ИПМ им.М.В.Келдыша РАН

Абрау-2019 • Труды конференции

Д.А. Варламов, В.Е. Туманов

Извлечение экспериментальных данных по химической кинетике из открытых источников в сети Интернет

Рекомендуемая форма библиографической ссылки

Варламов Д.А., Туманов В.Е. Извлечение экспериментальных данных по химической кинетике из открытых источников в сети Интернет // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23-28 сентября 2019 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2019. — С. 188-197. — URL: http://keldysh.ru/abrau/2019/theses/46.pdf doi:10.20948/abrau-2019-46

Размещена также презентация к докладу

Извлечение экспериментальных данных по химической кинетике из открытых источников в сети Интернет

Д.А. Варламов[1], В.Е. Туманов[2]

[1] Институт экспериментальной минералогии РАН
[2] Институт проблем микроэлектроники и особо чистых материалов РАН

Аннотация. В статье рассматривается процесс интеллектуального поиска и извлечения экспериментальных данных по химической кинетике из открытых источников в сети Интернет. Предложен подход к организации вышеуказанного процесса, который включает в себя последовательность следующих этапов: формирование электронного корпуса документов электронный версий научных журналов, как в открытом доступе, так и коммерческих журналов, разработка онтологии предметной области для построения тезауруса, интеллектуальный поиск и формирование корпуса библиографических ссылок, интеллектуальный анализ библиографических ссылок на основе кластерного анализа, скачивание, преобразование в тестовый формат и классификация документов байесовской нейронной сетью, извлечение данных и их складирование. Разработанные авторами технологии позволяют работать с англоязычными и русскоязычными текстами. Алгоритм автоматической классификаций использует три метода кластеризации. Автоматическая классификация публикаций использует специально построенный трехуровневый информационный базис, который разбивает множества статей по характеру содержания. На основе полученного разбиения строится корпус электронных документов, который содержит экспериментальные данные по реакционной способности органических соединений в химических реакциях. Извлеченные характеристики химических реакций сохраняются в разработанной базе данных для дальнейшего использования.

Ключевые слова: интеллектуальный поиск данных, извлечение данных, открытые источники Интернет, электронный корпус документов, предметная онтология, искусственные нейронные сети, химическая кинетика, экспериментальные данные.

Extraction of experimental data on chemical kinetics from open sources in the Internet

D.A. Varlamov[1], V.E. Tumanov[2]

[1] Institute of Experimental Mineralogy Russian Academy of Sciences

[2] *Institute of Microelectronics Technology and High-Purity Materials of the Russian Academy of Sciences*

Abstract. In article process of intellectual search and extraction of experimental data on chemical kinetics from open sources in the Internet is considered. Approach to the organization of the above-stated process which includes the sequence of the following stages is offered: forming of a computer corpus of documents electronic versions of scientific journals, as in open access, and commercial journals, development of a domain ontology for creation of the thesaurus, intellectual search and forming of the electronic corpus of bibliographic links, the intellectual analysis of bibliographic links on the basis of cluster analysis, downloading, conversion to a test format and document classification by a Bayesian neural network, extraction of data and their warehousing. The technologies developed by authors allow to work with English-language and Russian-language texts. The algorithm automatic classifications uses three methods of a clustering. Automatic classification of publications uses specially constructed three-level information basis which breaks sets of articles on the nature of contents. On the basis of the received splitting the case of electronic documents which contains experimental data on reaction activity of organic compounds in chemical reactions is under construction. The retrieved characteristics of chemical reactions remain in the developed database for further use.

Keywords: intellectual data retrieval, data extraction, Internet open sources, computer corpus of documents, subject ontology, artificial neural networks, chemical kinetics, experimental data.

Экспериментальные данные о реакционной способности органических соединений в химических реакциях имеют большое значение как для фундаментальных знаний, в том числе в живых системах, так и для химических и биологических технологий. С помощью эмпирических и теоретических моделей химических реакций они позволяют оценить фундаментальные характеристики органических соединений, такие как энергии диссоциации связей или константы скорости химических реакций. Для химической технологии и органического синтеза они позволяют лучше проектировать и управлять процессами химических превращений.

Таким образом, поиск, сбор, хранение, извлечение, интеллектуальный анализ и публикация экспериментальных данных о реакционной способности органических соединений в электронных коллекциях для последующего использования является актуальной научно-практической задачей.

Целью настоящей работы является исследование возможных схем интеллектуального поиска и извлечения экспериментальных данных о реакционной способности органических соединений из открытых источников в сети Интернет. Разработка программного обеспечения и реализация процесса интеллектуального поиска позволит повысить надежности извлечения информации из известных источников, а автоматизация извлечения данных приведет к снижению трудоемкости актуализации разработанной базы данных

по реакционной способности органических соединений в радикальных реакциях в жидкой фазе.

Вопросам интеллектуального поиска и извлечения данных из электронных документов посвящено большое количество публикаций в различных сферах человеческой деятельности. Текущее состояние данной проблемы в химии и биологии представлено в обзоре [1].

Развитие информационного общества привело к тому, что появилось большое количество электронных источников информации, в том числе электронных коллекций научных и научно-практических публикаций, как в библиографическом исполнении, так и полнотекстовых документов в открытом доступе в сети Интернет. Публикации, содержащие экспериментальные данные о реакционной способности органических соединений, разбросаны по различным электронным ресурсам. С другой стороны, интенсивный рост, как количества электронных источников информации, так и их объема приводит к тому, что ни в одной из поисковых машин нет индексации всех электронных публикаций.

Данный факт приводит к проблеме организации поиска информации. Использование технологий метапоиска позволяет охватить как можно более широкий круг электронных ресурсов. Однако при этом встает вопрос о релевантности полученной информации, так и надежности извлекаемых экспериментальных данных. Проблема релевантности может быть решена через построение предметных онтологий. Проблема надежности при автоматическом извлечении данных из текстов на естественном языке является более сложной для решения. Например, в широко известной у химиков-кинетиков базе кинетических данных NIST (США) [2] приведены данные о реакционной способности органических и неорганических соединений в газовой фазе, полученных как экспериментальными методами, так и путем квантово-химических расчетов. Причем разброс значений для отдельно взятого соединения может достигать несколько порядков. Разработанная в ИПХФ РАН база данных по константам скорости органических соединений в радикальных реакциях в жидкой фазе, которая включает в себя экспериментальные данные, также показывает разброс значений в широком диапазоне для многих органических соединений [3]. Это обстоятельство требует дополнительного изучения источника информации (страна, авторы, организация, импакт - фактор, метод исследования, условия проведения эксперимента и т.д.).

В настоящее время для выбранной предметной области нами предлагается использование методов интеллектуального поиска данных на основе предметной онтологии и ограничение числа используемых электронных ресурсов.

Таким образом, решение задачи поиска и извлечения экспериментальных данных о реакционной способности органических соединений в химических реакциях можно разбить на следующие этапы:

- Формирование электронного корпуса документов;

- Построение онтологии предметной области или тезауруса;
- Организация интеллектуального поиска и формирование ссылок;
- Предварительный анализ содержания документа по ссылкам;
- Скачивание и преобразование документов;
- Классификация документов;
- Извлечение данных и их складирование.

1. Формирование электронного корпуса документов

Достаточно полная классификация электронных источников химической информации приведена в обзоре [1]. Согласно предложенной классификации все источники по характеру доступа можно разбить на две большие группы – коммерческие и находящиеся в открытом доступе. С точки зрения полноты содержания можно выделить четыре группы источников. Из них нас интересуют:

1. «Журналы»:
    - библиографические базы данных с аннотациями;
    - полнотекстовые базы данных статей в журналах;
2. «Оперативные публикации»:
    - научные отчеты;
    - диссертации;
    - материалы конференций;
3. «Электронные коллекции документов»:
    - скомпилированные базы данных электронных документов (монографии, учебная литература и т.п).

Базы данных, содержащие информацию о патентах и их описаний, не включают в себя экспериментальные значения о реакционной способности органических соединений в химических превращениях.

Вышеуказанные электронные ресурсы представляют собой ту основу, которая позволяет сформировать электронный корпус документов, потенциально содержащий информацию о реакционной способности органических соединений, при этом импакт-факторы журналов позволяют судить, в том числе, и о надежности извлекаемых данных.

2. Использование предметных онтологий для построения тезауруса

Построением и применением онтологий в химии, в частности для представления свойств химических реакций, активно занимаются группы зарубежных и российских ученых [4-11]. Результаты исследований, полученные в работах [8-10], позволяют построить тезаурус (как разновидность онтологии), который удобно использовать при формировании запросов к

электронному корпусу документов в автоматическом режиме с помощью программного агента как в [12].

3. Организация интеллектуального поиска, формирование и анализ ссылок

В обзоре [1], монографии [13] и конкретно по реакционной способности в работе [14] рассмотрены проблемы организации интеллектуального поиска в электронных ресурсах сети Интернет.

Сформированные на основе тезауруса поисковые запросы к электронному корпусу документов позволяют создать корпус ссылок на них. Использование методов интеллектуального анализа данных [15] позволяет провести классификацию ссылок и приступить к скачиванию документов в PDF формате.

4. Извлечение и складирование данных

Для дальнейшего анализа содержания документов их целесообразно преобразовать в текстовый формат и классифицировать на основе специального информационного базиса, описывающего как характер работы, так и представленные в ней химические объекты.

Технология, по которой может быть построена такая полуавтоматическая обработка, предложена в работах [16-18]. Для классификации по информационному базису была использована байесовская нейронная сеть[1].

Размещенные в базе данных классифицированные и ранжированные документы позволяют в полуавтоматическом режиме извлечь из публикации информацию о реакционной способности органических соединений в структурированной форме: химическая реакция и ее класс, реагент(ы) и их характеристики (структурные и числовые), условия протекания химического превращения (среда, давление, температурный режим), методы измерения и методы аналитического контроля, специфические характеристики проведения опыта. Реляционная модель данных такой базы данных представлена в работе [19].

5. Извлечение и складирование данных

Предложенная в работе схема организации интеллектуального поиска и извлечения данных о реакционной способности органических соединений в химических превращениях является одним из возможных вариантов решения поставленной задачи. Она основывается на отработанных технологических решениях, сочетание которых приводит к желаемому результату.

Дальнейшие исследования в этом направлении должны учитывать представление электронных документов в различных форматах, исключение

[1] Программный модуль для реализации данной сети был разработан Д.Ю. Лазаревым (будет опубликовано).

приведения документов к одному формату для проведения интеллектуального анализа данных и использование возможностей искусственных нейронных сетей.

6. Заключение

В настоящей работе рассмотрен один из возможных подходов к решению задачи интеллектуального поиска и извлечения экспериментальных данных о реакционной способности органических соединений в химических реакциях.

Предложенный подход включает в себя последовательность следующих этапов: формирование электронного корпуса документов электронный версий научных журналов, как в открытом доступе, так и коммерческих журналов, разработка онтологии предметной области для построения тезауруса, интеллектуальный поиск и формирование корпуса библиографических ссылок, интеллектуальный анализ библиографических ссылок на основе кластерного анализа, скачивание, преобразование в тестовый формат и классификация документов байесовской нейронной сетью, извлечение данных и их складирование.

Разработанные технологии позволяют работать с англоязычными и русскоязычными текстами. Алгоритм автоматической классификаций использует три метода кластеризации. Автоматическая классификация публикаций использует специально построенный трехуровневый информационный базис, который разбивает множества статей по характеру содержания. На основе полученного разбиения строится корпус электронных документов, который содержит экспериментальные данные по реакционной способности органических соединений в химических реакциях. Извлеченные характеристики химических реакций сохраняются в разработанной базе данных для дальнейшего использования.

Литература

1. Krallinger M., Rabal O., Lourenço A., Oyarzabal J., Valencia A. Information Retrieval and Text Mining Technologies for Chemistry // Chem. Rev. 2017, V. 117. P. 7673−7761.
2. NIST Chemistry WebBook. NIST Standard Reference Database Number 69. [Digital resource]. Access mode: http://webbook.nist.gov/chemistry/
3. Tumanov V., Gaifullin G. Subject-oriented science intelligent system on physical chemistry of radical reactions // Modern Advances in Intelligent Systems and Tools (Proc. Of IEA/AIE 2012). Ed. Wei Ding, He Jaing, Moonis Ali, Mingchu Li. Berlin; Heidelberg: Springer. 2012. P. 121-127.
4. Артемьева И.Л., Рештаненко Р.В., Цветников В.А. Описание свойств реакций в модели онтологии химии // Интеллектуальные системы. 2006, № 1 С. 132-143.

5. Артемьева И.Л., Высоцкий В.И., Рештаненко Н.В. Модульная модель онтологии органической химии. Свойства органических соединений. Информатика и системы управления. 2006. №1. С.121-132.
6. Химические объекты биологического интереса (ХОБИ) [Электронный ресурс] – Режим доступа: https://www.ebi.ac.uk/chebi/ свободный – Загл. с экрана
7. Adams, E. Cannon, P. Murray-Rust. ChemAxiom – An Ontological Framework for Chemistry in Science. [Электронный ресурс] – Nature Precedings, 2009 – Режим доступа: http://dx.doi.org/10.1038/npre.2009.3714.1 свободный – Загл. с экрана
8. Sankar P., Aghila G. Design and Development of Chemical Ontologies for Reaction Representation. // J. Chem. Inf. Model. 2006. Vol.46. P.2355-2368.
9. Амосова Е.С., Туманов В.Е. Преставление химических реакций, реагентов и их термохимических свойств в интеллектуальной системе по физической химии радикальных реакций в жидкой фазе с использованием онтологической модели предметной области. // Бутлеровские сообщения. 2014. Т. 39. № 7. СС. 39-46.
10. Амосова Е.С., Берзигияров П.К. Разработка представления онтологической модели по физической химии радикальных реакций отношениями реляционной базы данных и реализация его в предметно-ориентированной системе научной осведомленности. // Бутлеровские сообщения. 2016. Т.45. № 1. С. 152-158.
11. Прохоров А.И., Варламов Д.А., Амосова Е.С., Берзигияров П.К., Туманов В.Е. Внедрение предметных онтологий в систему научной аналитики по физической химии радикальных реакций // Научный сервис в сети Интернет: труды XVIII Всероссийской научной конференции (19-24 сентября 2016 г., г. Новороссийск). — М.: ИПМ им. М. В. Келдыша, 2016. (344 с.) с. 298-302.
12. Туманов В.Е., Прохоров А.И., Соловьева М.Е. Технология интеллектуальных агентов в предсказании физико-химических свойств органических молекул. // Современные проблемы науки и образования. – 2013. – № 6; URL: http://www.science-education.ru/113-11073
13. Banville D.L. Chemical Information Mining, Facilitating Literature-Based Discovery // CRC Press, Taylor & Francis Group 2009.
14. Hakenberg J., Schmeier S., Kowald A., Klipp E., Leser U. Finding kinetic parameters using text mining // A Journal of Integrative Biology. 2004. V. 8. N. 2. P. 131-152.
15. Мохов А.С., Толчеев В.О., Туманов В.Е. Классификация научных публикаций в области химической физики по русскоязычным и англоязычным названиям // Научный сервис в сети Интернет: труды XVIII Всероссийской научной конференции (19-24 сентября 2016 г., г. Новороссийск). — М.: ИПМ им. М. В. Келдыша, 2016. (344 с.) с.279-283.
16. Наумец А.А., Соколов В.Н., Туманов В.Е. Информационно-аналитическая система РАДАР: мониторинг материалов научных и научно-технических

конференций. // Приложение к журналу «Вестник Московского университета им. С.Ю. Витте. Серия 1: Экономика и управление». Материалы XLI Международной конференции, XI Международной конференции молодых ученых «Информационные технологии в науке, образовании, телекоммуникации и бизнесе (IT +SE`2013). Майская сессия. 2013. Ялта-Гурзуф, Украина, Крым. 264 с. С. 66-68.
17. Naumets A.A., Sokolov V.N., Tumanov V.E. Approccio allo sviluppo di metodi analisi scientometricisull'esempio la pubblicazione di conferenze scientifiche. // *Italian Science Review*. 2015; 8(29). PP. 30-39. Available at URL: http://www.ias-journal.org/archive/2015/august/Naumets.pdf
18. Наумец А.А., Соколов В.Н., Туманов В.Е. Предметно-ориентированная информационно-аналитическая система мониторинга научных исследований по публикациям конференций // Фундаментальные исследования. 2016. № 4. Часть. 3. С. 529-534.
19. Prokhorov A., Tumanov V., Amosova E. Classical potential barrier of liquid phase radical reactions and its simulation based on the experimental kinetic and thermochemical data using fuzzy neural networks // In Proceedings 2018 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO). Prague, Czech Republic. 19-21 May 2018, pp. 121–126.

References

1. Krallinger M., Rabal O., Lourenço A., Oyarzabal J., Valencia A. Information Retrieval and Text Mining Technologies for Chemistry // Chem. Rev. 2017, V. 117. P. 7673−7761.
2. NIST Chemistry WebBook. NIST Standard Reference Database Number 69. [Digital resource]. Access mode: http://webbook.nist.gov/chemistry/
3. Tumanov V., Gaifullin G. Subject-oriented science intelligent system on physical chemistry of radical reactions // Modern Advances in Intelligent Systems and Tools (Proc. Of IEA/AIE 2012). Ed. Wei Ding, He Jaing, Moonis Ali, Mingchu Li. Berlin; Heidelberg: Springer. 2012. P. 121-127.
4. Artem'eva I.L., Reshtanenko R.V., Cvetnikov V.A. Opisanie svojstv reakcij v modeli ontologii himii // Intellektual'nye sistemy. 2006, № 1 S. 132-143.
5. Artem'eva I.L., Vysockij V.I., Reshtanenko N.V. Modul'naja model' ontologii organicheskoj himii. Svojstva organicheskih soedinenij. Informatika i sistemy upravlenija. 2006. №1. S.121-132.
6. Chemical Entities of Biologigical Interest (ChEBI) [Digital resource]. Access mode: https://www.ebi.ac.uk/chebi/
7. Adams, E. Cannon, P. Murray-Rust. ChemAxiom – An Ontological Framework for Chemistry in Science. [Электронный ресурс] – Nature Precedings, 2009 – Режим доступа: http://dx.doi.org/10.1038/npre.2009.3714.1

8. Sankar P., Aghila G. Design and Development of Chemical Ontologies for Reaction Representation. // J. Chem. Inf. Model. 2006. Vol.46. P.2355-2368.
9. Amosova E.S., Tumanov V.E. Prestavlenie himicheskih reakcij, reagentov i ih termohimicheskih svojstv v intellektual'noj sisteme po fizicheskoj himii radikal'nyh reakcij v zhidkoj faze s ispol'zovaniem ontologicheskoj modeli predmetnoj oblasti. // Butlerovskie soobshhenija. 2014. T. 39. № 7. CC. 39-46.
10. Amosova E.S., Berzigijarov P.K. Razrabotka predstavlenija ontologicheskoj modeli po fizicheskoj himii radikal'nyh reakcij otnoshenijami reljacionnoj bazy dannyh i realizacija ego v predmetno-orientirovannoj sisteme nauchnoj osvedomlennosti. // Butlerovskie soobshhenija. 2016. T.45. № 1. S. 152-158.
11. Prohorov A.I., Varlamov D.A., Amosova E.S., Berzigijarov P.K., Tumanov V.E. Vnedrenie predmetnyh ontologij v sistemu nauchnoj analitiki po fizicheskoj himii radikal'nyh reakcij // Nauchnyj servis v seti Internet: trudy XVIII Vserossijskoj nauchnoj konferencii (19-24 sentjabrja 2016 g., g. Novorossijsk). — M.: IPM im. M. V. Keldysha, 2016. (344 s.) s. 298-302.
12. Tumanov V.E., Prohorov A.I., Solov'eva M.E. Tehnologija intellektual'nyh agentov v predskazanii fiziko-himicheskih svojstv organicheskih molekul. // Sovremennye problemy nauki i obrazovanija. – 2013. – № 6; URL: http://www.science-education.ru/113-11073 .
13. Banville D.L. Chemical Information Mining, Facilitating Literature-Based Discovery // CRC Press, Taylor & Francis Group 2009.
14. Hakenberg J., Schmeier S., Kowald A., Klipp E., Leser U. Finding kinetic parameters using text mining // A Journal of Integrative Biology. 2004. V. 8. N. 2. P. 131-152.
15. Mohov A.S., Tolcheev V.O., Tumanov V.E. Klassifikacija nauchnyh publikacij v oblasti himicheskoj fiziki po russkojazychnym i anglojazychnym nazvanijam // Nauchnyj servis v seti Internet: trudy XVIII Vserossijskoj nauchnoj konferencii (19-24 sentjabrja 2016 g., g. Novorossijsk). — M.: IPM im. M. V. Keldysha, 2016. (344 s.) c.279-283
16. Naumec A.A., Sokolov V.N., Tumanov V.E. Informacionno-analiticheskaja sistema RADAR: monitoring materialov nauchnyh i nauchno-tehnicheskih konferencij. // Prilozhenie k zhurnalu «Vestnik Moskovskogo universiteta im. S.Ju. Vitte. Serija 1: Jekonomika i upravlenie». Materialy XLI Mezhdunarodnoj konferencii, XI Mezhdunarodnoj konferencii molodyh uchenyh «Informacionnye tehnologii v nauke, obrazovanii, telekommunikacii i biznese (IT +SE`2013). Majskaja sessija. 2013. Jalta-Gurzuf, Ukraina, Krym. 264 s. S. 66-68.
17. Naumets A.A., Sokolov V.N., Tumanov V.E. Approccio allo sviluppo di metodi analisi scientometricisull'esempio la pubblicazione di conferenze scientifiche. // *Italian Science Review*. 2015; 8(29). PP. 30-39. Available at URL: http://www.ias-journal.org/archive/2015/august/Naumets.pdf
18. Naumec A.A., Sokolov V.N., Tumanov V.E. Predmetno-orientirovannaja informacionno-analiticheskaja sistema monitoringa nauchnyh issledovanij po

publikacijam konferencij // Fundamental'nye issledovanija. 2016. № 4. Chast'. 3. S. 529-534.
19. Prokhorov A., Tumanov V., Amosova E. Classical potential barrier of liquid phase radical reactions and its simulation based on the experimental kinetic and thermochemical data using fuzzy neural networks // In Proceedings 2018 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO). Prague, Czech Republic. 19-21 May 2018, pp. 121–126.