An Experiment on Mining Chemical Reaction Databases

Sandra Berasaluce, Claude Laurenço, Amedeo Napoli, Gilles Niel

An Experiment on Mining Chemical Reaction Databases

Sandra Berasaluce[1,2,3], Claude Laurenço[1,2], Amedeo Napoli[3] and Gilles Niel[1]

1 *Laboratoire des Systèmes d'Information Chimique, LSIC – ENSCM, 8, rue de l'Ecole Normale, 34296 Montpellier*

Email: Sandra.Berasaluce@wanadoo.fr

2 *LSIC and Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, LIRMM, 161, rue Ada, 34392 Montpellier*

Email: cl@lirmm.fr

3 *Laboratoire LOrrain de Recherche en Informatique et ses Applications, LORIA, BP 239, 54506 Vandœuvre-lès-Nancy*

Email: Amedeo.Napoli@loria.fr

1 *LSIC – ENSCM, 8, rue de l'Ecole Normale, 34296 Montpellier*

Email: niel@cit.enscm.fr

ABSTRACT *In this paper, we present an experiment on knowledge discovery in chemical reaction databases. Chemical reactions are the main elements on which relies synthesis in organic chemistry, and this is why chemical reactions databases are of first importance. From a problem-solving process perspective, synthesis in organic chemistry must be considered at several levels of abstraction: mainly a strategic level where general synthesis methods are involved, and a tactic level where actual chemical reactions are applied. The research work presented in this paper is aimed at discovering general synthesis methods from chemical reaction databases in order to design generic and reusable synthesis plans. The knowledge discovery process relies on frequent levelwise itemset search and association rule extraction, but also on chemical knowledge involved within every step of the knowledge discovery process. Moreover, the overall process is supervised by an expert of the domain.*

KEYWORDS knowledge discovery, data mining, frequent level-wise itemset search, association rule, knowledge-based system.

1. Introduction

In this paper, we present an experiment on the application of knowledge discovery algorithms for mining chemical reaction databases. Chemical reactions are the main elements on which relies synthesis in organic chemistry, and this is why chemical reaction databases are of first importance. From a problem-solving process perspective, synthesis in organic chemistry must be considered

at several levels of abstraction: mainly a strategic level where general synthesis methods are involved, and a tactic level where actual chemical reactions are applied. The research work presented in this paper is aimed at discovering general synthesis methods from chemical reaction databases in order to design generic and reusable synthesis plans. This can be understood in the following way: mining reaction databases at the tactic level for finding synthesis methods at the strategic level. This knowledge discovery process relies on the one hand on mining algorithms, i.e. frequent levelwise itemset search and association rule extraction, and, on the other hand, on domain knowledge, that is involved at every step of the knowledge discovery process.

This research work is carried out within a long-term project for designing chemical information systems whose goal is to help a chemist building a synthesis plan [**6,9**]. Synthesis planning is mainly based on retrosynthesis, i.e. a goal-directed problem-solving approach, where the target molecule is iteratively transformed by applying reactions for obtaining simpler fragments, until finding accessible starting materials. For a given target molecule, a huge number of starting materials and reactions may exist, e.g. thousands of commercially available chemical compounds. Thus, exploring all the possible pathways issued from a target molecule leads to a combinatorial explosion, and needs a strategy for choosing reaction sequences to be used within the planning process.

At present, reaction database management systems are the most useful tools for helping the chemist in synthesis planning. One aspect of this research is to study how data mining techniques may contribute to knowledge extraction from reaction databases, and beyond that, to the structuring of these databases and the improvement in their querying. This paper presents a preliminary experiment carried on two commercial reaction databases[1] using frequent itemset search and association rule extraction [**1,7,8**]. This study is original and novel within the domain of organic synthesis planning, and is of first importance, with respect to chemical researches. Regarding the knowledge discovery research, we stress the fact that knowledge extraction in an application domain has to be guided by knowledge domain if substantial results have to be obtained.

The paper is organized as follows. First, we briefly introduce the chemical context, describing the synthesis problem. Then, we detail the application of frequent itemsets search and association rule extraction to the data. Finally, we conclude the paper with a discussion on the first results of this experiment and on the research perspectives.

2. The chemical context

Actually, the main questions for the synthesis chemist are related to chemical families to which a target molecule belongs, and to the reactions or sequence

[1]Supplied by Molecular Design Ltd – MDL, `http://www.mdli.com`

of reactions building structural patterns, to be used for building these families. Two main categories of reactions may be distinguished: reactions building the skeleton of a molecule –the arrangement of carbon atoms on which relies a molecule–, and reactions changing the functionality of a molecule, i.e. changing a function into another function. Hereafer, we are mainly interested in reactions changing the functionality, and especially in the following question: what are the reactions allowing the transformation of a function $\mathtt{F_i}$ into a function $\mathtt{F_j}$?

The experiment reported hereafter has been carried out on two reaction databases, namely the "Organic Syntheses" database ORGSYN-2000 including 5486 records, and the "Journal of Synthetic Methods" database JSM-2002 including 75291 records. The purpose of the preprocessing step of data mining is to improve the quality of the selected data by cleaning and normalizing the data. The information items in databases such as ORGSYN-2000 and JSM-2002 may be seen as a collection of records, where every record contains one chemical equation involving structural information, that can be read, according to the reaction model, as the transformation of an initial state –or the set of reactants– into a final state –or the set of products– associated with an atom-to-atom mapping between the initial and final states (see fig. 1).

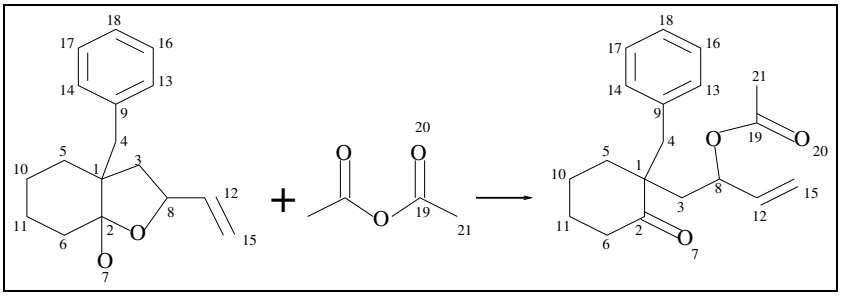

Figure 1: The structural information on a reaction with the associated atom-to-atom mapping (reaction #13426 in the JSM-2002 database).

In our framework, data preprocessing has mainly consisted in exporting and analyzing the structural information recorded in the databases for extracting and for representing the functional transformations in a target format that has been processed afterwards. The considered transformations are functional modifications, functional addition and deletion, i.e. adding or deleting a function. The reactions have been considered at an abstract level, the so-called block level as shown in figure 2.

In the following, we discuss the whole process of chemical reaction databases manipulation, involving data transformation and data mining, for retrieving and organizing chemical reaction databases.

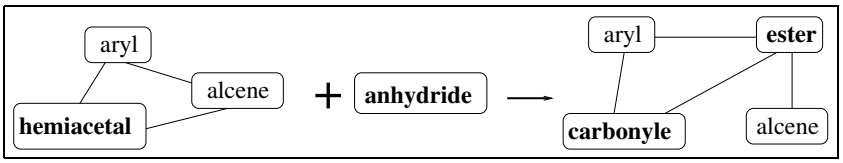

Figure 2: The representation of the reaction #13426 in the JSM-2002 database at the block level.

3. Knowledge discovery in reaction databases

The knowledge discovery process in chemical reaction databases is considered as an interactive and iterative experimental process. An expert of the data domain, called hereafter the analyst, plays a central role in this process since he is in charge of controlling all the steps of the process (as discussed in e.g. [**4**]). According to given synthesis objectives, the analyst selects first the data to be analyzed, applies data mining modules for extracting knowledge units from data, and finally interprets and validates the units having a sufficient plausibility for being reused. Hence, in our approach, the knowledge discovery process is, first of all, guided by the analyst and domain knowledge. The knowledge discovery process itself is based on frequents itemsets search and association rules extraction. Practically, the Close and the Pascal algorithms have been used for data processing [**1,7,8**].

3.1 The modeling of reactions for knowledge discovery

The RESYN-ASSISTANT system [**9**] has been used for recognizing the building blocks of reactions. Based on the atom-to-atom mapping, the system establishes the correspondence between the recognized blocks of the same nature, and determines their role in the reaction. A function may be present in a reactant, in a product, or in both. In the last case, the function is unchanged. In the two other cases, the function in the reactant is destroyed, or the function in the product is formed. During a reaction, either one or more reactant functions may contribute to form the functions in the products. At the end of the preprocessing step, the information obtained by the recognition process is incorporated into the representation of the reaction.

For allowing the application of the algorithms Close and Pascal for frequent itemsets search, the data on reaction have to be transformed into a Boolean table. Thus, the representation of a molecule as a composition of functional blocks cannot be used in a straightforward way. Moreover, a reaction can be considered from two main points of view (see figure 3 showing the correspondences between blocks in the reaction #13426 of the JSM-2002 database, and figure 4 for the associated Boolean table):

Reactions/Blocks	Destroyed	Formed	Unchanged
no correspondence one entry	anhydride, hemiacetal	ester, carbonyle	alcene, aryle
with correspondence two entries	anhydride, hemiacetal hemiacetal	ester carbonyle	alcene, aryle alcene, aryle

Figure 3: The preparation of the original data: the correspondence bteween blocks within the reaction #13426 of the JSM-2002 database.

Reactions/Blocks	Destroyed		Formed		Unchanged	
	anhydride	hemiacetal	carbonyle	ester	alcene	aryle
Single entry **R**	x	x	x	x	x	x
entry $\mathbf{R}_1$	x	x		x	x	x
entry $\mathbf{R}_2$		x	x		x	x

Figure 4: Boolean transformation of the data respectively not taking and taking into account the atom mapping, i.e. one single line or two lines in the Boolean table.

- a global point of view on the functionality interchanges leads to consider a single entry R corresponding to an analyzed reaction, to which is associated a list of properties, i.e. formed and/or destroyed and/or unchanged functions,
- a specific point of view on the functionality transformations that is based on the consideration of a number of different entries $\mathtt{R}_\mathtt{k}$ corresponding to the different functions being formed.

Both correspondences have been used during the experiment, and in both cases, the spatial information on the graph structure of the molecules is lost.

3.2 The search for itemsets and the extraction of association rules

The Close and the Pascal algorithms have been applied to Boolean tables for generating first itemsets, i.e. sets of functions (with an associated support), and then association rules. The study of the extracted frequent itemsets may be done with different points of view. Firstly, studying frequent itemsets of length

2 or 3 enables the analyst to determine basic relations between functions. For example searching for a formed functions $\mathtt{F_f}$ ($\_\mathtt{f}$ for formed) deriving from a broken function $\mathtt{F_d}$ ($\_\mathtt{d}$ for destroyed) leads to the study of the itemsets $\mathtt{F_d} \sqcap \mathtt{F_f}$, where the symbol $\sqcap$ stands for the conjunction of items or functions. Moreover, searching for formed functions $\mathtt{F_f}$ deriving from two destroyed functions $\mathtt{F_{d1}}$ and $\mathtt{F_{d2}}$ leads to the study of the itemsets $\mathtt{F_{d1}} \sqcap \mathtt{F_{d2}} \sqcap \mathtt{F_f}$. In some cases, a reaction may depend on functions present in both reactants and products that remain unchanged ($\_\mathtt{u}$ for unchanged) during the reaction application, leading to the study of frequent itemsets such as $\mathtt{F_f} \sqcap \mathtt{F_u} \sqcap \mathtt{F_d}$. This kind of itemsets can be searched for extracting a "protection function" supposed to be stable under given experimental conditions.

The extraction of association rules gives a complementary perspective on the knowledge extraction process. For example, searching for the more frequent ways to form a function $\mathtt{F_f}$ from a function $\mathtt{F_d}$ leads to the study of rules such as $\mathtt{F_f} \longrightarrow \mathtt{F_d}$: indeed, this rule has to be read in a retrosynthetic way, i.e. if the function $\mathtt{F_f}$ is formed then this means that the function $\mathtt{F_d}$ is destroyed. Again, this rule can be generalized in the following way: determining how a function $\mathtt{F_f}$ is formed from two destroyed functions $\mathtt{F_{d1}}$ and $\mathtt{F_{d2}}$, knowing say that the function $\mathtt{F_{d1}}$ is actually destroyed, leads to the study of the association rules such as $\mathtt{F_f} \sqcap \mathtt{F_{d1}} \longrightarrow \mathtt{F_{d2}}$. It must be noticed that for the sake of simplicity, the examples have been kept formal here. Concrete examples can be found either in [**2,3**].

4. Discussion on the knowledge discovery process on chemical reactions

A whole set of results of the application of the data mining process on the ORGSYN-2000 and JSM-2002 databases is given in [**3**]. It can be noticed that only a few research works hold on the application of data mining methods on reaction databases. A study on the lattice-based classification of dynamic knowledge units has been a valuable source of inspiration for the present work [**5**], leading to the division of functions in three categories, formed, destroyed, and unchanged. A number of topics are discussed hereafter regarding the experiment presented in this paper.

The abstraction of reactions within blocks and the separation in three kinds of blocks, namely formed, destroyed, and unchanged blocks, is one of the most original idea in that research work, that is responsible of the good results that have been obtained. This idea of the separation into three families may be reused in other contexts involving dynamic data. However, the transformation into a Boolean table has led to a loss of information, e.g. the connection information on reactions and blocks.

Frequent items or association rules are generic elements that can be used

either to index (and thus organize) reactions or to retrieve reactions. Termed in another way, this means that frequent itemsets or extracted association rules may be in certain cases considered as a kind of meta-data giving meta-information on the bases that are under study.

Knowledge is used at every step of the knowledge extraction process, e.g. the coupling of the knowledge extraction process with the RESYN-ASSISTANT system, and domain ontologies such as the function ontologies, the role of the analyst,... Indeed, and this is one of the major lesson of this experiment: the knowledge discovery process in a specific domain such as organic synthesis has to be knowledge-intensive, and has to be guided by domain knowledge, and an analyst as well, for obtaining substantial results. The role of the analyst includes fixing the thresholds, and interpreting of the results. The thresholds must be chosen in function of the objectives of the analyst, and in function of the content of the databases (it can be noticed that a threshold of 1% for an item support means that for a thousand of reactions, ten may form a reaction family, and this is not a bad hypothesis).

Moreover, the use of data mining methods such as frequent itemsets search or association rule extraction has proven to be useful, and has provided encouraging results. It could be interesting to test other (symbolic) data mining methods, e.g. namely sequential patterns, closed itemsets and icebergs, olap technology, cluster analysis, or Bayesian network classification, knowing that numerical methods such as hidden Markov models or neural networks are not adapted to the kind of data that are considered in our experiment.

5. Conclusion

In this paper, we have presented an experiment on knowledge discovery in chemical reaction databases. Two databases have been deeply studied and mined, namely the orgsyn-2000 and the jsm-2002 databases, using frequent levelwise itemset search and association rule extraction. The main topic of interest in the reactions is related with functionality interchanges. Thus, the reactions in the databases have been abstracted in terms of three kinds of building blocks for molecules involved in the reactions, namely formed, destroyed and unchanged blocks. This categorization of blocks has been the basis for building the Boolean tables on which data mining algorithms such as Close and Pascal have been applied. From a chemical point of view, the results are very encouraging, and provide a set of meta-data for organizing and retrieving chemical reaction according to given synthesis objectives. From a knowledge discovery point of view, a number of questions can be discussed, such as the value of the thresholds (usually lower than in marketing analysis), on the processing of the data, and on the interpretation of the results. Moreover, two major elements have to be pointed out, and can be reused in other contexts :

the categorization of dynamic data such as reactions into three families, here, formed, destroyed and unchanged functions, and the use of knowledge at every stage of the knowledge discovery process. Indeed, in a domain such as organic synthesis, the knowledge discovery process has to be fully guided by domain knowledge, and the analyst, an expert of the domain, as well. There are a number of research perspectives following the present work, including the adaptation of sequential pattern algorithms to chemical reactions, taking actually into account the structures of the molecules involved in reactions, and working in the spirit of concept analysis for lattice-based classification of the data.

REFERENCES

[1] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal,*Mining frequent patterns with counting inference, ACM SIGKDD Explorations*, 2(2):66–75, 2000.

[2] S. Berasaluce,Fouille de données at acquisition de connaissances à partir de bases de données de réactions chimiques, Thèse de chimie informatique et théorique, Université Henri Poincaré – Nancy 1, 2002.

[3] S. Berasaluce, C. Laurenço, A. Napoli, and G. Niel,*Data mining in reaction databases: extraction of knowledge on chemical functionality transformations,* Technical Report A04-R-049, LORIA, Nancy, 2004.

[4] R.J. Brachman and T. Anand,*The Process of Knowledge Discovery in Databases,* In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 37–57, Menlo Park, California, 1996. AAAI Press / MIT Press.

[5] B. Ganter and S. Rudolph.,*Formal Concept Analysis Methods for Dynamic Conceptual Graphs,* In H.S.Delugach and G. Stumme, editors, *Conceptual Structures: Broadening the Base – 9th International Conference on Conceptual Structures, ICCS-2001, Stanford*, Lecture Notes in Artificial Intelligence 2120, pages 143–156. Springer, Berlin, 2001.

[6] A. Napoli, C. Laurenço, and R. Ducournau,*An object-based representation system for organic synthesis planning, International Journal of Human-Computer Studies*, 41(1/2):5–32, 1994.

[7] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal,*Pruning closed itemset lattices for association rules, International Journal of Information Systems*, 24(1):25–46, 1999.

[8] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal,*Discovering frequent closed itemsets for association rules,* In C. Beeri and P. Buneman, editors, *Database Theory – ICDT'99 Proceedings, 7th International Conference, Jerusalem, Israel*, Lecture Notes in Computer Science 1540, pages 398–416. Springer, 1999.

[9] P. Vismara and C. Laurenço,*An abstract representation for molecular graphs, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 51:343–366, 2000.