ИНФОРМАЦИО́ННО-ПОИСКО́ВЫЕ СИСТЕ́МЫ

Авторы: О. В. Кириллова

ИНФОРМАЦИО́ННО-ПОИСКО́ВЫЕ СИСТЕ́МЫ (ИПС) по химии, автоматизированные поисковые системы, реализованные на средствах электронной вычислит. техники и предназначенные для сбора, поиска, обработки, хранения и выдачи пользователям химич. информации по заданным критериям. Многообразие объектов химии и сложность её языка привели к выделению ИПС по химии в самостоят. класс информац. систем. ИПС представляет собой совокупность информационно-поискового языка (языков), программных средств и правил перевода текстов на этот язык (индексирования), обеспечения поиска и критериев соответствия. Материализованное представление об ИПС включает в себя информац. массивы, их носители (магнитные, оптические и т. п.), программные и технич. средства. Осн. информац. массивами ИПС являются базы данных (БД) и банки данных (БнД), а также информац. массивы специализир. интернет-систем. В зависимости от характера информации, включаемой в БД, различают документографич., или документальные, ИПС (ДИПС; содержат библиографич. описания, ключевые слова, рефераты статей из журналов и сборников, монографий, патентов, стандартов, отчётов о н.-и. работе и т. д.); фактографич. ИПС (ФИПС; содержат унифициров. фактографич. данные об объектах предметной области, свойствах материалов и веществ, структурные формулы соединений, уравнения химич. реакций, данные физико-химич. исследований, сведения об областях применения, стоимости и др.); документально-фактографич. (интегрированные) ИПС. Большинство ИПС по химии представлены в Интернете или распространяются на твёрдых магнитных носителях (CD-ROM, DVD и др.).

Ключевой информац. составляющей в ИПС по химии является объект – химич. вещество. Информация о химич. веществах может быть представлена в виде простых текстовых характеристик (название, синонимы, молекулярная формула и др.), числовых значений конкретных свойств, рефератов и полнотекстовых статей, графических и/или табличных, оптич., УФ-, ИК-, ЭПР-, ЯМР- и др. спектров или зависимостей, плоских и трёхмерных изображений и т. д. Главной и специфической является информация о структуре молекул химич. соединений. Осн. способом представления структуры молекулы химич. соединения является структурная формула. Структурная формула – гл. источник информации о структуре молекулы конкретного химич. соединения и его однозначный идентификатор. Способом представления химич. информации в памяти ЭВМ и вне её служат форматы данных. Наиболее крупными разработчиками форматов данных в химии считаются следующие зарубежные информац. центры: Chemical Abstracts Service (CAS), Molecular Design Limited (MDL), Daylight, TRIPOS. Осн. форматами данных являются линейные коды (SLN/SMILES, SYBYL, Висвессера и др.), текстовые файлы ASCII (MOL, SDF, RDF и др.), файлы форматов XML (CML).

Поиск (идентификация) химич. веществ в ИПС может производиться на основе сравнения разл. характеристик и свойств: регистрац. кодов, названий и синонимов, молекулярных данных (масса, формула, структура), библиографич. данных и др. Ключевой является задача поиска химич. веществ по структурной формуле. Для реализации поиска структурной химич. информации применяются специально разработанные информационно-поисковые языки (классификаторы, справочники, словари, тезаурусы, рубрикаторы и т. п.) и алгоритмы

обработки, решаемые как стандартными средствами систем управления БД, так и специализир. программными приложениями. В общем случае алгоритмы обработки структурной химич. информации основаны на обработке молекулярных графов. Разработка первых алгоритмов, решающих проблему изоморфизма графа, относится к 1950–60-м гг. Эти алгоритмы были основаны на применении поатомного сопоставления молекулярных графов химич. веществ (алгоритмы Рэя и Кирча, Ульмана). Алгоритм Ульмана основан на использовании рекурсивного алгоритма, методики глубокого просмотра и булевских матриц, хранящих состояние сопоставления графов; алгоритм Ульмана адаптирован для поиска по формулам Маркуша. Для увеличения производительности (скорости) поиска в больших ИПС используется метод предварит. фильтрации на основе разл. дескрипторов молекулярной структуры (топологич. индексы, физико-химич. дескрипторы, структурные дескрипторы), преобразования структуры в редуцированный граф посредством замещения определённых больших структур спец. метками. На завершающем этапе допускается применение алгоритмов поатомного сопоставления.

Структурная информация в совр. ИПС (БД) может быть представлена как в виде обычных химич. структур для индивидуальных соединений, так и в виде формул Маркуша для обобщённых структур (характеризуются переменными молекулярными заместителями-радикалами, переменными местами замещения и количеством групп замещений, общими и частными названиями заместителей). Преимущественное распространение формулы Маркуша получили в патентной области, где их использование позволяет существенно расширить и защитить права авторов новых химич. соединений. Формулы Маркуша могут применяться как гибкие классификаторы химич. веществ. Существуют три крупные общедоступные (коммерческие) ИПС для поиска Маркуш-структур: Derwent World Patents Index (WPI), CAS MARPAT, INPI Merged Markush Service (MMS) в сотрудничестве с Derwent Information Ltd. (Markush DARC).

ИПС различаются охватом (числом) обрабатываемых источников (наполнением), структурой данных, функциональными и поисковыми возможностями. Наиболее крупные ДИПС по химии: в России – БД Химия ВИНИТИ РАН (Всерос. ин-т науч. и технич. информации РАН; пополнение – 120 тыс. документов/год); за рубежом – БД Chemical Abstracts (Chemical Abstracts Service, США; пополнение – 1 млн. документов/год, относится к политематич. БД), БД Index Chemicus (Thomson Scientific, США; включает ок. 2,5 млн. химич. структур, опубликованных в лит-ре с 1993). Существенная часть химич. информации включена в политематич. БД– Science Citation Index (SCI, Thomson Scientific, США), SCOPUS (Elsevier, Нидерланды), отчётов о НИОКР и диссертаций (Всерос. науч.-технич. информац. центр, Россия), патентные БД (Федеральный ин-т пром. собственности, Derwent, Europatent). Осн. фактографич. ресурсом для сопровождения химико-синтетич. исследований является ИПС CrossFire Beilstein, в области металлоорганич. и неорганич. химии – CrossFire Gmelin (включает лит-ру начиная с 1772). Значит. часть зарубежных БД по химии представлена в Интернете в онлайновых системах-агрегаторах электронных информац. ресурсов: STN International, DIALOG и др., а также на платформах издательств (MDL, Elsevier, CAS и др.).

Наиболее крупные и специализир. полнотекстовые ИПС – коллекции по химии, наукам о материалах и др. смежным областям на интернет-платформах Elsevier (ScienceDirect), Wiley (полнотекстовые и фактографич. ИПС: e-EROS – Encyclopedia of Reagents for Organic Synthesis, e-Proxemis – Organic Reactions, Organic Syntheses, Organic-Chemical Drugs и др.). Значительно число проблемно ориентированных и узкотематич. химич. ИПС.

В Интернете ДИПС получают новое развитие, выходят за рамки БД и преобразуются в информац. системы с расширенным спектром функциональных возможностей и услуг. Примеры новых типов ИПС – SciFinder, ChemNet

и др. Большие возможности поиска БД и химич. информации в целом предоставляют поисковые интернет-системы Scirus, Google, MSN и др., для Рунета – Yandex, Rambler, Google.ru. Большое количество химич. информации находится в Интернете в открытом доступе.

Значит. часть ИПС имеет печатные аналоги в виде реферативных журналов и индексов: реферативный ж. «Химия» ВИНИТИ РАН, Chemical Abstracts, ChemInform (Wiley), Chemistry Citation Index (Thomson) и др.

Литература

Лит.: Влэдуц Г. Э., Гейвандов Э. А. Автоматизированные информационные системы для химии. М., 1974; Гордон А., Форд Р. Спутник химика. Физико-химические свойства, методики, библиография. М., 1976.