The structure of handwritten chemical formulas can be divided into three levels: formula level, molecule level and text level. As showed in Fig.1, both formula level and text level are reflected as the one-dimensional structure in the meaning of grammar, and molecule level belongs to two-dimensional structure. Meanwhile, the symbols in handwritten chemical formula can be separated as operator, text and bond. Operator belongs to the formula level; bond belongs to the molecule level, both of which are the key symbols in structure analyzing of corresponding level, respectively.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.70

Table 1. Symbol set of OHCFs

Text	1 2 3 4 5 6 7 8 9 0 A B C D E F G H I K L M N O P R S T U V W X Y Z a b c d e f g h i k l m n o p r s t u y	
Bond	— ═ ≡ ○(Ring)	
Operator	1 generator	↔ → =
	2 separator	+
	3 State-assisted Symbols	↑ ↓ + -
	4 Conditions	Δ C °
	5 Structure-assisted symbols	( )

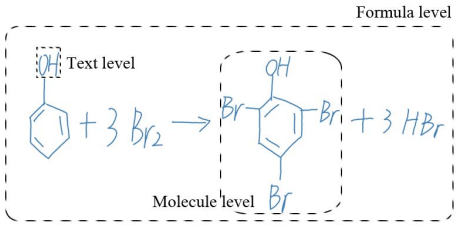

Figure 1. Different levels of chemical formula

Among the formula level and text level, the grammatical meaning between the symbols can be described as the one-dimension combination, the definition scope of which is restricted in its neighbor symbols or structures. That means the current symbols can only be merged with its neighbor symbols or structure to form a new structure. In this one-dimension structure of grammar, there might be some particular one-dimension instances, such as superscript, subscript, and up-down combination, which we call the layout analyzing function to deal with. Moreover, these particularities of the grammar in the special Layout structure also give us a big favor in analyzing current level structure.

In the molecule level, symbols are arranged in the two-dimension flat plane, the connection and combination of symbol groups and bond represent the specific chemistry meaning which is shown in Fig 2.

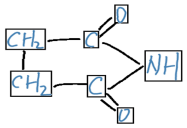

Figure 2. Molecule expression structure.

The symbol group in the rectangular areas is defined as a text group and the line segments are the bond. We can use the text group as a vertex and bond as an edge to construct a graph which is called Molecule Spanning Graph (MSG). The intersection of two bonds is actually a hidden Carbon dot which is invisible in the graph and we will record it and separate it from other Carbon points. MSG is stored using adjacency matrix. This definition is very important in the areas of recording the meaning of substance itself, analyzing process of document structure and representing of analyzing results.

During the analyzing of molecule level, two layout problems should be considered: bond layout analyzing and the detection of connections between symbol groups and bonds. Firstly, a start and an end of every bond stroke, added with the sharp point in the bond stroke are called end points, all of which are marked as three kinds: free ends, junctions (where 3 or more bonds meet), connections (2 bonds meeting). Secondly, every free end will connect with one specific atom of a text group. If there's no text found during the detection, the free end is marked as a hidden carbon.

3. Understanding and structure analyzing for OHCFs

3.1. System Model

We propose an approach for understanding and analyzing the structure of OHCF document, according to the structural characteristics. The approach is based on grammatical rules, combined with layout information, timing information, and semantic rules. The structure of OHCF is analyzed following the 3-level process below.

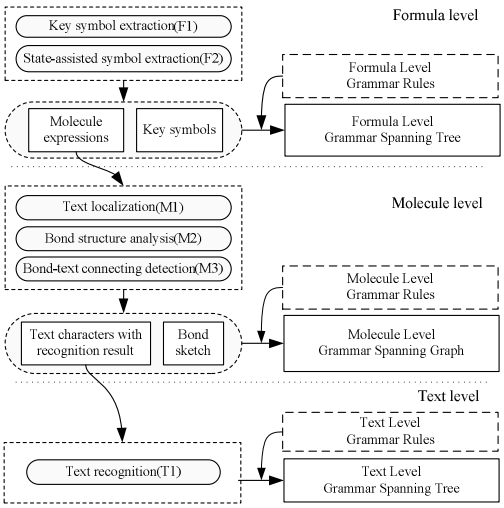

Figure 3. Understanding for OHCFs

Step 1: Layout analysis, get all the formula level symbols and sub-expressions (reactants and products), use formula level productions to analyze the current set, generate a grammar spanning tree.

Step 2: For each molecule level expression, get all the molecule level symbols and sub-expressions (text groups), analyze the layout information and adjacent

relation of the bonds and the text groups, and generate a compound adjacency matrix.

Step 3: For each text group, use text level productions to analyze the current set to obtain the text symbols' property and the combination relationship of them.

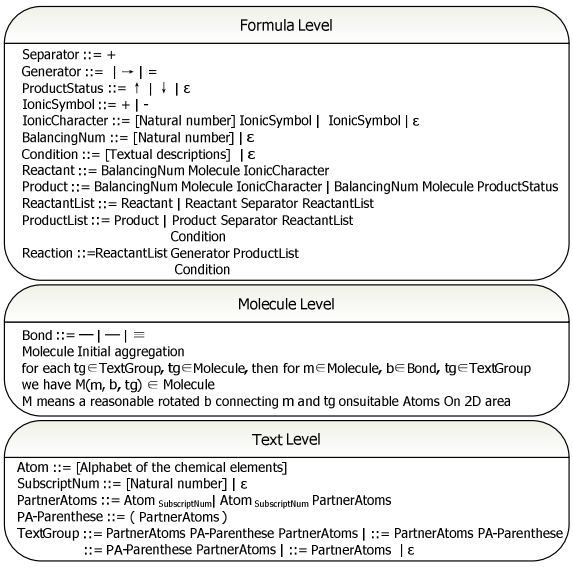

Figure 4. Grammar rules production

3.2. Formal description of grammatical rules

The structure analysis process is motivated by 3 levels grammatical rules. Key symbols from different level are extracted in different orders according to the level of certain grammatical rule; the process of analyzing structures from different level consider certain level grammatical rules as a core guidance; the store and representation of the analyzing result from different level are directly linked to the certain level grammatical rules: formula level and text level analysis performances a grammar spanning tree, while a molecule spanning graph is generated from analysis.

The formula grammar we use can be formally defined as a 5-tuple: $G = (T, N, P, M, S)$

(1) T is a set of terminal symbols
(2) N is a set of non-terminal symbols
(3) P is a finite set of productions
(4) M is a reasonable mapping in N
(5) S is the start of $G(S \in N)$

This grammar is composed of context-free rewriting rules as indicated in the figure 4, and they are designed with ambiguity elimination.

3.3. Formula level structure analysis

Key Operator Extraction. Firstly, we extract the formula level key operators which have been mentioned as level 1, 2 operators in table 1. The extraction process start with analyzing the horizontal baseline （HBL）of a reaction which has the most symbols or connected components on it in the horizon direction. After the cursory segmentation, we narrow the range of the key operator candidates on the HBL, with checking their characteristics such as: the number of strokes, the size, the layout information, etc. Then we use our chemistry symbol recognition module to confirm the final result.

State-assisted Symbol Extraction. After extracting key operator, we use the key operator spread method to extract the state-assisted symbols which have been mentioned as level 3 operators in table 1. State-assisted symbol only appears around the reactant and the product, in other words, they appears in the adjacent structure of the key operators. Therefore, we can get their grammatical attributes on the corresponding formula level by checking out these symbols through our chemistry symbol recognition module. Of course, this assisted operator extraction process is combined with the molecule level analysis for further validation.

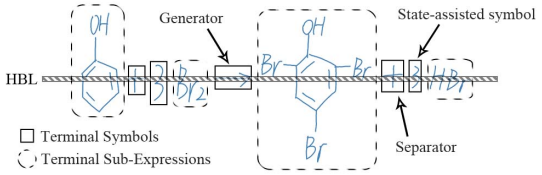

Figure 5. Formula level layout structure analysis

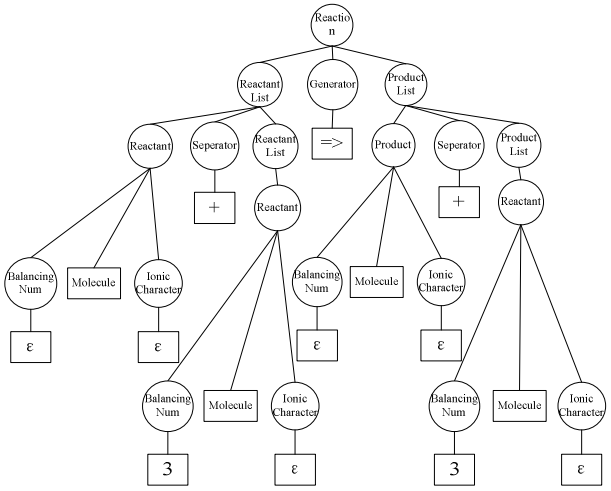

Figure 6. Formula level grammar spanning tree

Formula level structure analysis. We analyze the formula level structure with all the key operators and state-assisted operators extracted, and a formula level spanning tree is generated applying the following method:

Step 1: Scan and markup all the terminal symbols and terminal sub-expressions in the current level, and generate a terminal sequence S.
Step 2: Select all productions in current level as Set P.
Step 3: Chose the highest level p in Set P which can be used to produce S(only one p can be found because of our ambiguity-eliminated design for the grammatical rules) with the form of $S ::= [t_1..t_n, n_1..n_m]$, apply production p to sequence S, and generate a grammatical spanning tree $S <= (t_1..t_n, n_1..n_m)$.
Step 4: For each S_i, use the routine Step 3 until all the symbols or sub-expressions in S_i are terminal symbols, and get a grammar spanning tree at last.

3.4. Molecule level structure analysis

Text localization. The symbols appearing in the molecule level is limited to bond and text. Separating them from each other effectively improves the accuracy of the structure analyzing result. In order to upgrade the system operating efficiency, text localization in molecule level and text recognition in character level are carried out together in our system. In the molecule level, the size, shape, arrangement of the segments, and other segment features are representative of the texture of molecule structure. Using this information, the model makes a decision and classifies each segment into one of the following categories: text (Specific content will be given by the recognition module), bond, and non-decision. For non-decision segment, studying its size and position in relation to the text segments already localized, leads to a comparison with the context, and certain ambiguities can thus be resolved.

Bond structure analysis. Starting with cleaning up the bonds set separated from the original molecule level structure, including removing duplicated points, eliminating hooks, smoothing data, connecting broken strokes, and handling containing relationship of the ring structure, etc. After these preprocessing, we analyze the composition of the clean bonds Set. Our objective at this stage is to markup the type of every end point on the bonds. All the end points can be divided into 3 categories: the start and the end of a stroke, and the sharp point in a stroke. A sharp point is at the peak of the wavelet where the writing direction has changed. The algorithm for exploring sharp points is developed based on the changed-angles of pen motion [7].

All the end points are detected through the above-mentioned methods. We group these end points into several divisions by their coordinates and some specific rules (like end points in the same stroke cannot be in the same group, etc.) The grouping method for determining whether an end point belongs to a Group A is calculated as follows:

$$Include(x_0, y_0) = \begin{cases} 1, & d_0 < t \max d_x + \partial \\ 0, & else \end{cases} \quad (3.1)$$

Include(x_0, y_0) represents the grouping result for adding p_0 into Group A, where

$$d_k = \sqrt{(x_k - x_a)^2 + (y_k - y_a)^2} \quad (3.2)$$

$$x_a = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad y_a = \frac{1}{n}\sum_{k=1}^{n} y_k \quad (3.3)$$

Figure 7. shows the specific process.

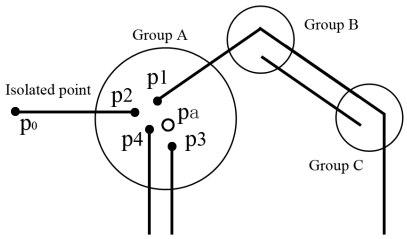

Figure 7. End points grouping

Every end points group are marked up as free end, connections and junctions by the total number of the start point , end point, and sharp point in it.

Bond connection detecting. The text groups are always connected by a bond. We have to detect this situation for the structure analyzing. Free ends are the only concerned type of end points in this stage. The connecting situation is calculated as follow:

$$Con(x_v, y_v) = \begin{cases} 1, & \min x - tW < x_v < \max x + tW \\ & \min y - tH < y_v < \max y + tH \\ 0, & else \end{cases} \quad (3.4)$$

The text group is connected by the bond when Con (x_v,y_v)=1. And t is a threshold here.

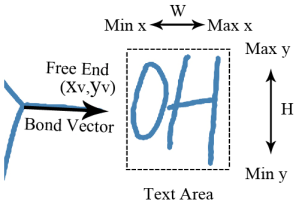

Figure 8. Detecting free end connection

Molecule level structure analysis. After marking up all the end points and detecting every free end connection, we start analyzing molecule level structure to get grammar spanning graph and adjacency matrix.
Initial Status: Bond Set B = Φ, Text Group Set T = Φ, adjacency matrix M = 0.
Step 1: Scan out every each bond b from the symbols, add b into Set B, markup all end points of the bonds.

Step 2: For every junctions and connections in end points, add a t of a hidden carbon into Set T.
Step 3: For every each text group t, add t into Set T.
Step 4: Init the size and content of M with every t in T.
Step 5: Scan the both end points of every bond, if a free end is detected, then find the connected character group, else if a junction or a connection is detected, then find the related hidden carbon, add the content of the bond into the corresponding position in M. At last, a adjacency matrix is generated.

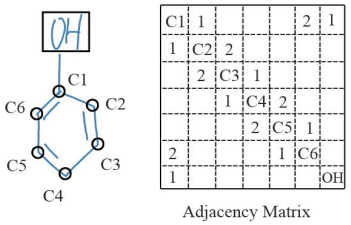

Figure 9. Molecule level structure analysis

3.5. Text level structure analysis

Text Symbols Recognition. In our text recognition model, we built an HMM for each symbol, and obtained top-3 accuracy of 98.7% on a dataset containing 5,670 train samples and 2,016 test samples.

Text level analyzing. Text level analyzing is a similar process with the formula level. We use the corresponding level method to analyze the text level structure with the text level grammatical rules. Also, a text level grammar spanning tree is generated. Thus, we can facilitate access the grammar attributes and combination relationship of all the text symbols.

4. Experiment and Conclusion

In this paper, we presented a novel approach for understanding and analyzing online handwritten chemical formulas, and proposed an efficient and creative processing model.

Table 2. Experimental results

Level	Step	Target Symbols	Num	Acc(%)
Formula	F1	key symbols	60	98.3%
	F2	state-assisted	48	100%
Molecule	M1	text area	313	100%
	M2	bond end grouping	431	98.8%
	M3	free end-text conn	117	100%
Text	T1	text recognition	313	98.7%

25 formulas were chosen randomly from the 1250 formula data samples to test the model proposed above, among which we have 484 symbols including 25 generators, 35 separators, 48 state-assisted symbols, 313 text symbols, 43 rings, 66 independent bonds, 431 bond end groups, and 117 free end-text connections. The result is shown in table 2.

We have shown that the approach is capable of understanding not only inorganic chemistry notations but also common organic ones. And it gives a good solution for a relatively high degree of complex organic molecules showing as the reactant or product. In addition, the analyzing results retained the various online information of the original handwritten document by writers' intentions and habits. Also, the structure analysis results can be conveniently converted to visual editing tools or other standard format such as CML, used for subsequent analysis and operation.

We will make in depth study in the future based on the work of this paper, with a focus on better optimizing the algorithm in the key processing parts, and establishing a more complete grammatical rule-based recognition validation system.

Acknowledgement

This project is supported by Microsoft Research Asia Mobile Computing in Education Theme.

References

[1] J. F. Yang, G. S. Shi, and Q. R. Wang, "Recognition of On-line Handwritten Chemical Expressions", *Proceedings of the International Joint Conference on Neural Networks*, 2008, pp. 2360-2365.
[2] Barnard and M. John, "Draft Specification for Revised Version of the Standard Molecular Data (SMD) Format", *Journal of Chemical Information and Computer Sciences*, vol. 30(1), 1990, pp.81-96.
[3] Casey, R. Boyer, S. Healey, P. Miller, A. Oudot, and B. Zilles, "Optical recognition of chemical graphics", *Proceedings of the Second International Conference on Document Analysis and Recognition*, 1993, pp. 627-631.
[4] Joe R. McDaniel and Jason R. Balmuth, "Automatic interpretation of chemical structure diagrams", *Graphics Recognition Methods and Applications*, vol. 1072, 1996, pp. 148-158.
[5] Jean-Yves Ramel, Guillaume Boissier, and Hubert Emptoz, "Automatic Reading of Handwritten Chemical Formulas from a Structural Representation of the Image", *Proceedings of the International Conference on Document Analysis and Recognition*, 1999, pp. 83-86.
[6] T. Y. Ouyang and R. Davis, "Recognition of Hand Drawn Chemical Diagrams", *Proceedings of the national conference on artificial intelligence*, 2007, pp. 846-852.
[7] A. Nair and C. G. Leedham, "Preprocessing of Line Codes for Online Recognition Purposes", *Electronics Letters*. vol. 27, 1991, pp. 1-2.