Engineering Applications of Artificial Intelligence 39 (2015) 215–222

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A lattice-based approach for chemical structural retrieval

Peng Tang [a,*], Siu Cheung Hui [a], Alvis C.M. Fong [b]

[a] School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798
[b] School of Computing Science, University of Glasgow, Lilybank Gardens, Glasgow, Scotland

ARTICLE INFO

Article history:
Received 6 October 2014
Received in revised form
9 December 2014
Accepted 11 December 2014
Available online 9 January 2015

Keywords:
Formal concept analysis
Chemical structural similarity retrieval
Latticed-based information retrieval
Chemical concept lattice

ABSTRACT

Searching for chemical structures with similar structural and functional information of organic chemicals is an important part of the drug discovery process. However, the current chemical structural retrieval methods have focused mainly on finding chemicals with similar structures to the input chemical structural query, and tend to ignore the functional features which are important for determining the chemical property and activity of the chemicals. In this paper, we propose a lattice-based approach for chemical structural retrieval. The proposed lattice-based approach is based on Formal Concept Analysis. It retrieves chemical structures that have functional groups and interactions between functional groups similar to the chemical structural query. The performance of the proposed lattice-based approach is evaluated and its promising performance results have shown that the proposed approach is effective for chemical structural retrieval.

1. Introduction

In many drug discovery projects, it is often required to search for similar chemical structures of drug-like compounds that are worthy for further synthetic or biological investigation (Sheridan and Kearsley, 2002). A number of chemical structural similarity retrieval methods (ChemInfoSystem, 2014; Brown and Martin, 1996; Ewing et al., 2006; Fechner et al., 2005; Hagadone, 1992; Hert et al., 2004; Rarey and Dixon, 1998; Schuur et al., 1996) have been proposed for chemical structural retrieval. These chemical structural retrieval methods are mainly based on the assumption that chemicals which are globally similar in structure to each other are more likely to have similar chemical properties and activities. Therefore, these methods have focused mainly on finding chemicals with similar structures to the input structural query.

However, the current chemical structural retrieval methods tend to overfit on the structure of the chemicals and ignore the functional features such as functional groups and interactions between functional groups that are hidden inside the chemical structure. Functional features in a chemical structure can determine the chemical property and activity of the chemicals. Tang et al. (2012) proposed an approach to extract functional features from chemical structures. The extracted features are then used for chemical structural retrieval based on a Vector Space Model. This paper proposes a lattice-based approach for chemical structural retrieval. The proposed lattice-based approach is based on Formal Concept Analysis (FCA) (Ganter et al., 1999). It retrieves chemical structures that have functional groups and interactions between functional groups similar to the chemical structural query.

Formal Concept Analysis has been applied to different information retrieval applications (Koester, 2006; Poshyvanyk, 2007; Muangon and Intakosum, 2009; Nauer and Noussaint, 2009). FCA-based information retrieval techniques consider the inter-document similarity and support context-dependent ranking of the documents. In addition, it does not suffer from the drawbacks of the vocabulary problem (Furnas et al., 1987) and heuristics strategies (Carpineto and Romano, 2000) of conventional information retrieval systems such as best matching and clustering-based ranking.

In this paper, we will present the proposed lattice-based approach for chemical structural retrieval and its performance evaluation. The proposed approach has made the following contributions: (1) chemical functional features including functional groups and their interactions are extracted for chemical structural retrieval; and (2) formal concept analysis is proposed for chemical structural retrieval. Experimental results show that the proposed lattice-based approach achieves promising performance, and outperforms state-of-the-art chemical structural retrieval methods.

The rest of the paper is organized as follows. Section 2 reviews the related work on chemical structural retrieval and formal concept analysis. Section 3 presents the proposed lattice-based approach for chemical structural retrieval. Section 4 gives the performance evaluation of the proposed lattice-based approach. Finally, Section 5 concludes the paper.

* Corresponding author.

http://dx.doi.org/10.1016/j.engappai.2014.12.006
0952-1976/

2. Related work

2.1. Chemical structural retrieval

ChemSpider (Pence and Williams, 2010) is a free online aggregated chemical structure database providing fast text and structural search access to over 26 million structures from hundreds of data sources. PubChem (Wang et al., 2009) is a free database of chemical structures of small organic molecules and information on its biological activities. ChEMBLdb (Gaulton et al., 2012) is a manually curated chemical database of bio-active molecules with drug-like properties. eMolecules (eMolecules, 2014) is a search engine for chemical molecules supplied by commercial suppliers. All these publicly available chemical search systems support database query retrieval for chemical structures including exact search, substructure search and similarity search. However, the structured query search method may tend to overfit the chemical structures and fail to recognize chemicals that are more similar in chemical functionalities.

Apart from structured query search, similarity retrieval methods have been under development for decades to find relevant chemical structures. To compute the similarities between chemical structures, different chemical structural representations and similarity measures have been proposed. Currently, there are three major chemical similarity retrieval methods, namely superposition-based similarity methods, histogram-based similarity methods and descriptor-based similarity methods.

The superposition-based similarity methods map one chemical structure onto another. It treats two molecular structures as graphs and aims to find the correspondence between the atoms in one structure and the atoms in another structure (Hagadone, 1992; Rarey and Dixon, 1998). Histogram-based methods (Schuur et al., 1996) transform chemical structures into one or more spectra or histograms, and then calculate the overlapping between the histograms of two chemical structures for similarity measurement. Descriptor-based methods are most popular for chemical structural similarity search. A molecule is represented as a set of descriptors or numbers. As such, a molecule can be considered as a point in a multidimensional descriptor space. This method is computational efficient. However, in contrast to the superposition-based methods, the equivalence of sub-structures (or parts) between one molecule and another is lost.

In particular, the fingerprinting (ChemInfoSystem, 2014; Brown and Martin, 1996; Ewing et al., 2006; Fechner et al., 2005; Hert et al., 2004) method, which is the most popular descriptor-based method, uses a set of user-defined 2D substructures and their frequencies to represent molecules. The substructures are used as the descriptors. In the fingerprinting method, only the presence or absence of a descriptor is captured. The substructure descriptors are considered as the fingerprints of the chemical structure. The similarity is defined based on the number of descriptors that the two molecules have in common and normalized by the number of descriptors in each molecule. The fingerprinting method is efficient because it is computationally inexpensive to compare two lists of pre-computed descriptors.

Different from the current chemical similarity retrieval methods which focused mainly on finding chemicals with similar structures to the input structural query, Tang et al. (2012) proposed a functional-based chemical structural retrieval approach to retrieve functionally relevant chemical structures based on functional groups and their interactions. In the approach, the Vector Space Model was used. A new *ctf-icf* weighting scheme which is defined as the product of chemical term frequency *ctf* and inverse chemical frequency *icf*, and a new chemical similarity measure were proposed for the retrieval and ranking of the similar chemical structures.

2.2. Formal Concept Analysis

Formal Concept Analysis (FCA) is a well established technique in mathematics and computer science. FCA can be used to identify meaningful groupings of objects that share common attributes, and analyze the hierarchies of these groupings (Aswani Kumar et al., 2012; Graña, 2012). FCA is particularly suitable for the information retrieval task (Carpineto and Romano, 2004b). In conventional document retrieval systems, the relationship between documents and terms is represented by a term-document matrix. Each element of the matrix is a binary relation between a term and a document. It specifies whether the term is occurred in the document or not. Godin et al. (1989) developed a textual information retrieval system based on a term-document lattice. The proposed system was compared with other retrieval methods and found that its performance is better than hierarchical classification (Godin et al., 1993).

Cheung and Vogel (2005) proposed to apply Formal Concept Analysis to transform a term-document matrix to a formal context of a lattice, and map each document to an object and each term to an attribute. Similarly, Messai et al. (2006) proposed a lattice-based information retrieval system called BR-Explore. It constructs a concept lattice based on a term-document matrix. Based on the terms extracted from a query, the query will be inserted into the lattice. The documents which share the same terms with the query will be considered as relevant to the query. Rajapakse and Denham (2006) proposed to extract unit-concepts from documents to construct a concept lattice. Unit-concept is an object–attribute pair which represents the occurrence of an attribute in an object. Relevance is evaluated by comparing the nodes of the query lattice with the nodes of the document lattice.

3. Proposed lattice-based approach

Chemical Functional Groups (CFGs) are specific groups of atoms within molecules that determine the characteristics of chemical

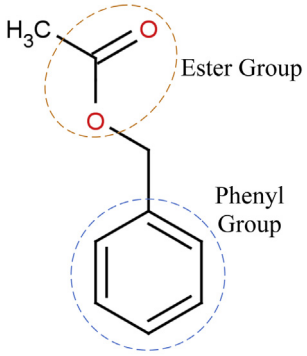

Fig. 1. Benzyl acetate.

Table 1
Common basic functional groups.

Functional group	Structural formula	SMILES	Functional group	Structural formula	SMILES
Alkenyl	CC	CC	Alkynyl	C≡C	C#C
Fluoro	−F	F	Iodo	−I	I
Chloro	−Cl	Cl	Bromo	−Br	Br
Hydroxyl	−OH	O	Ether	−O−	O
Carbonyl	−C(=O)−	CO	Aldehyde	−C(=O)H	CO
Amine	−N−	N	Primary Imine	−CNH	CN
Secondary Imine	−CN−	CN	Nitrile	−C≡N	C#N
Sulfhydryl	−SH	S	Sulfide	−S−	S
Phosphino	−P−	P			